Hajipour *et al. EURASIP Journal on Wireless Communications and Networking* (2015) 2015:263
DOI 10.1186/s13638-015-0481-4

EURASIP Journal on Wireless Communications and Networking
a SpringerOpen Journal

RESEARCH Open Access

Utility-based efficient dynamic distributed resource allocation in buffer-aided relay-assisted OFDMA networks

Javad Hajipour[1*], Amr Mohamed[2] and Victor C. M. Leung[1]

Abstract

In this paper, we study resource allocation in buffer-aided relay-assisted OFDMA networks. We consider utility-based stochastic optimization framework where there are constraints to be met either instantaneously or in average sense. Using the well-known Lyapunov drift-plus-penalty policy, we extract the instantaneous problem that needs to be solved in each slot to control the data admission and allocate the time slots, power, and subchannels. We propose the parameters that should be taken into account in utilizing the drift-plus-penalty policy in relay-assisted cellular networks, for providing fair data admission and satisfying the average power constraints. We introduce a low-complexity strategy for power and subchannel allocation and propose distributed and centralized algorithms to utilize it. Specifically, the proposed efficient dynamic distributed resource allocation (EDDRA) scheme is suitable for use in practice as it imposes less overhead on the system and splits the resource allocation tasks among the base station (BS) and the relays. Extensive simulation results show the effectiveness of the proposed parameters in meeting the objective and the constraints of the studied problem. We also show that the proposed EDDRA scheme has close performance to the proposed centralized one and outperforms an existing centralized scheme.

Keywords: OFDMA, Regenerative buffering relays, Distributed resource allocation, Low complexity

1 Introduction

Relay-assisted orthogonal frequency division multiple access (OFDMA) networks are the promising solutions for providing high-speed data services in wide coverage areas and therefore, they have been accepted in the standardization bodies such as IEEE 802.16 [1] and long term evolution-advanced (LTE-A) [2] for providing wireless access to the customers. Resource allocation is an important factor in utilizing the capacities of these networks; while the combination of OFDMA and relaying techniques results in high benefits and opportunities, it also brings challenges and issues that need to be addressed for exploiting those opportunities [3]. There have been extensive works in this area in the recent years. In [4], the authors aimed at utilizing cross layer optimization framework for resource allocation in cooperative decode-and-forward (DF) relaying networks. For that, they introduced virtual links and nodes to embed the cooperation mechanism into the optimization framework. They assumed half-duplex relaying and also considered spatial reuse of the spectrum among the links with lower mutual interference on each other. Using dual method, then, they maximized the balanced end-to-end throughput. On the other hand, [5] considered reusing the OFDMA resources only among the different access links (from the relays to their users) and also suggested adaptive segmentation of the frame for transmissions on the base station (BS)-to-relay links and relay-to-user links. To optimize these, the authors proposed linear-programming-based and greedy algorithms which led to significant improvements in the system capacity. However, the proposed algorithms in [4, 5] were centralized which impose high computational burden on the BS and high signaling overhead for reporting channel state information (CSI) of the links. To alleviate these drawbacks, several other works have studied distributed resource allocation [6–9]. In [6], cross layer scheduling in an OFDMA amplify-and-forward (AF) relay

*Correspondence: hajipour@ece.ubc.ca
[1] ECE Department, The University of British Columbia, Vancouver, Canada
Full list of author information is available at the end of the article

network was studied to maximize the received goodput for the relayed users, taking into account the effects of imperfect CSI at the transmitter. Based on dual decomposition, a distributed algorithm was proposed for power and subcarrier allocation. Similarly, [7] proposed a distributed algorithm for power and subchannel allocation based on dual decomposition, with the difference that the authors studied multiple-input multiple-output (MIMO) transmissions in a system with the possibility to dynamically select full-duplex or half-duplex, and AF or DF relaying. Pan et al. [8] investigated distributed power allocation algorithms in the presence of *cognitive* relays, based on convex optimization. They considered the cases with and without fairness considerations, with and without low signal-to-noise ratio (SNR) limitation on the BS-to-relay links, assuming fixed subcarrier allocations. Moreover, they proposed a distributed scheme for joint power and subcarrier allocation. In contrast with the above works, [9] proposed a low-complexity semi-distributed algorithm for power and subcarrier allocation, by considering complete/limited information about the CSI of the relay-to-user links at the relays/BS and assigning the subcarriers based on them. More other works studied frequency reuse schemes to improve the system capacity [10–12].

The common assumption in most of the literature in this area is that the relays do not have buffer to store packets for later transmission. Therefore, they have to forward their received data immediately in the following transmission interval. Recently it has been shown that the use of buffers in relays can improve the system capacity [13–15]. This is achieved as a result of more flexibility in transmissions from relays. In other words, buffering capability in relays enables them to postpone the data forwarding for a user if its channel is not good and use the wireless resources for the links with higher quality of channel.

Even though using buffers helps in compensating for the effect of channels in wireless networks, it also brings new challenges. In general, for better utilization of buffers, it is needed to take into account the queue dynamics in them. In this regard, different problems arising in different parts of these networks have been addressed in the literature. Yang et al. [16] studied adaptive media playout for video streaming, taking into account the queue dynamics in the receiver. By monitoring the queue size and its variation, a model was developed for buffer underflow probability estimation which was exploited to smoothly control frame separations of video traffic. In [17], a novel framework was proposed for online source rate control, where a buffer overflow probability (BOP) estimator monitors the queue sizes and its variation at the BS, and sends feedback about the estimated BOP to the video source to control its bit rate.

With the introduction of buffers in relays, several works also studied the challenges that arise for resource allocation in these networks. In [18], the authors studied resource allocation in a system with quasi full-duplex (quasi-FD) relays, where each relay is able to simultaneously receive and transmit on orthogonal channels. They considered symmetric traffic for the users and proposed centralized algorithms for joint routing and subchannel allocation, to provide load balance among the cell nodes and fairness among the users. Compared with that, [19] studied a system with half-duplex (HD) relays where each relay can only receive in the first half of the frame and transmit in the second half. Then, using queue-length coupling across subframes, the authors proposed centralized joint routing and subchannel allocation for each subframe and studied the system's performance under both symmetric and asymmetric data traffic. In [20], the authors considered quasi-FD relaying and formulated a convex optimization problem for joint power and subchannel allocation. Assuming a single power constraint for the whole system, a distributed resource allocation framework was proposed based on dual decomposition. On the contrary, [21] considered HD relays with individual peak and average power constraints for the BS and relays, and studied joint optimization for subframe, subchannel, and power allocation in LTE-A systems, where each subframe can either be used for transmissions on the BS-to-relays and BS-to-users links or the BS-to-users and relays-to-users links. Using utility-based stochastic network optimization [22, Chapter 5], and dual decomposition, optimal solution was provided for data admission into the network, and an iterative distributed algorithm was proposed for reaching the optimal resource allocation. In Table 1, we have classified the most relevant references cited above, based on the usage of buffer and centralized or distributed approach of resource allocation algorithms.

The work in [21] is a pioneer in utilizing the Lyapunov drift-plus-penalty framework [22, Chapter 5] for data admission and resource allocation in OFDMA relay networks. However, it does not take into account some constraints and challenges that arise in practice in such networks. In particular, the proposed iterative algorithm to get the optimal solution incurs very low convergence rate due to the separate power constraints for the BS

Table 1 Classification of the most relevant resource allocation references

Relaying without buffering	Centralized	[4, 5]
	Distributed	[6–9]
Relaying with buffering	Centralized	[18, 19]
	Distributed	[20, 21]

and relays. This is not suitable for the practical scenarios where the scheduling is performed in the units of millisecond. Also, it does not take into account the constraint that might be imposed on the buffer capacity in relays. Other than that, in the drift-plus-penalty framework, the average of a variable is defined over infinite time horizon, which requires some considerations for achieving the desired objectives and satisfying the constraints in practical systems. To the best of our knowledge, none of the existing works on OFDMA relay networks has studied resource allocation with the abovementioned constraints altogether. This paper aims at addressing these issues and filling the gaps highlighted above.

In summary, we study low-complexity utility-based resource allocation in buffer-aided relay-assisted OFDMA networks, with HD relays, based on stochastic optimization framework presented in [22, Chapter 5]. We consider the network utility as a function of average data admission of the users and aim at maximizing it subject to the long term and instantaneous constraints. Note that in our previous work [20], data admission was not studied. Also, the resource allocation problem only had a single power constraint for the whole system, and there was no constraint on the transmission rates of the subchannels (like finite data availability and limited buffer capacity). Therefore, it was possible to convert it to a convex optimization problem and use dual decomposition for power and subchannel allocation. However, in the current work, we consider several practical constraints which necessitate a new solution approach. In addition to those constraints, the contributions of our current paper can be classified into two categories: one category is the identification of important factors that should be taken into account in the instantaneous problem formulation. The other one is the design of low-complexity algorithms for solving the resource allocation subproblem. Specifically, the main contributions are as follows:

- We identify the factors that need to be taken into account for adapting the Lyapunov drift-plus-penalty policy for relay-based cellular networks. In particular, we propose to consider an importance parameter for average power constraint, to satisfy that constraint in a reasonable time period for practical scenarios. Also, we propose to add extra weight for the BS-to-relays and relays-to-users links, in the cases that the fairness is also an objective in the utility-based data admission control.
- We aim at low-complexity algorithms for time slot, subchannel, and power allocations and highlight the challenges even in such algorithms due to the lack of a priori knowledge about the subchannel sets and total power usage of the BS and relays in each time slot. Then, we propose a low-complexity strategy for breaking the ties and making the correlations tractable, which can be used in both centralized and distributed resource allocation implementations.
- We focus on distributed mechanism for resource allocation and propose low-complexity algorithms for deciding about the type of time slot, subchannel sets of the nodes, and subchannel and power allocations to the links of the nodes. Based on that, we also present a low-complexity centralized mechanism which needs more signaling overhead and can be used as a baseline.
- We take into account practical constraints such as HD relaying operation, average and peak power constraint for each of the BS and relays, as well as limits on data availability and buffer capacity.
- Using extensive simulations, we demonstrate the effectiveness of the introduced parameters and also evaluate the performance of the proposed algorithms. We observe that the distributed scheme has very close performance to the centralized one and outperforms an existing centralized scheme proposed in [19]. Also, we show that our proposed algorithms have similar or even better performance compared with the iterative algorithm proposed in [21] for reaching the optimal solution.

The rest of the paper is organized as follows. Section 2 describes the system model and the stochastic problem formulations. In Section 3, we state the subproblems and challenges as well as the proposed parameters and algorithms. Numerical results are provided in Section 4, with the conclusion finally presented in Section 5.

2 Preliminaries

In this section, we present the system model and the stochastic problem formulation. Then, we present the transformed version of the problem and introduce the virtual queues which make it possible to exploit the Lyapunov drift-plus-penalty policy in the next section. Hereafter, for easiness, we will use the term "drift-plus-penalty" instead of "Lyapunov drift-plus-penalty".

2.1 System model

We consider the downlink of a single cell relay-assisted OFDMA network, as shown in Fig. 1. It is assumed that each user is connected to either the BS or one of the relays, meaning that it receives service from only one of them. This is decided at the beginning of users' connection to the network and through handshaking procedures between the BS, relays, and users about the signal strengths that users can receive from the BS and relays. Users, relays, and subchannels are indexed, respectively,

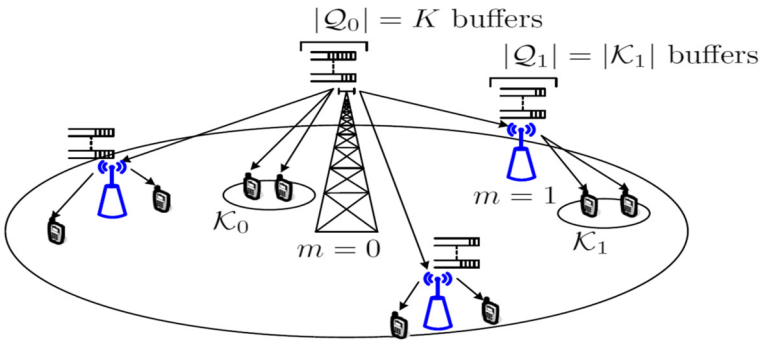

Fig. 1 System model for buffer-aided relay-assisted network

by $k \in \mathcal{K} = \{1, \ldots, K\}$, $m \in \mathcal{M} = \{1, \ldots, M\}$ and $n \in \mathcal{N} = \{1, \ldots, N\}$. Table 2 presents the key notations used throughout this paper. We use the term "serving node" or simply "node" to refer to any of the BS or relays and show the set of all nodes by $\mathcal{B} = \{0, 1, \ldots, M\}$, where $m = 0$ indicates the BS. Also, we use $\mathcal{K}_m$ to denote the set of users that have a direct link to node $m \in \mathcal{B}$. On the other hand, $m(k)$ is used to refer to the node directly serving user k.

We assume that time is divided into the units of slot, where each time slot can be either type A or type B. In type A slots, the BS transmits to users directly connected to it, or to the relays; in type B slots, the BS and relays can transmit only to the users connected to them and therefore, there is no transmission from the BS to relays. This transmission format is based on LTE-A with type 1 relays where the BS-to-relays transmissions and relays-to-users transmissions use the same bandwidth but over different time slots, to prevent the interference between transmit and receive antennas.

We assume that the MAC layers of the BS and relays are equipped with buffers, where the BS has one for each user but every relay has one for each of the users connected to it. We denote the set of the users that have a buffer in node $m \in \mathcal{B}$ by $\mathcal{Q}_m$; therefore, we have $\mathcal{Q}_0 = \mathcal{K}$ and $\mathcal{Q}_m = \mathcal{K}_m, \forall m \in \mathcal{M}$. These notations are defined to make the formulations and algorithms shorter. The data admitted into a BS buffer are queued until transmission to the corresponding direct user, or to the corresponding buffer in the relay serving the user. Similarly, data arrived at the relays' buffers are queued until transmission to their users.

Note that in the following, when we use the term "the link of user k from node m", we mean "the link that serves the queue of user k in node m", which might be a *direct* link between the BS and a user, a *feeder* link between the BS and a relay or an *access* link between a relay and a user connected to it. We use $e^m_{kn}(t)$ for the link of user k from node m to denote the channel gain-to-noise ratio at the receiver side on subchannel n in time slot t. It is assumed that the channel conditions of the links vary over time and frequency, but remain constant during one time slot and over one subchannel. In the following, when we remove the (t) argument from the variables, we imply them in a general transmission incident. We assume that the BS and relays use M-ary QAM modulation for their transmissions; therefore, the achievable transmission rate on the link of user k from node m on subchannel n in time slot t can be computed as follows [23]:

$$r^m_{kn}(t) = B \log_2 \left(1 + \frac{p^m_n(t) e^m_{kn}(t)}{\Gamma_k}\right), \tag{1}$$

where B is the bandwidth of a subchannel. Γ_k is the SNR gap due to the limited number of coding and modulation schemes and is related to the bit error rate of user k (BER_k), through equation $\Gamma_k = -\frac{\ln(5BER_k)}{1.5}$ [23]. $p^m_n(t)$ denotes the power allocated by node m on subchannel n in time slot t. We indicate the total power used by node m in time slot t as $P_m(t) = \sum_{n=1}^{N} p^m_n(t)$. Using (1), the total transmission rate on the link of user k from node m can be written as $r^m_k(t) = \sum_{n=1}^{N} x^m_{kn}(t) r^m_{kn}(t)$, where $x^m_{kn}(t) \in \{0, 1\}$ denotes the subchannel allocation indicator which will be one if subchannel n is used for transmission on the link of user k from node m in time slot t, and zero otherwise. Note that for any n, in type A time slot, x^m_{kn} should be set to zero for $m \in \mathcal{M}, k \in \mathcal{K}_m$ and in type B time slot, x^0_{kn} should be set to zero for $k \in \mathcal{K} - \mathcal{K}_0$.

In each time slot, a resource allocation policy determines the type of time slot, subchannel, and power allocations for the different links of the system. Based on that, the BS and relays transmit data from their queues, and at the end, the queue sizes are updated as follows:

$$Q^m_k(t+1) = \min\left[L^m_k, \max\left[Q^m_k(t) - T r^m_k(t), 0\right] + a^m_k(t)\right], \quad \forall m \in \mathcal{B}, k \in \mathcal{Q}_m \tag{2}$$

Table 2 Notation Summary

Notation	Description
$\mathcal{N}, N$	Set and total number of subchannels, respectively
$\mathcal{M}, M$	Set and total number of relays, respectively
$\mathcal{K}, K$	Set and total number of users, respectively
$\mathcal{B}$	Set of all the serving nodes, including the BS and relays
$\mathcal{K}_m$	Set of users connected to node m
$\mathcal{Q}_m$	Set of users that have a buffer in node m
$m(k)$	Serving node of user k
T	Duration of a time slot
B	Bandwidth of a subchannel
Γ_k	SNR gap for user k
$e_{kn}^m(t), x_{kn}^m(t)$	Channel gain-to-noise ratio and subchannel allocation indicator of subchannel n in time slot t, respectively, for the link of user k from node m
$r_{kn}^m(t), \tilde{r}_{kn}^m(t)$	Achievable transmission rate and estimated achievable transmission rate on subchannel n in time slot t, respectively, for the link of user k from node m
$r_k^m(t)$	Total transmission rate on the link of user k from node m in time slot t
$p_n^m(t)$	Power used by node m on subchannel n in time slot t
$P_m(t), \hat{P}_m, P_m^{av}$	Total power to be used in time slot t, peak power constraint and average power constraint, respectively, for node m
$L_k^m, Q_k^m(t)$	Capacity of MAC layer buffer of user k in node m and the queue size in it in time slot t, respectively
$a_k^m(t)$	Size of data arrived at the MAC layer buffer of user k in node m in time slot t
$\hat{a}$	Upper bound of data admission into a buffer in the MAC layer of the BS
$J_k, Y_k(t), A_k(t)$	Capacity, queue size and the arrived data size in time slot t, respectively, in the top layer buffer of user k in the BS
$\mathcal{U}(.), V$	Utility function and the value coefficient for it, respectively
$\gamma_k(t), G_k(t)$	Auxiliary variable corresponding to data admission of user k and its corresponding virtual queue size, respectively, in time slot t
$Z_m(t), I$	Virtual power queue size of node m in time slot t and the importance factor, respectively, corresponding to average power constraints
ρ_k^m, W_e	Indicator variable for the link of user k from node m and the extra weight, respectively, for providing fair data admission
$\tilde{N}_m, \hat{N}_m$	Estimation of node m for its number of subchannels and the upper bound considered by the BS on the number of subchannels for node m, respectively
D_n^m, D^m	Average demand of relay m on subchannel n and average total demand of node m, respectively
D_n^{0a}, D^{0a}	Average demand of the BS on subchannel n and average total demand of the BS, respectively, for type A time slot
D_n^{0b}, D^{0b}	Average demand of the BS on subchannel n and average total demand of the BS, respectively, for type B time slot
D^A, D^B	Total demand for type A and type B time slots, respectively

where T is time slot duration, and L_k^m and $Q_k^m(t)$, respectively, denote the buffer capacity of user k in node m and the size of data queued in it in time slot t. Data arrival process $a_k^m(t)$ into a relay buffer is in fact the departure process from the BS, and therefore, we have $a_k^m(t) = \min\left[Q_k^0(t), Tr_k^0(t)\right], \forall m \in \mathcal{M}, k \in \mathcal{K}_m$. On the other hand, the arrival processes in the BS buffers are managed through a data admission control policy in the MAC layer, which decides to admit data or not to admit from the queues of the top layer buffers in the BS; this will be clarified in Section 3.2 and through (10). The size of the queues in top layer are updated as follows:

$$Y_k(t+1) = \min\left[J_k, \max\left[Y_k(t) - a_k^0(t), 0\right] + A_k(t)\right], \ \forall k \in \mathcal{K} \tag{3}$$

where J_k and $Y_k(t)$, respectively, denote the buffer capacity of user k in the top layer of the BS and the size of data queued in it in time slot t. $A_k(t)$ is the amount of data arrived in time slot t for user k, according to an exogenous stochastic process, which is assumed to be stationary and ergodic. We assume that due to the processing limitations, an upper bound of $\hat{a}$ is imposed on the amount of data admitted into the MAC layer buffers, and therefore, we have $a_k^0(t) \leq \hat{a}, \forall k, t$. Note that in the above, no specific assumptions have been considered for queueing mechanisms other than the ordinary first-in-first-out (FIFO) operation and the layer-based architecture of data management, which has been considered in [24] and [21] as well. In this paper, for simplicity, we have only considered the top layer and MAC layer in the BS and the MAC layer in the relays. However, the discussions can be easily extended to the cases where the queueing in different layers are also taken into account.

2.2 Stochastic problem formulation

We note that in a realistic scenario, the BS buffers are not infinitely backlogged and are fed by stochastic data arrivals. This makes it necessary to take into account the queue dynamics, in the design of resource allocation algorithms, in addition to the randomness caused by the wireless channels. Therefore, the *average* performance metrics become important for network operators, and especially, the average throughput and average power constraints are the issues that need to be managed. In the following, we will explain these in more detail.

We define the time average expectation of a stochastic variable $\nu(t)$ as $\overline{\nu} = \lim_{\tau\to\infty} \frac{1}{\tau} \sum_{t=0}^{\tau-1} E[\nu(t)]$, according to [22, Chapter 4]. Considering the abovementioned, we aim at controlling the data traffic and resource allocation,

by addressing the following stochastic optimization problem:

$$\max_{\boldsymbol{a}^0,\boldsymbol{x},\boldsymbol{p}} \sum_{k=0}^{K} \mathcal{U}\left(\overline{a}_k^0\right), \tag{4a}$$

$$\text{s.t.} \quad C1: \overline{P}_m \le P_m^{av}, \quad \forall m \in \mathcal{B}, \tag{4b}$$

$$C2: r_k^m(t)T \le Q_k^m(t), \quad \forall m \in \mathcal{B}, k \in \mathcal{Q}_m, \tag{4c}$$

$$C3: r_k^0(t)T \le (L_k^m - Q_k^m(t)), \forall m \in \mathcal{M}, k \in \mathcal{K}_m, \tag{4d}$$

$$C4: r_{kn}^m(t) \le B\hat{s}, \quad \forall m \in \mathcal{B}, k \in \mathcal{Q}_m, \forall n, \tag{4e}$$

$$C5: a_k^0(t) \le \min\left[\hat{a}, Y_k(t)\right], \quad \forall k \in \mathcal{K}, \tag{4f}$$

$$C6: a_k^0(t) \le \left(L_k^0 - Q_k^0(t)\right) \quad \forall k \in \mathcal{K}, \tag{4g}$$

$$C7: P_m(t) \le \hat{P}_m, \quad \forall m \in \mathcal{B}, \tag{4h}$$

$$C8: \sum_{m\in\mathcal{B}} \sum_{k\in\mathcal{Q}_m} x_{kn}^m(t) \le 1, \quad \forall n \in \mathcal{N}, \tag{4i}$$

$$C9: \{x_{kn}^m(t)\} \text{ comply to the transmission rules of either type A or type B slot} \tag{4j}$$

where $\mathcal{U}(.)$ is the utility function and $C1$ is to limit the average power consumption of each node. $C2$ shows that a finite amount of data can be transmitted from each queue and $C3$ is to prevent the incidents of more transmissions to relay buffers than they can accommodate. $C4$ indicates the limit on the availability of modulation schemes, where $\hat{s}$ is the spectral efficiency of the highest order modulation in the system; considering it helps in controlling the power allocation and preventing overflows from relays' buffers (This will be explained clearly in Section 3). $C5$ and $C6$, respectively, show the limit on the data admission from the top layer buffers of the BS and the limit on the available buffer space in the MAC layer of the BS. $C7$ indicates the maximum instantaneous power, $\hat{P}_m$, that node m can use for transmissions, $C8$ shows that each subchannel can be allocated to only one link, and $C9$ is to use the feasible values for subchannel allocation variables $\{x_{kn}^m\}$.

The utility function in (4a) makes it possible to control the data admission for the users based on the objective of the network operator. For example for maximizing the total throughput, $\mathcal{U}(z) = z$ can be used or for providing proportional fairness, $\mathcal{U}(z) = \log(z)$ can be considered [22, Chapter 5]. We assume that $\mathcal{U}(z)$ is a concave and continuous function of z.

Note that the problem (4) has two types of constraints. While $C1$ needs to be satisfied in long term, $C2 - C9$ state the constraints that must be met in each time slot. In particular, P_m^{av} is different from $\hat{P}_m$, as the former can be set to limit the power consumption costs or the circuit heating but the latter is imposed by the system hardware (such as power amplifiers' linear operation characteristics or maximum available instantaneous power) and is larger than P_m^{av}.

2.3 Transformed problem and virtual queues

We note that the objective in problem (4) is a function of time average of users' data admission rate. In order to utilize the drift-plus-penalty policy, it is needed to have the objective function as a time average expression. Similar to [22, Chapter 5], we define auxiliary variables $0 \le \gamma_k(t) \le \hat{a}, k = 1, \ldots, K$, corresponding to each $a_k^0(t), k = 1, \ldots, K$. Then, the problem (4) can be transformed into the following equivalent problem, in which the objective is the time average of a function:

$$\max_{\boldsymbol{a}^0,\boldsymbol{x},\boldsymbol{p},\boldsymbol{\gamma}} \sum_{k=0}^{K} \overline{\mathcal{U}(\gamma_k)}, \tag{5a}$$

$$\text{s.t.} \quad C1 - C9, \tag{5b}$$

$$C10: \overline{\gamma}_k \le \overline{a}_k^0, \forall k \in \mathcal{K}, \tag{5c}$$

$$C11: 0 \le \gamma_k \le \hat{a} \tag{5d}$$

We also define the virtual power queues $Z_m(t)$ and virtual auxiliary queues $G_k(t)$, respectively, corresponding to the constraints $C1$ and $C10$, with the following updating equations:

$$Z_m(t+1) = \max[\, Z_m(t) + P_m(t) - P_m^{av}, 0\,], \quad \forall m \in \mathcal{B} \tag{6a}$$

$$G_k(t+1) = \max[\, G_k(t) + \gamma_k(t) - a_k^0(t), 0\,], \quad \forall k \in \mathcal{K} \tag{6b}$$

Based on the abovementioned, we are able now to define the instantaneous problem (with the objective and constraints stated in terms of the instantaneous values of the variables) and study the algorithms for solving it, which will be presented in the next section.

3 Cross layer traffic control and resource allocation

In this section, we first describe the instantaneous problem to be addressed in each time slot and propose some parameters that need to be included in it to make it suitable for relay-assisted cellular networks. Then, we present the data admission subproblem and discuss the factors influencing it. After that, we highlight the issues in solving the resource allocation subproblem and propose a low-complexity strategy to address them. Then, we provide a low-complexity distributed scheme which uses the proposed strategy through four steps to decide about the allocation of time slots, power and subchannels. Finally, we present a low-complexity centralized algorithm and describe its required steps.

3.1 Instantaneous problem

To address the problem (5), based on the drift-plus-penalty policy [22, Chapter 5], we define the "instantaneous" problem in time slot t as follows:

$$\max_{\boldsymbol{a}^0,\boldsymbol{x},\boldsymbol{p},\boldsymbol{\gamma}} V\sum_{k=0}^{K}\mathcal{U}(\gamma_k)+\sum_{k=0}^{K}G_k(t)\left[a_k^0(t)-\gamma_k(t)\right] + I\sum_{m=0}^{M}Z_m(t)\left[P_m^{av}-P_m(t)\right] + \sum_{m=0}^{M}\sum_{k\in\mathcal{Q}_m}\left(Q_k^m(t)+\rho_k^m W_e\right)\left[r_k^m(t)T-a_k^m(t)\right], \quad (7a)$$

$$\text{s.t.} \quad C2-C9, C11 \quad (7b)$$

where $V>0$ is the value that can be given to the objective (5a), and by that, we can trade off higher utility to larger queue sizes [22]. This will be clarified later. $I>0, \rho_k^m\in\{0,1\}$ and $W_e>0$ are the parameters that we propose in this paper, to adapt the drift-plus-penalty policy to relay-assisted cellular networks. I is the importance factor that we give to average power constraint, through which we can prevent the continuous growth of the virtual power queues and consequently, we can meet the average power constraints in shorter time. W_e is an extra positive weight, that can be given to the feeder links from the BS to relays and the access links from relays to users, in the cases that the fair admission of users' data is of our concern. ρ_k^m is the indicator variable to specify the cases to apply W_e; in particular, it is set to zero unless when the fair data admission is desired and the corresponding queue (either in the BS or in a relay) belongs to an indirect user, i.e., $k\in\mathcal{K}-\mathcal{K}_0, m\in\mathcal{B}$. Proposition of I, ρ_k^m and W_e is one of the main contributions of our paper and will be discussed later. Based on the above, we can see that the work in [21] is in fact a special case of our work, in which $I=1$, $\rho_k^m=0$, $W_e=0$ and no limitations on buffer capacities or modulation schemes are considered.

It is observed from (7a) that, using the drift-plus-penalty policy, the instantaneous objective in each time slot includes four terms: the first term corresponds to the long term objective (5a) and the rest correspond to serving the actual queues and stabilizing the virtual queues (to meet the constraints $C1$ and $C10$). We note that due to the limited buffer capacities, the actual queues of the system are always stable. However, using drift-plus-penalty policy provides a useful framework for channel- and queue-aware resource allocation which takes into account both the channel states and the data availability in the system's actual queues. It also makes it possible to stabilize the virtual queues $\{G_k(t)\}$ and $\{Z_m(t)\}$. It is shown in [22, Chapter 5] that solving the instantaneous problem obtained from drift-plus-penalty policy in the case of infinite buffer capacities (i.e., when the constraints $C3$ and $C6$ are not imposed) provides a utility (4a) that has a gap of $\frac{D}{V}$ from its optimal value, where D is a constant related to sum of the squared data arrivals and squared transmission rates. If the solution algorithm for instantaneous problem (7) also leads to an approximate value of the objective function (7a) within distance C from its optimal value (e.g., due to a suboptimal solution), then the aforementioned gap will be $\frac{D+C}{V}$. Therefore, there is an $O(1/V)$ gap between the optimal utility of the stochastic problem and the utility obtained from solving the instantaneous problem, which can be arbitrarily reduced by choosing a large V. However, this will lead to larger queue sizes and delays due to the fact that the upper bound on the queue sizes has an $O(V)$ expression. Recently, [25] has shown that in the case of finite buffer capacity of β, the utility obtained from drift-plus-penalty policy is within $O(1/V)+O(e^{-\beta})$ of its optimal value. Note that [25] assumes that each node is allowed to transmit to the next hop even if it does not have enough buffer space, in which case the received packet is dropped. However, this is not allowed in the system considered in our paper, due to the constraint $C3$, as it will waste the (expensive) frequency resources. This makes the mathematical analysis of the drift-plus-penalty policy difficult in our system model and can be investigated in future works.

It is worth mentioning that considering V, I, ρ_k^m, and W_e in (7a) facilitates reaching our goals for system utility and constraints in different scenarios; neglecting them is in fact like setting them based on a fixed scenario (i.e., $V=I=1$ and $W_e=0$) which would lower the usefulness of the drift-plus-penalty policy. Note that V and I can be tuned easily by considering the range of values for weighted rates in (7a), affected by packet sizes and transmission rates. On the other hand, ρ_k^m can be easily set as stated above, when fairness is an objective; then, W_e can be tuned by increasing its value from zero towards the values in the range of queue sizes in relays, depending on how much we trade off between data admission for the direct and indirect users. These will get clearer in the next section, when we discuss their effects.

Similar to [22], by rearranging the terms in the instantaneous problem, it can be divided into three instantaneous Subproblems (SPs) which will be presented in the next subsections.

3.2 Traffic control and data admission

The first subproblem of (7) is related to the auxiliary variables as follows:

$SP1$ (auxiliary variable subproblem) :

$$\max_{\boldsymbol{\gamma}}\sum_{k=1}^{K}\Big(V\mathcal{U}(\gamma_k(t))-G_k(t)\gamma_k(t)\Big), \quad (8a)$$

$$\text{s.t.}\ 0\le\gamma_k(t)\le\hat{a}, \forall k\in\mathcal{K} \quad (8b)$$

SP1 is a simple convex optimization problem and therefore, the BS can solve it easily. As an example, for the proportional fairness, i.e., when $\mathcal{U}(\gamma_k) = \log(\gamma_k)$, after taking the derivatives, the BS can determine the optimal auxiliary variables in time slot t through $\frac{1}{\gamma_k(t)} = \frac{G_k(t)}{V} \Longrightarrow \gamma_k(t) = \min\left(\hat{a}, \frac{V}{G_k(t)}\right), \forall k$. After computing the solution, the BS can update the corresponding virtual queue sizes based on (6b).

The second subproblem of (7) is related to the flow control and data admission into the users' buffers in the MAC layer of the BS, as follows:

*SP*2 (flow control subproblem) :

$$\max_{\boldsymbol{a}^0} \sum_{k=1}^{K} \Big(G_k(t) - Q_k^0(t)\Big) a_k^0(t), \tag{9a}$$

$$\text{s.t. } 0 \le a_k^0(t) \le \hat{a}, \forall k \in \mathcal{K}, \tag{9b}$$

$$a_k^0(t) \le \min\left[Y_k(t), L_k^0 - Q_k^0(t)\right], \forall k \in \mathcal{K} \tag{9c}$$

which is a linear problem; by solving it, the BS can determine the optimal data admissions in time slot t as:

$$a_k^0(t) = \begin{cases} \min\left[Y_k(t), \min\left[\hat{a}, L_k^0 - Q_k^0(t)\right]\right], & Q_k^0(t) \le G_k(t) \\ 0 & \text{, otherwise} \end{cases} \tag{10}$$

Based on the above, for deciding about data admission for user k, the BS needs the information about its top layer queue size $Y_k(t)$, MAC layer queue size $Q_k^0(t)$, auxiliary variable's virtual queue size $G_k(t)$, maximum admissible data size $\hat{a}$, and the MAC layer's buffer capacity L_k^0. All of these information are accessible to the BS as the related buffers and queues are implemented in the BS. $\hat{a}$ and L_k^0 are fixed parameters; $Y_k(t)$, $Q_k^0(t)$ and $G_k(t)$ are in fact the input variables to the data admission procedure in the MAC layer of the BS and the decision about the size of admitted data is the output. This is clarified in Fig. 2. After determining data admissions and transmissions (discussed later), the BS can use (2) and (3) to update the affected queue sizes. It is worth mentioning that the data admission procedure presented above is based on [22, Chapter 5], and its inclusion here is for completeness of the discussions presented later.

Note that based on (10), whenever the size of an actual queue at the BS, $Q_k^0(t)$, is larger than the virtual queue $G_k(t)$, no data is admitted into the corresponding BS buffer. This can happen for several time slots, in buffer-aided relay-assisted cellular networks, due to the admission of a large packet and low service rate of the queue of an indirect user at the BS (which is caused by the differential backlog terms described in the next subsection). Consequently, even using a utility function with fairness property and large value for parameter V does not necessarily lead to fair data admission, and therefore, more considerations are needed in resource allocation for serving the queues. This is the motivation for proposing ρ_k^m and the extra weight W_e (explained clearly later, in Remark 1), which help to improve the fair data admission for indirect users.

3.3 Resource allocation challenges

By substituting (1) in (7a) and removing the constant terms, the last and most important subproblem, which is to decide about the time slot, subchannel and power allocations, can be stated as

*SP*3 (resource allocation subproblem) :

$$\max_{\boldsymbol{p},\boldsymbol{x}} \sum_{n=1}^{N} \sum_{m=0}^{M} \left(\sum_{k \in \mathcal{Q}_m} BT w_k^m(t) \log_2\left(1 + \frac{p_n^m(t) e_{kn}^m(t)}{\Gamma_k}\right) - IZ_m(t) p_n^m(t) \right), \tag{11a}$$

$$\text{s.t.} C2 - C4, C7 - C9 \tag{11b}$$

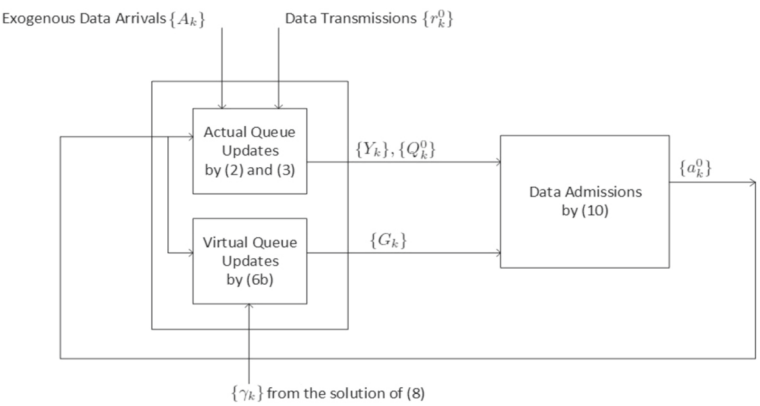

Fig. 2 The diagram of the inputs and output of data admission procedure

where $w_k^0(t) = Q_k^0(t), k \in \mathcal{K}_0$ (weights for the direct links from the BS to its users; recall that ρ_k^0 is always equal to 0 for these links), $w_k^m(t) = Q_k^m(t) + \rho_k^m W_e, m \in \mathcal{M}, k \in \mathcal{K}_m$ (weights for the access links from relays to their users), and $w_k^0(t) = Q_k^0(t) - Q_k^{m(k)}(t) + \rho_k^0 W_e, \forall k \in \mathcal{Q}_0 - \mathcal{K}_0$ (weights for the feeder links of indirect users from the BS to relays). The differential backlog term $Q_k^0(t) - Q_k^{m(k)}(t)$ in the weight of a feeder link is resulted by switching the sums and considering the fact that for the buffers of the relays, the arrivals are upper bounded by the transmission rates from the BS to relays, i.e., $a_k^m(t) \leq r_k^0(t), \forall m \in \mathcal{M}, k \in \mathcal{Q}_m$. As explained in the previous subsection, and later in Remark 1, these differential backlog terms lead to unfair data admission for indirect users, and their effect can be reduced by using $\rho_k^m W_e$ terms.

We note that SP3 is a mixed integer and nonlinear programming and needs an exhaustive search to find its optimum solution. One common approach for these types of problems is to relax the subchannel allocation variables x_{kn}^m whenever this relaxation converts the problem into a convex one. Then, using the dual decomposition, optimal solution can be found, if the duality gap is zero. This approach was used in our previous work in [20], based on which we proposed a dynamic distributed resource allocation. However, in the current paper, due to the finite data and limited buffer capacity constraints, i.e., $C2$ and $C3$, the resulted problem after relaxation of x_{kn}^m variables will be non-convex. In addition, we note that in [20], there was only one power constraint for the whole system which made it possible to have high convergence speed for the proposed algorithm. In a more realistic system like the one we consider in the current paper, the BS and relays have separate power constraints. Therefore, even if we remove the constraints $C2, C3$ and relax x_{kn}^m variables to make it convex, a dual-based iterative algorithm will need many iterations and a long time to converge. This is not suitable for use in practical scenarios where each time slot is in the order of a millisecond, and the resource allocation decision needs to be made in a small fraction of time.

Due to the abovementioned, we aim at low-complexity suboptimal algorithms which can be easily implemented in practical systems. For this purpose, we consider equal power allocation on subchannels and allocate them in a greedy way, based on the queue sizes and achievable transmission rates of the links. This is not only for making the resource allocation practical, but it is also reasonable, because when adaptive transmission rates are used (as in our system by (1)), optimal power allocation results in marginal gains [26].

However, even considering equal power distribution on subchannels and computing the achievable transmission rates of the links is not trivial here, due to the following two issues:

a) *Unknown number of subchannels for each node.* For deciding about the allocation of subchannels, we need to know the achievable transmission rates of the links on the subchannels, and for that we need to know the power allocations on the subchannels. However, before subchannel allocation, it is not clear how many subchannels will be allocated to the BS and relays, and consequently, it is not known on how many subchannels their total powers will be distributed equally.
b) *Unknown total powers to be used by each node.* The total powers used by each of the BS and relays need to satisfy the average and peak power constraints. This is controlled in SP3 by constraint $C7$ and the objective function which is the sum of increasing and decreasing functions of power. Based on that, the total power used by each node can vary in each time slot between zero and its peak value, depending on the subchannel allocations and the size of the corresponding virtual power queue. Therefore, even if we make an assumption on the number of subchannels to be used by each node, it is not clear that how much total power will be distributed equally on them.

To address the above issues, we propose a low-complexity suboptimal strategy, as shown in subchannel and power allocation strategy (SPAS), which breaks the interdependence between power allocation and subchannel assignment. At the beginning, SPAS assumes that

Subchannel and power allocation strategy (SPAS)

- Assume the number of subchannels that node m will use for its transmissions, $\tilde{N}_m$, will be proportional to the number of its queues.
- Assume that each node m will use its *peak* power (which will be equally distributed on its subchannels, i.e., it will use $\frac{\hat{P}_m}{\tilde{N}_m}$ for transmission on each of its subchannels).
- Estimate the achievable transmission rates of the links of the nodes based on their channel conditions and the abovementioned assumption for powers.
- Determine the type of time slot and allocate the subchannels to the system links, based on the estimated achievable transmission rates and actual queue sizes.
- Adjust the total power that node m should use, by considering the size of its actual data queues and virtual power queue; then, distribute it equally on the subchannels assigned in the previous step to the links of node m.

the number of subchannels that each node will use for its transmissions will be proportional to its number of queues, and the total power each node will distribute on its subchannels will be its peak power. Based on that, SPAS estimates the transmission rates of the links to be used in making a decision about the type of time slot and subchannel allocation. At the end, it adjusts the total powers, considering the size of actual and virtual queues. The details for these are presented in the next subsections.

Note that SPAS in fact provides a plan which can be utilized for designing the resource allocation algorithms in a centralized or distributed way. In the following, we will present the distributed implementation, as it is of more interest to the research and industrial bodies; later, we will describe the centralized resource allocation which can be used as a baseline for the proposed distributed one.

3.4 Efficient dynamic distributed resource allocation

In this subsection, we propose the EDDRA method which performs resource allocation in each time slot, through four steps. In the first step, every node reports estimates of its subchannel demands to the BS and based on them, the BS decides about the type of time slot. In the second step, the BS determines and reports the subchannel sets that each of the BS and relays can use. Then, in the third and fourth steps, in a distributed way, each node first assigns the subchannels to its users and then adjusts the total power it can distribute over its subchannels.

Step 1) Slot Type Determination (STD). At the end of each time slot, first, the BS needs to specify the type of the next slot. For this, relays report an estimate of their average demand for each subchannel to the BS. These demands are computed based on the assumptions on the number of subchannels they can get and the total power they can use. Then, the BS uses the reported demands from the relays as well as its own demands to estimate the system's total demands for type A and type B slots, and based on that, it decides about the type of the next time slot. This is outlined in STD algorithm and the details are described in the following.

Based on SP3, we define the estimated average demand of node $m \in \mathcal{M}$ on subchannel n as

$$D_n^m = \frac{1}{|\mathcal{K}_m|} \sum_{k \in \mathcal{K}_m} w_k^m \tilde{r}_{kn}^m, \forall m \in \mathcal{M}, \forall n \in \mathcal{N} \tag{12}$$

where $\tilde{r}_{kn}^m = \log_2\left(1 + \frac{\hat{P}_m e_{kn}^m}{\tilde{N}_m \Gamma_k}\right)$ is the estimated transmission rate of the link of user k from node m on subchannel n. It is computed in node m, $m \in \mathcal{M}$, assuming that the number of subchannels it will get, $\tilde{N}_m$, is proportional to the ratio of its number of queues ($|\mathcal{K}_m|$) and the total number of queues that can be considered for service in type B slot $\left(\sum_{m \in \mathcal{B}} |K_m| = K\right)$; i.e.,

$$\tilde{N}_m = N \frac{|\mathcal{K}_m|}{K}, \forall m \in \mathcal{B}, \tag{13}$$

Since the BS knows the number of queues in each relay, it can easily estimate their total demands as in line 2 of STD algorithm. For itself, the BS needs to compute separate demands for type A and type B slots. Noting that in type B slots, it can only transmit to its direct users while sharing subchannels with relays, its average demands are computed similar to relays and based on the weights and rates of the links of direct users (assuming the transmission power on each subchannel equal to $\frac{\hat{P}_0}{\tilde{N}_0}, \tilde{N}_0 = N\frac{|\mathcal{K}_0|}{K}$), i.e., $D_n^{0b} = \frac{1}{|\mathcal{K}_0|} \sum_{k \in \mathcal{K}_0} w_k^0 \log_2\left(1 + \frac{\hat{P}_0 e_{kn}^0}{\tilde{N}_0 \Gamma_k}\right)$.

Algorithm 1 Slot Type Determination (STD)

1: Each relay $m \in \mathcal{M}$ reports to the BS, its estimated average demands on all subchannels, i.e., D_n^m, $\forall n \in \mathcal{N}$.
2: The BS estimates the total demand of each relay m as $D^m = |\mathcal{K}_m| \sum_{n=1}^{N} D_n^m, m \in \mathcal{M}$
3: The BS estimates its own demand for type A slot as $D^{0a} = K \sum_{n=1}^{N} D_n^{0a}$
4: The BS estimates its own demand for type B slot as $D^{0b} = |\mathcal{K}_0| \sum_{n=1}^{N} D_n^{0b}$
5: The BS estimates the total demand for type A slot as $D^A = D^{0a}$, and for type B slot as $D^B = D^{0b} + \sum_{m=1}^{M} D^m$
6: **if** $D^A > D^B$
7: The BS sets the type of slot to A
8: **else**
9: The BS sets the type of slot to B
10: **end if**

On the other hand, we note that in a type A slot, only the BS can transmit and *all* the queues in the BS (including those of indirect users) can be served using *all* the subchannels; thus, its total demand is computed based on the weights and rates of all of its links and assuming $\frac{\hat{P}_0}{N}$ power on each subchannel, i.e., $D_n^{0a} = \frac{1}{K} \sum_{k \in \mathcal{K}} w_k^0 \log_2\left(1 + \frac{\hat{P}_0 e_{kn}^0}{N \Gamma_k}\right)$.

Note that for computing the demands in the BS, it needs to know the queue sizes of the relays as well (to be used in the weights of the feeder links from the BS to relays). For this purpose, relays also report the information about their modified queue sizes, to the BS. Considering the fact that in each time slot, at most N different queues can be served, the maximum number of modified queue sizes in

relays is $\min(K - |\mathcal{K}_0|, N)$). Therefore, in EDDRA, the total overhead of signaling about the demands and the modified queue sizes is of $O(\min(K - |\mathcal{K}_0|, N) + MN)$.

Remark 1. Here, we explain the reason for using $\rho_k^m W_e$ in the weights of the links of indirect users from the BS and relays. Without that, due to the low powers of relays and low transmission rates, their demands would not be comparable to the demands of the BS for direct users, unless the queues in relays grew large. On the other hand, for the links serving the queues of indirect users in the BS, we would have $w_k^0 = Q_k^0 - Q_k^{m(k)}, \forall k \in \mathcal{K} - \mathcal{K}_0$. As a result, these would not have enough impact on computing the average demands for indirect users and providing service for them (in the cases that the queue sizes of an indirect user in the BS and relay have the same size, $Q_k^0 = Q_k^{m(k)}$, the impact would be zero). Consequently, the queues of indirect users in the BS would usually have larger sizes than the queues of direct users, and therefore, data admission for them would be less. This would degrade the usefulness of drift-plus-penalty for cellular networks, because fairness is usually one of the main concerns in these networks. To prevent that, ρ_k^m should be set to 1 for the feeder and access links and W_e should be applied as mentioned in subsection 3.1. This will compensate for the effect of low power of relays on the demands of access links and the effect of differential-backlog-based weights on the demands of feeder links. Similar effect holds also in the subchannel sets determination and subchannel allocation steps which will be described later.

Step 2) Subchannel Sets Determination (SSD). We note that in a type B slot, due to sharing subchannels among all the nodes, the resource allocation for the links of different nodes are tied together which is reflected in (11). In this step, the goal is to break this tie and specify the subchannel sets to be used for transmissions from the BS and relays. This allows to have the resource allocation in a distributed manner at each node.

For the above purpose, when the slot is decided to be type A, the BS notifies the relays about it and they know they have no transmissions. In the case of a type B slot, the BS determines the subchannel sets of the relays and notifies them to transmit on them. SSD algorithm shows the whole procedure in detail. Since in type B slots, the BS can only transmit to the direct users, line 5 of the algorithm defines its demands based on the estimations for type B slot, explained before. Line 6 sets $\hat{N}_m$, as the upper bound for the number of the subchannels that each node can get, and the next lines assign the subchannels to the nodes that have not reached their limit on the number of subchannels and have higher average demands on the subchannels.

Algorithm 2 Subchannel Sets Determination (SSD)

1: **if** the slot type is A
2: The BS determines subchannel sets as $\mathcal{N}_0 = \mathcal{N}$ and $\mathcal{N}_m = \emptyset, m \in \mathcal{M}$
3: **else**
4: The BS specifies subchannel sets, based on the relays' demands as well as its own, as follows
5: Set $D_n^0 = D_n^{0b}$
6: Set $\hat{N}_m = \left\lceil N \frac{|\mathcal{K}_m| \sum_{n=1}^{N} D_n^m}{\sum_{m=0}^{M} \sum_{n=1}^{N} |\mathcal{K}_m| D_n^m} \right\rceil, \forall m \in \mathcal{B}$
7: Initialize $\mathcal{N}' = \mathcal{N}, \mathcal{B}' = \mathcal{B}, \mathcal{N}_m = \emptyset, \forall m \in \mathcal{B}'$
8: **while** $\mathcal{N}' \neq \emptyset$ and $\mathcal{B}' \neq \emptyset$
9: find $(m^*, n^*) = \arg \max_{m \in \mathcal{B}', n \in \mathcal{N}'} D_n^m$
10: $\mathcal{N}_{m^*} = \mathcal{N}_{m^*} \cup \{n^*\}$
11: $\mathcal{N}' = \mathcal{N}' - \{n^*\}$
12: **if** $|\mathcal{N}_{m^*}| = \hat{N}_{m^*}$
13: $\mathcal{B}' = \mathcal{B}' - m^*$
14: **end if**
15: **end while**
16: **end if**
17: The BS notifies relays about their subchannel sets

Note that setting the limit $\hat{N}_m$ for the size of the subchannel set of node m is to prevent node m from getting subchannels more than it needs. For example, a relay might have only one user with high *average* demands on the subchannels while another relay with several users might have a little lower *average* demands on the subchannels. In such a case, without considering the total number of users and the limit for subchannel set sizes, the relay with a single user would overshadow the other relay in all the iterations of subchannel assignments through line 9.

The computational complexity of the SSD algorithm is of $O((M+1)N^2)$, which is obtained by ignoring the insignificant computations and considering the number of iterations needed for performing line 9.

Step 3) Subchannel Allocation (SA). After determining the subchannel sets, the resource allocation subproblem (11) can be further decomposed into separate *subsubproblems*, as follows:

SSP (resource allocation subsubproblems) :

$$\max_{\boldsymbol{p},\boldsymbol{x}} \sum_{n \in \mathcal{N}_m} \sum_{k \in \mathcal{Q}_m} \left(BTx_{kn}^m(t) w_k^m(t) \log_2 \left(1 + p_n^m(t) e_{kn}^m(t)\right)\right) - \sum_{n \in \mathcal{N}_m} IZ_m(t) p_n^m(t), \forall m \in \mathcal{B}, \tag{14a}$$

$$\text{s.t.} C2 - C4, C7 - C8 \tag{14b}$$

where each node knows its set of subchannels and can decide individually about allocating them to its own links, considering the related subset of the constraints $C2 - C4$,

$C7 - C8$. For this purpose, following the SPAS strategy, we propose to have subchannel allocations by each node based on using $\frac{\hat{P}_m}{|\mathcal{N}_m|}$ (i.e., assuming $Z_m(t) = 0$) for computing the achievable transmission rates. Then in the power adjustment step, considering the real value of $Z_m(t)$, each node can decide about the total power it should use and distribute it on its subchannels.

Algorithm 3 Subchannel Allocation (SA) in the BS

1: if slot type is A, set $\mathcal{Q}' = \mathcal{Q}_0$, otherwise, set $\mathcal{Q}' = \mathcal{K}_0$
2: Initialize $q_k^0 = Q_k^0, r_{kn}^0 = B\log_2\left(1 + \frac{\hat{P}_0 e_{kn}^0}{|\mathcal{N}_0|\Gamma_k}\right), k \in \mathcal{Q}', n \in \mathcal{N}_0$.
3: **while** $\mathcal{N}_0 \neq \emptyset$ and $\left(\sum_{k\in\mathcal{Q}'} q_k^0 > 0\right)$
4: Compute $w_k^0 = q_k^0, k \in \mathcal{K}_0$
5: Compute $w_k^0 = \left(q_k^0 - Q_k^{m(k)}\right), k \in \mathcal{Q}' - \mathcal{K}_0$
6: if $Q_k^0 > BT\hat{s}, w_k^0 = w_k^0 + \rho_k^0 W_e, k \in \mathcal{Q}'$
7: if $L_k^{m(k)} - q_k^{m(k)} < BT\hat{s}, w_k^0 = -1, k \in \mathcal{Q}' - \mathcal{K}_0$
8: Compute $D_{kn}^0 = w_k^0 r_{kn}^0, k \in \mathcal{Q}', n \in \mathcal{N}_0$
9: Find $(k^*, n^*) = \arg\max_{k\in\mathcal{Q}', n\in\mathcal{N}_0} D_{kn}^0$
10: $x_{k^*n^*}^0 = 1$
11: $\mathcal{N}_0 = \mathcal{N}_0 - \{n^*\}$
12: if $k^* \in \mathcal{Q}' - \mathcal{K}_0$, then $q_{k^*}^{m(k^*)} = q_{k^*}^{m(k^*)} + \min\left(q_{k^*}^0, Tr_{k^*n^*}^0\right)$
13: $q_{k^*}^0 = \max\left(q_{k^*}^0 - Tr_{k^*n^*}^0, 0\right)$
14: **end while**

Noting that the BS has more constraints than the other nodes (the constraint $C3$ is only enforced on the feeder links from the BS, which is to prevent transmitting data to the relays more than their empty buffer spaces), we provide the subchannel allocation by the BS and then explain its use for relays. SA algorithm shows the details in allocating the subchannels by the BS. The procedure is done in an iterative way with $|\mathcal{N}_0|$ iterations. In each iteration, the weights of the links and the resulted demands are computed, and the pair of subchannel and queue with the highest corresponding demand is selected. Then, the selected subchannel is assigned to the link serving the selected queue and the size of the affected queues are updated virtually. Since the actual queue sizes, Q_k^m, are only updated at the end of transmission intervals, we have used q_k^m variables to prevent ambiguity about the updating during the algorithm iterations. Note that these updates are done to meet the constraints $C2$ and $C3$. Line 6 is for applying the extra weight W_e, described before. However, before adding it, by comparing the queue size with $BT\hat{s}$, we make sure that there are enough data such that they can utilize the subchannel completely if it is allocated. Line 7 is to meet the constraint $C3$ and prevent overflow, by giving a negative weight in case the remaining empty space in a relay buffer is less than the possible maximum transmission size on a subchannel. If a feeder link gets negative weight, then it will not be considered for subchannel allocation and this will prevent transmitting data to the corresponding relay buffer.

Remark 2. Note that the rate computations in SA are based on the assumption of equal distribution of peak powers on the subchannel sets. This way, we will be sure that when in Step 4, the total power is adjusted (which certainly will be equal or less than the peak power), the transmission rate for each link will be less than the amount considered in SA algorithm, and therefore, the constraints $C2$ and $C3$ will not be violated.

In type B time slots, in parallel to the BS, any relay also uses the SA algorithm with the difference that all the superscripts/subscript 0 are replaced by the corresponding m. Note that in this case we have $\mathcal{Q}' = \mathcal{K}_m$ and $\mathcal{Q}' - \mathcal{K}_m = \emptyset$. Hence, the lines 5, 7, 12 are not executed when SA algorithm is used by relays. Based on the abovementioned, the subchannel allocation task in EDDRA is split among the serving nodes, where the computational complexity of the SA algorithm in any node $m \in \mathcal{B}$ is of $O(|\mathcal{N}_m|^2|\mathcal{Q}_m|)$.

Step 4) Total Power Adjustment (TPA). After assigning the subchannels to the links, the BS and relays decide about the total power that they can distribute on their subchannels to meet the constraints $C4, C7$. For this, based on SSP in (14), each node m solves the following problem, which we refer to as the total power adjustment, to find the total power, P_m, that it can use.

TPA (total power adjustment problems) :

$$\max_{P_m} \sum_{n\in\mathcal{N}_m} \left(BTw_{k(n)}^m(t)\log_2\left(1 + \frac{P_m(t)e_{k(n)n}^m(t)}{|\mathcal{N}_m|\Gamma_k}\right)\right) - IZ_m(t)P_m(t), \forall m \in \mathcal{B} \quad (15a)$$

$$\text{s.t. } 0 \leq P_m(t) \leq \hat{P}_m \quad (15b)$$

In the above, $k(n)$ indicates the index of the user, to the queue of which the subchannel n has been allocated. The TPA problem is a convex problem with one variable. Thus, the optimal value, P_m^*, can be found easily by using an iterative one-dimensional search such as the Golden Section method [27, Appendix C.3], which has the computational complexity of $O(\log(1/\epsilon_1))$, where ϵ_1 is the desired relative error bound.

Remark 3. As explained before, the constraint $C1$ is enforced over time through the virtual queues, $\{Z_m\}$,

defined for that purpose. In fact, based on (6a), having a nonzero Z_m means that in the past time slots, there have been the events of transmission with the total power, P_m, larger than the average power limit, P_m^{av}. Therefore, in TPA problem, Z_m applies a kind of negative feedback to use less power than $\hat{P}_m$. The proposed importance factor I is in fact for amplifying this negative feedback to adjust the total power use in a short period of time. Without it, the second term in the objective (15a) would not be comparable to the first one over a large period of time slots, before Z_m becomes big enough to impact the objective value. This is due to the fact that the values of the power variables are very small (in the order of 1–10 watts) compared to the values of queue sizes multiplied by transmission rates (in the order of tens of megabits).

After finding P_m^* as described above, each node computes the power on its subchannels, considering equal power distribution and noting that the rate on each subchannel can not be larger than $B\hat{s}$ (due to the limited spectral efficiency of modulation schemes in practice); i.e.,

$$p_n^m = \min\left(\frac{P_m^*}{|\mathcal{N}_m|}, \frac{(2^{\hat{s}}-1)\Gamma_{k(n)}}{e_{k(n)n}^m}\right), \forall m \in \mathcal{B}, n \in \mathcal{N}_m \tag{16}$$

The reason for considering the term with $\hat{s}$ in (16) is to prevent using the power more than needed for maintaining the desired bit error rate. It is obtained based on (1) and *C*4.

After the above steps, based on the variables x_{kn}^m and p_n^m, each node notifies its users about the subchannel allocations and the assigned transmission rates. Then, it transmits to them and updates its actual and virtual queues using (2) and (6).

Remark 4. As discussed in subsection 3.3, the resource allocation subproblem (11) is not a convex optimization problem; therefore, the existence of global optimal is unknown. The algorithms proposed in the STD, SSD, SA, and TPA steps provide a suboptimal solution which have low overhead and low computational complexity, and, as shown in Section 4 (Figs. 5, 6, 7, 8, 9, and 10), lead to better performance compared with an existing suboptimal algorithm. Moreover, to the best of our knowledge, even when the constraints *C*3 and *C*6 are not imposed (i.e., assuming infinite buffer capacities), there is not a clear method to compute mathematically the distance of our proposed scheme from the optimal solution of the corresponding convex optimization problem, as it is heuristic. However, as shown in Section 4 (Figs. 13, 14, 15, and 16), depending on the value of V in this case, our proposed algorithms can lead to higher or slightly lower system utility (at the cost of higher queue sizes) compared with an existing algorithm which uses subgradient method to reach the optimal solution.

Remark 5. Note that the Steps 1 to 4 are executed only once in each time slot. Also, STD, SSD, and SA algorithms have a fixed number of iterations/operations after which they terminate, and there is no need to analyze their convergence. On the other hand, as stated in Step 4, TPA problem is a single-variable convex optimization problem and therefore, any one-dimensional search is guaranteed to terminate when reaching the specified tolerance of ϵ_1.

3.5 Efficient dynamic centralized resource allocation (EDCRA)

In this subsection, we briefly describe the EDCRA method, in which the BS performs all the procedures for resource allocation. In a centralized scheme, the BS needs to get notified about the channel states of all the links in the system over all the subchannels[1]. For this purpose, since the indirect users do not have connection to the BS, the relays report to the BS about the channel conditions of the access links (which already have been reported by the indirect users to their serving relays). This imposes a signaling overhead of $O((K-|\mathcal{K}_0|)N)$ from relays to the BS. Considering the fact that in practice the number of users is remarkably more than the number of the relays, the signaling overhead in EDCRA is a lot more compared with EDDRA[2] (which is of $O(\min(K-|\mathcal{K}_0|, N)+MN)$).

Having all the information about the channel states and the queue sizes, the BS performs STD procedure and if the slot type is set to A, it uses the SA algorithm as in EDDRA. However, if the slot type is set to B, the BS does not need to run SSD algorithm. Instead, considering the set of all the subchannels, it uses SA algorithm as follows. The queues in relays are assumed to be located in the BS, and their corresponding access links are assumed as direct links starting from the BS to the indirect users; however, the weights and channel rates are considered the same as those of actual access links. Then, SA algorithm is exploited to decide about the subchannel allocation to the different links in the system, which incurs the computational complexity of $O\left(N^2\left(\sum_{m\in\mathcal{B}}|\mathcal{K}_m|\right)\right) = O(N^2K)$. After that, based on the corresponding subchannel allocations for all the nodes, i.e., $\{x_{kn}^m\}$, the BS specifies the powers to be used by each node by performing the total power adjustments for them. Finally, the BS informs all the relays about the subchannels and powers they can use.

Remark 6. We note that the algorithms proposed in this section do not consider any strict quality of service (QoS) requirement such as packet delay thresholds for the users. Therefore, they are mostly suitable for the data admission and resource allocation in the case of best effort services. However, the heuristics used here can provide insights

for future works to design data admission and resource allocation algorithms in the presence of the services with strict QoS requirements.

4 Performance evaluation and discussion

To evaluate the performance of the proposed algorithms, we have conducted extensive simulations for a system with M=6 relays, which are located in the distance of 2/3 of the cell radius from the BS and in an equal angular distance from each other. The simulation parameters are as in Table 3, unless otherwise specified. For the links from the BS to relays, channels are assumed line-of-sight (LOS)-based and therefore, Rician channel model, with κ factor equal to 6 dB, is considered; for the links between the BS/relays and users, channels are assumed non-LOS (NLOS)-based and therefore, Rayleigh channel model is used [28]. For computing the path loss of the links, we have considered the COST231 Hata urban propagation model which uses the following equation [29]:

$$\begin{aligned} PL = {} & (44.9 - 6.55\log_{10}(h_{tx}))\log_{10}\left(\frac{d}{1000}\right) + 45.5 \\ & + (35.46 - 1.1h_{rx})\log_{10}(f_c) - 13.82\log_{10}(h_{tx}) \\ & + 0.7h_{rx} + 3, \end{aligned} \tag{17}$$

where PL is the path loss in dB, h_{tx} is the transmitter antenna height in meters, h_{rx} is the receiver antenna height in meters, f_c is the carrier frequency in MHz, and d is the distance between the transmitter and receiver in meters. The above model for path loss has been considered in [29] for urban Macrocell environment, where the distance between the BS of the adjacent cells is larger than 1 km. Due to the fact that using relays in cellular networks makes it possible to have a large coverage area for a single cell, served through the BS and relays, we have assumed the network environment as urban macrocell and have used the above equation. We acknowledge the fact that, in reality, the channel behaviors in cellular networks might be different from the models considered in our simulations. However, considering the fact that the same models have been used in simulating the behavior of the baseline algorithms existing in the literature, the relative performance improvements of our proposed algorithms, presented later, are expected to hold.

Table 3 Simulation parameters

Parameter name	Setting
Cell radius	1000 m
BS antenna height	15 m
Relay antenna height	10 m
User antenna height	1.5 m
Carrier frequency	2500 MHz
Subhannel bandwidth	180 KHz
Time slot duration	1 ms
Noise power spectral density	−174 dBm/Hz
BER requirement	10^{-6}
Traffic model	Poisson
Packet size	5 KBits

For utility function, we have considered $\mathcal{U}(z) = \log(z)$ to provide proportional fairness. Due to the large packet sizes which resulted in large queue sizes, based on the observations from simulation results, we have chosen V to be 10^7. This gives high value for utility function in (7a) to be comparable to the terms related to the weighted transmission rates. The buffer capacities at the BS and relays are considered equal to 100 and 10 packet sizes, respectively. The highest order for modulation is assumed to be 64 QAM which has the spectral efficiency of 6 bits/sec/Hz.

In the following, we first consider a special scenario with the settings as follows. The data arrival rate of each user is 100 packets per second, or equivalently 500 kbps. The peak power of the BS is equal to $\hat{P}_0 = 34$ dBm and the peak power of the relays are equal to $\hat{P}_m = 25$ dBm, $m \in \mathcal{M}$. The average power constraint of the nodes are half of their peak power constraints, i.e., 31 dBm = 1259 mW for BS and 22 dBm = 158 mW for relays. The number of subchannels is considered equal to $N = 7$. There are 12 users in the system, 6 of them connected directly to the BS and the rest connected to relays, one user per relay. The distances of the direct users from the BS and the indirect users from the corresponding relays are 300 m. Note that this special scenario, with the mentioned settings, is to provide an example to show the necessity of considering the importance parameter I in practical scenarios. The values of the different parameters are selected specifically for simulation purpose. However, in practice, these values are possible. For instance, the indirect users' location are considered 300 m from relays to simulate the case that those users are close to the cell edge, and the direct users location is selected 300 m from the BS to simulate an average location between the BS and relays. On the other hand, the number of subchannels is selected equal to seven to have, on average, one subchannel for each of the serving nodes (the BS and six relays). Even though seven subchannels might not correspond to an explicit standard bandwidth, it can happen in practical systems. For example, the operators might allocate the system resources separately for different classes of users and reserve specific number of subchannels for each class of users, using the number of serving nodes as a rule of thumb.

Figure 3 shows the average power consumed by each node, over 20,000 time slots, with different values for importance factor I, and Fig. 4 depicts the virtual queue

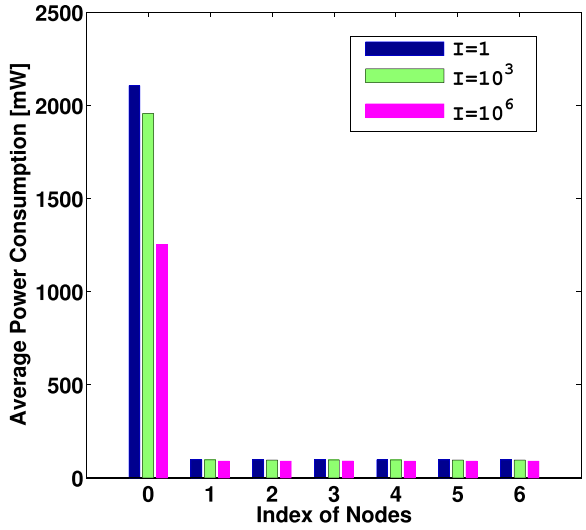

Fig. 3 Effect of parameter *I* on average power consumption of the BS and relays; $N = 7, |\mathcal{K}_0| = 6, |\mathcal{K}_m| = 1, m \in \mathcal{M}$

size corresponding to the average power constraint of the BS during the mentioned period. It is observed that without considering a suitable I (e.g., when $I = 1$ or $I = 10^3$), the average power consumption of the BS is about 2000 mW, which is far beyond the constraint of 1259 mW, and the size of virtual power queue of the BS grows constantly over this period. This happens due to the fact that in (15a), the value of the second term is very small compared with the value of the first one and as a result, it does not have much effect on the optimization objective; therefore, the only thing that limits the total power usage is the peak power or the maximum spectral efficiency. The consequence of this is the steady use of the peak power of the BS in each time slot, which incurs the steady growth of its virtual power queue size according to (6a). Without a suitable I, this would continue for a long time, until the size of the virtual queue has grown so large that the second term in (15a) is comparable to the first one. However, as stated in Remark 3, using a suitably large I (e.g., $I = 10^6$ in this scenario) amplifies the effect of the virtual power queue sizes in (15a) and prevents continuous use of the peak powers. Therefore, as shown in Fig. 4, the virtual power queue of the BS gets bounded after about 300 time slots and, as displayed in Fig. 3, the average power consumption of the BS over the simulation period is about the specified constraint. It is worth mentioning that due to fewer transmissions from the relays, compared with the BS, their virtual power queues did not grow large and remained stable in all the above values for I and had similar graphs as that of the BS in the case of $I = 10^6$. Because of this similarity, their graphs were omitted.

To investigate the overall performance of the proposed algorithms in general scenarios, we consider a system with 25 users, which are uniformly distributed in the cell area and are connected to the node from which they receive higher signal strength. The simulations are conducted for

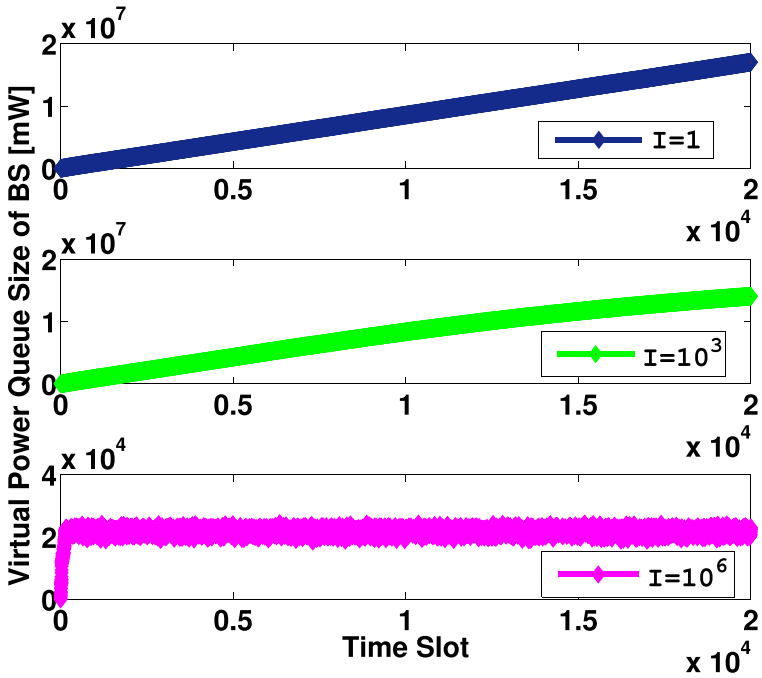

Fig. 4 Effect of parameter *I* on virtual power queue size of the BS over time; $N = 7, |\mathcal{K}_0| = 6, |\mathcal{K}_m| = 1, m \in \mathcal{M}$

100 runs, each over 10,000 time slots, to generate different realizations of users' locations. All the users have the data arrival rate of 20 packets per second or equivalently 100 kbps, and the buffer capacity in the BS and relays are, respectively, 100 and 10 packets per user. There are 14 subchannels in the system, the BS's peak power is 37 dBm, relays' peak powers are 28 dBm, and the average power constraints are half of the peak powers. As a baseline, we have adapted the low-complexity centralized algorithm proposed in [19] to our system model, which we refer to as fixed half-duplex relaying (FHDR) in the figures. With FHDR, the type of the time slots are fixed such that the odd-numbered time slots are used for the transmissions from the BS and the even-numbered slots for the transmissions only from the relays. The subchannel allocations in even-numbered slots are based on considering a minimum of $\lfloor N/M \rfloor$ subchannels for each relay and assigning them based on the Hungarian algorithm. For FHDR, the average power limit of each node is equally distributed over all the subchannels, considering the maximum spectral efficiency constraint. Also, we have implemented the data traffic control procedure in the FHDR to compare its utility with that of the proposed schemes.

We note that due to limited buffer capacities, all the queues are stable and their sizes are less than the corresponding buffer capacities. Therefore, in the following, we do not present any results about them and instead study the overflow performance. Figure 5 displays the cumulative distribution function (CDF) of the system utility. It is observed that all the algorithms have the same utility of data admissions. This indicates that all of them lead to similar amount of transmissions from the BS, and therefore, similar queue sizes in it which allow similar data admissions. However, as shown in Fig. 6, the total overflow from the buffers of relays is zero with the proposed EDCRA and EDDRA schemes whereas FHDR has the incidents of overflow. This is due to the fact that FHDR does not take into account the limited buffer capacities of the relays when deciding about the subchannel allocations. On the contrary, in the case of insufficient free space in a relay buffer, the proposed schemes do not allocate any subchannel to the corresponding link from the BS and this way, they prevent data transmissions to that buffer.

We have also presented the system average throughput, received by the users, in Fig. 7. It is observed that the EDCRA and EDDRA result in higher throughput than

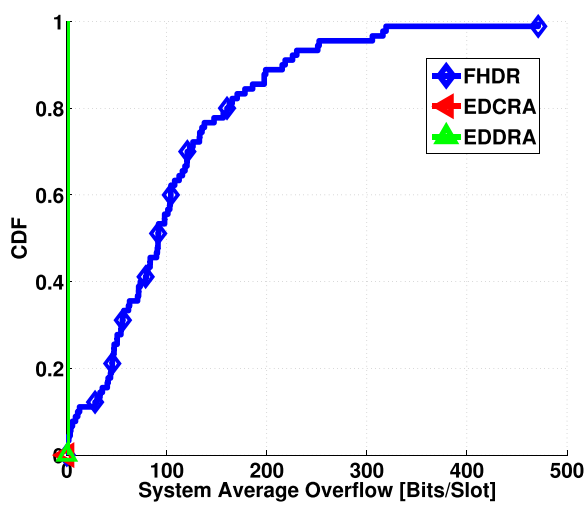

Fig. 6 CDF of system average overflow at the data arrival rate of 20 packets/s

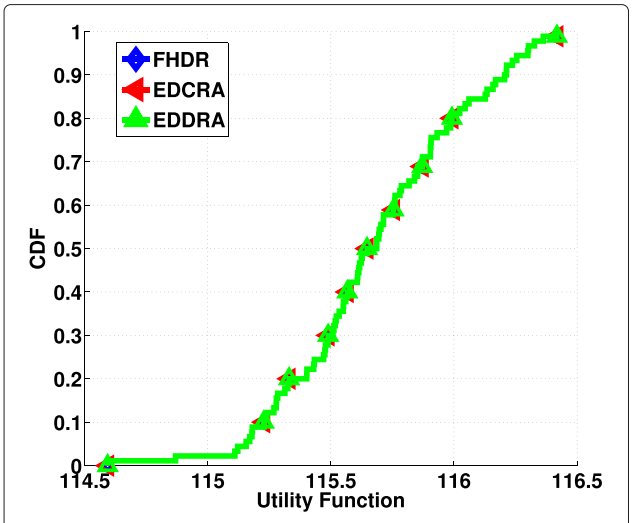

Fig. 5 CDF of system utility at the data arrival rate of 20 packets/s

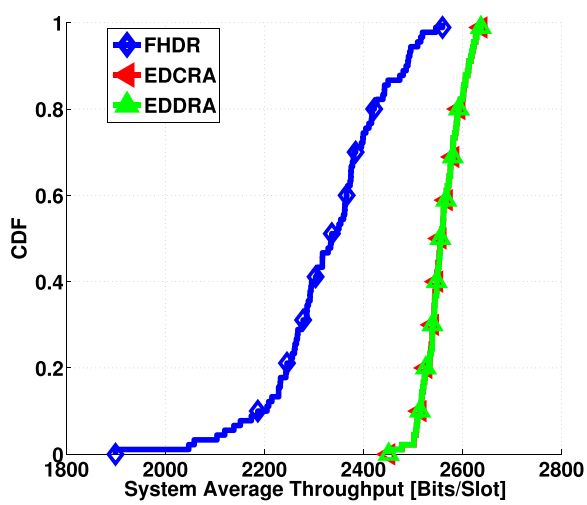

Fig. 7 CDF of system average throughput at the data arrival rate of 20 packets/s

the FHDR, which is a sign of efficient utilization of system resources (i.e., time slot, subchannel and power) by the proposed schemes. This happens due to several reasons. In FHDR, the transmission power of the subchannels are constant; the time intervals for the transmissions from the BS and relays are fixed and are done in the odd-numbered and even-numbered time slots, respectively. Also, FHDR considers a fixed minimum number of subchannels for each node in each time slot, which reduces the flexibility in allocating the subchannels. On the other hand, EDCRA and EDDRA decide about the type of the time slots and the subchannel allocations in an adaptive way and based on the demands of the different nodes, where the demands are based on achievable transmission rates and the actual queue sizes. Moreover, after subchannel allocations, EDCRA and EDDRA adjust the power usage on the subchannels depending on the virtual power queue sizes, peak power, and the maximum spectral efficiency constraints, which allow them to utilize the flexibility in power allocation.

Next, in the same system, we investigate the effect of increase in the data arrival rate on the performance of the proposed algorithms. Figures 8, 9, and 10 show, respectively, the system utility, system average overflow, and system average throughput versus data arrival rate. It is observed that as the data arrival rates increase, the utilities of the EDDRA and EDCRA also increase; but after the arrival rate of 60 packets/s, this increase is not much. This is firstly due to the fact that the utility function is a concave function which has diminishing returns as the arrival rates increase. The other reason is the fact that the system capacity is saturated in high packet arrival rates, leading to lower increase in throughput as shown in Fig. 10; therefore, the queues can not be served much and this prevents more admission of data into the system. Similar effect is observed about FHDR; however, since it does not determine the type of time slot based on the demands and does not use the subchannel and power resources efficiently, it leads to higher queue sizes in the BS and consequently, lower data admissions in high packet arrival rates. Due to this and the overflows from the relays' buffers, it also results in lower throughput.

Note that the performance of the EDCRA is a little better than that of EDDRA. This is due to the fact that in EDCRA, the BS has information about the channel states

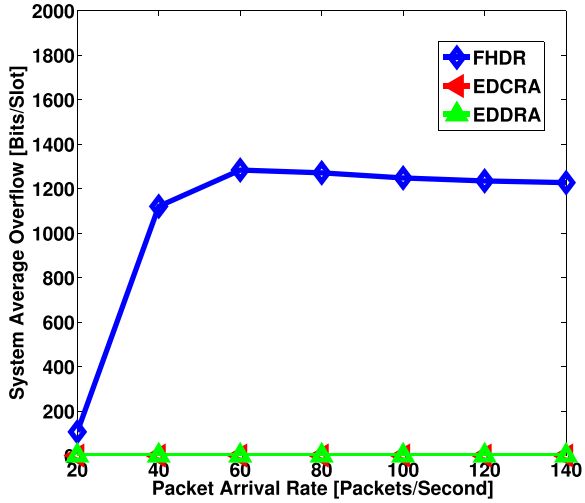

Fig. 9 Effect of increase in the packet arrival rate on the average overflow from the buffers of relays

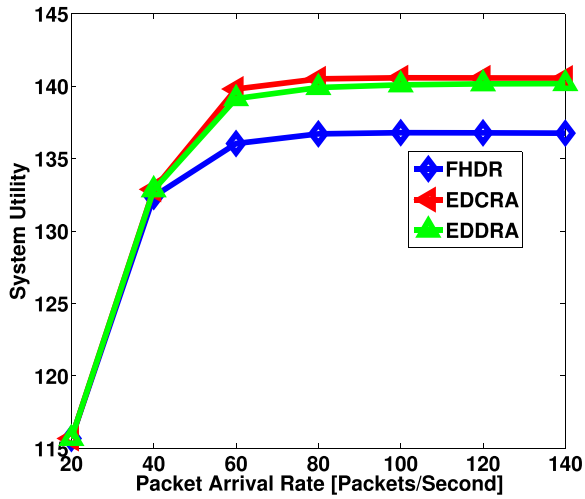

Fig. 8 Effect of increase in the packet arrival rate on the system utility

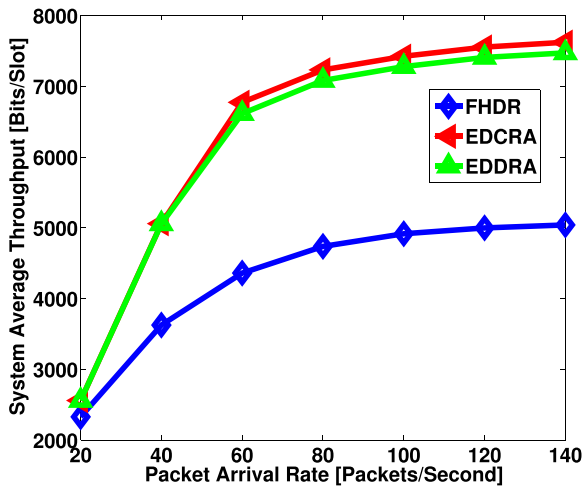

Fig. 10 Effect of increase in the packet arrival rate on the system average throughput

of all the links and performs subchannel allocation based on the individual demands of the relays' queues and its own queues; but in EDDRA, the BS determines the subchannel sets of the nodes based on their *average* demands and then each set of subchannels is only used to serve the set of the queues of the corresponding node. However, the degradation in the performance of EDDRA is not significant. The reason for this is the fact that there are usually several users for each node, with different channel conditions, that make it possible for each node to utilize its set of subchannels efficiently. Also considering an upper bound for the number of subchannels each node can get prevents wasting the resources. These observations show that using EDDRA, the system can allocate resources efficiently with less computations at the BS, low signaling overhead, and without remarkable reduction in the system performance.

Next, in order to show the effectiveness of the parameters ρ_k^m and W_e proposed in this paper, we show the data admissions for direct and indirect users. It is observed in Fig. 11 that as the packet arrival rates of the users increase, the data admissions for direct users increase but less data are admitted for indirect ones. As explained in Remark 1, this is due to the fact that the queues of indirect users at the BS get low weights and grow large, which result in lower data admissions for them. To prevent that, ρ_k^m can be set to one to apply an extra weight of W_e in the weights of the links from the BS to relays and the links from the relays to users, to increase the service provisioning for their corresponding queues and consequently lead to more data admissions into the corresponding buffers in the BS. Figure 12 shows this in the case of arrival rates equal to 140 packets per second for every user. It is observed that giving higher weights can increase the data admissions for indirect users; this comes at the cost of large drops in the data admissions of direct users. It is due to the fact that adding the extra weights results in the increase of subchannel allocations to the queues of indirect users in the BS and relays. Since the BS has higher power than relays, the less subchannel allocation to the direct users means that their queues loose higher transmission rates than those for the queues in relays. As a result, the queues of direct users in the BS grow quicker and limit the data admissions.

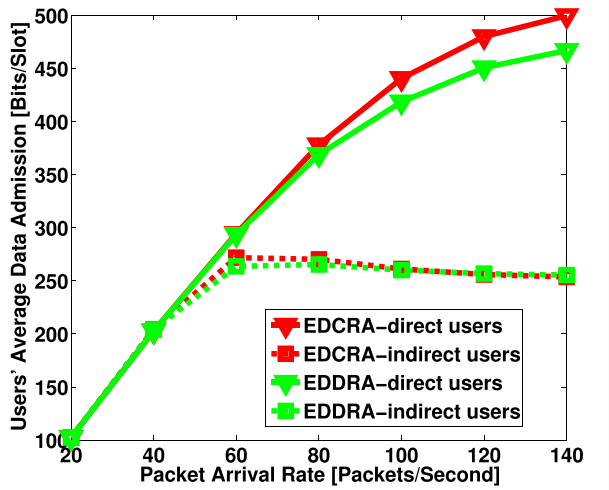

Fig. 11 Effect of increase in the packet arrival rate on the amount of data admitted for direct and indirect users

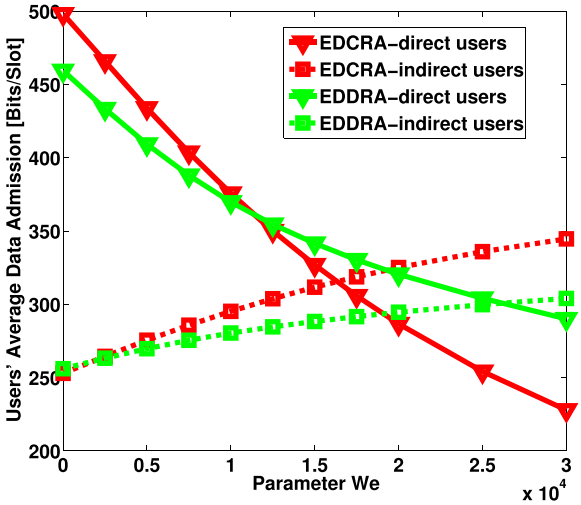

Fig. 12 Effect of parameter W_e on providing fair data admissions for direct and indirect users at the arrival rate of 140 packets/s for every user

Finally, we consider a scenario in which there is no limit on the transmission rates and buffer capacities (i.e., $C2-C4$ and $C6$ are removed from problem (4)), and actual queues need to be stabilized (i.e., remain bounded). In this case, as shown in [21], the resource allocation subproblem can be formulated as a convex optimization problem and therefore, a global optimal exists. In [21], the JFRP algorithm is proposed which uses subgradient-based iterations to get close to the optimal solution for resource allocation. To compare our proposed schemes with JFRP, we have run simulations for EDDRA and EDCRA with the settings considered in [21], over 10,000 time slots. The results are shown in Figs. 13, 14, 15, and 16, in which the graphs of JFRP are imported from the corresponding figures in [21].

In Figs. 13 and 14, the system utility and total queue size are shown versus mean data arrival rate, when $V = 60$. It is observed that EDCRA and EDDRA have almost the same or even higher utility than JFRP, which is a sign of more data admissions to the network. On the other hand, the proposed schemes lead to larger queue sizes in most of the arrival rates. Figures 15 and 16 show the system utility

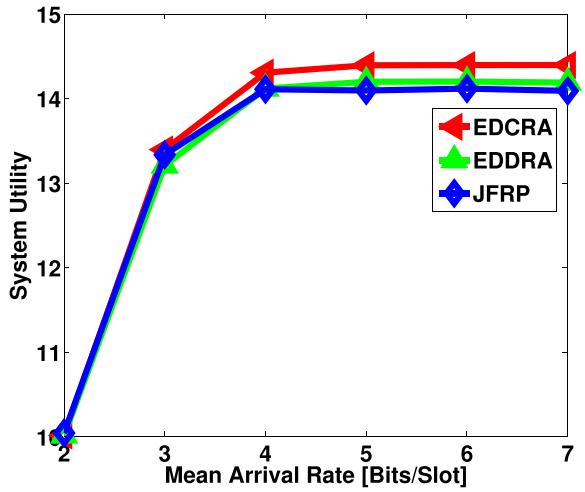

Fig. 13 Effect of increase in the data arrival rate on the system utility in the settings of [21]

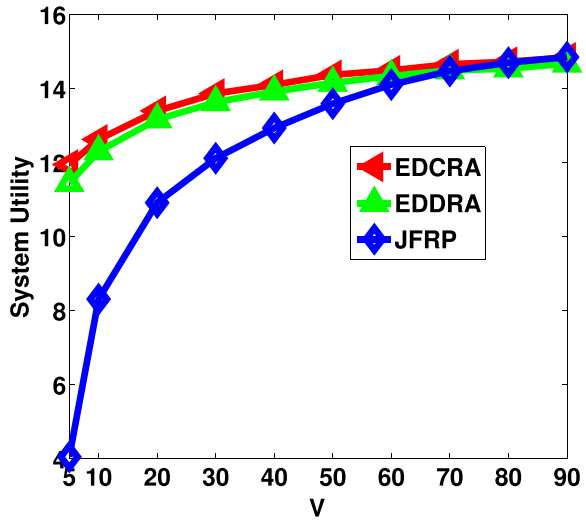

Fig. 15 Effect of value parameter V on the system utility in the settings of [21]

and total queue size versus V. It is observed that when $V < 70$, the proposed schemes outperform JFRP in terms of system utility, even though they result in larger queue sizes in almost all the values of V. It is stated in [21] that the lower utility of the JFRP in smaller values of parameter V is because the power is under-utilized. Also, we note that even though the JFRP algorithm tries to get a solution close to the optimal one, it leads to a suboptimal solution which can be due to the approximations through the iterations of subgradient method. Moreover, note that in these results, according to the settings of [21], infinite buffer capacities were assumed for the BS and relays, and therefore, instead of overflow, we have shown the results for queue sizes. It is observed that all the algorithms keep the queues stable which means that in long term, the data reception rates at the users are equal to the data admission rates into their buffers at the BS; thus, in this scenario, the system utility performance also indicates the throughput performance. However, in the scenarios with limited buffer capacities, JFRP can result in overflows and lower throughput similar to FHDR in Figs. 6, 7, 9, and 10. Also based on the Figs. 3, 4, 11, and 12, we note that since JFRP does not consider the parameters I, ρ_k^m and W_e, it can lead to unsatisfactory performance for the average power

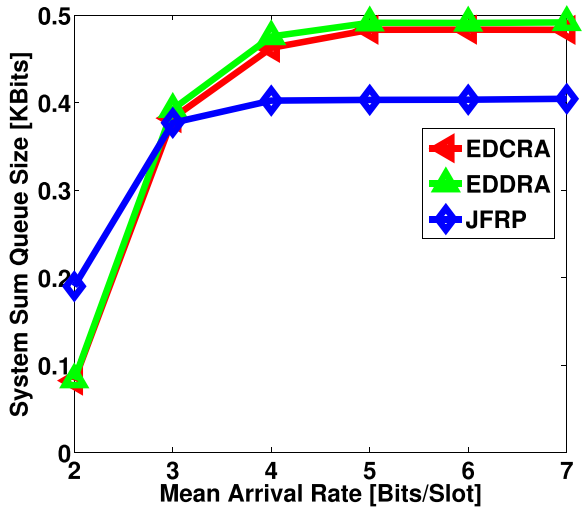

Fig. 14 Effect of increase in the data arrival rate on the system sum queue size in the settings of [21]

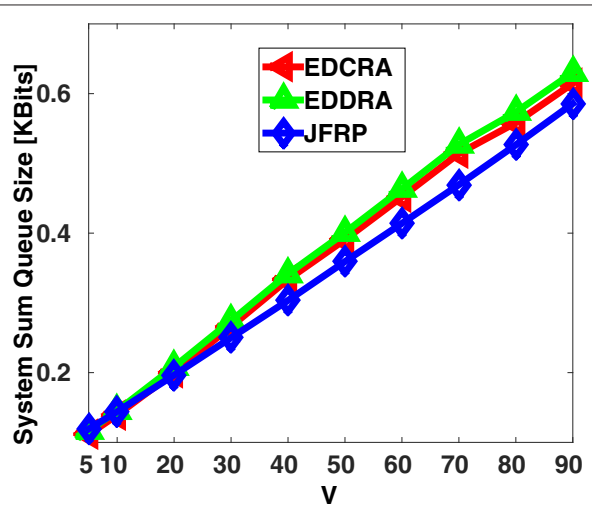

Fig. 16 Effect of value parameter V on the system sum queue size in the settings of [21]

constraints and the fair data admission for the users in the settings of practical systems.

In order to compare the computational complexities of JFRP and the proposed algorithms, we first note that the subgradient method requires number of iterations in the order of $O(1/\epsilon_2^2)$ to reach the ϵ_2 error bound [30]. Also we note that in each iteration, the dual variables used for updating the primal ones are not optimal. This can lead to ties in computing the subchannel allocation indicators during the iterations. Depending on the methods and approximations used for breaking the ties, JFRP might require even more iterations. This prohibits distributed implementation of JFRP, due to the messaging overheads imposed for updating the dual and primal variables in every iteration. Also, it is very expensive in terms of time it consumes, considering the fact that the iterations should be done for each time slot which is in the order of millisecond in practical systems. Considering the abovementioned and the main computations in each iteration, the overall computational complexity of JFRP algorithm is at least in the order of $O\left(KN/\epsilon_2^2\right)$. On the other hand, the EDDRA algorithm requires only one-time messaging from relays to the BS (about their queue sizes and demands) and one-time messaging from the BS to relays (about subchannel sets), and the rest of the processing is done locally in each node. Furthermore, the overall computational complexity of EDDRA is clear and is of $O(KN^2 + (M+1)\log(1/\epsilon_1))$ which is split among the serving nodes. Similarly, EDCRA requires one-time messaging from relays to the BS (about the users' channel conditions) and one-time messaging from the BS to relays (about the subchannel and power allocations). It has clear overall computational complexity, too, which is of $O(KN^2 + (M+1)\log(1/\epsilon_1))$. Based on the discussions in this section as well as the previous ones, EDDRA is more suitable for implementation in practice while EDCRA and JFRP provide good baselines for comparison purposes.

5 Conclusions

In this paper, we have studied data admission control and resource allocation in buffer-aided relay-assisted OFDMA networks. We have formulated time slot, subchannel, power allocation as a utility-based stochastic optimization problem, taking into account several practical constraints. Using the Lyapunov drift-plus-penalty policy, we have transformed the problem into instantaneous subproblems while introducing several parameters related to cellular networks. For practical considerations, we have proposed low-complexity strategy for the allocation of power and subchannels and have provided distributed and centralized schemes to utilize it. In particular, the proposed EDDRA policy is attractive for use in practice due to its low complexity as well as low signaling overhead. Numerical results confirm the effectiveness of the proposed parameters and also show that the proposed algorithms lead to significant improvement in data admission and throughput of the system.

Endnotes

[1] In a centralized implementation, the BS has information about all the queue sizes, due to the fact that it has the history of the transmissions from all the queues.

[2] In the above discussions, we excluded the signaling overhead of channel state feedbacks from the receiver side of any link to the transmitter side, due to the fact that it is the same in EDDRA and EDCRA.

Competing interests
The authors declare that they have no competing interests.

Acknowledgements
This work was supported by the Canadian Natural Sciences and Engineering Research Council through grants RGPIN-2014-06119 and RGPAS-462031-2014, and the National Natural Science Foundation of China through Grant No. 61271182. The work of Amr Mohamed was supported by NPRP 5-782-2-322 from the Qatar National Research Fund (a member of Qatar Foundation). The statements made herein are solely the responsibility of the authors.

Author details
[1] ECE Department, The University of British Columbia, Vancouver, Canada. [2] CSE Department, Qatar University, Doha, Qatar.

Received: 13 May 2015 Accepted: 12 November 2015
Published online: 22 December 2015

References

1. IEEE Std 802.16j-2009, *Part 16: Air interface for broadband wireless access systems-amendment 1: Multihop relay specification*. (IEEE, New York, USA, 2009)
2. 3GPP TR 36.814 V9.0.0, *Evolved universal terrestrial radio access (E-UTRA); further advancements for E-UTRA physical layer aspects*. (3GPP, Valbonne, France, 2010). http://www.3gpp.org/dynareport/36814.htm
3. M Salem, A Adinoyi, H Yanikomeroglu, D Falconer, Opportunities and challenges in OFDMA-based cellular relay networks: a radio resource management perspective. IEEE Trans. Veh. Technol. **59**, 2496–2510 (2010)
4. SJ Kim, X Wang, M Madihian, Optimal resource allocation in multi-hop OFDMA wireless networks with cooperative relay. IEEE Trans. Wirel. Commun. **7**, 1833–1838 (2008)
5. K Sundaresan, S Rangarajan, in *Proc. IEEE International Conference on Computer Communications*. Adaptive resource scheduling in wireless OFDMA relay networks (IEEE, New York, USA, 2012), pp. 1080–1088
6. DWK Ng, R Schober, Cross-layer scheduling for OFDMA amplify-and-forward relay networks. IEEE Trans. Veh. Technol. **59**, 1443–1458 (2009)
7. DWK Ng, ES Lo, R Schober, Dynamic resource allocation in MIMO-OFDMA systems with full-duplex and hybrid relaying. IEEE Trans. Commun. **60**, 1291–1304 (2012)
8. Y Pan, A Nix, M Beach, Distributed resource allocation for OFDMA-based relay networks. IEEE Trans. Veh. Technol. **60**, 919–931 (2011)
9. S Zhang, XG Xia, in *Proc. IEEE Wireless Communications and Networking Conference*. A high-efficiency semi-distributed resource allocation in OFDMA-based wireless relay networks (IEEE, New York, USA, 2013), pp. 3277–3281
10. O Oyman, Opportunistic scheduling and spectrum reuse in relay-based cellular networks. IEEE Trans. Wirel. Commun. **9**, 1074–1085 (2010)
11. J Liang, H Yin, H Chen, Z Li, S Liu, in *Proc. IEEE Vehicular Technology Conference*. A novel dynamic full frequency reuse scheme in OFDMA cellular relay networks (IEEE, New York, USA, 2011), pp. 1–5
12. H Mei, J Bigham, P Jiang, E Bodanese, Distributed dynamic frequency allocation in fractional frequency reused relay based cellular networks. IEEE Trans. Commun. **61**, 1327–1336 (2013)

13. N Zlatanov, R Schober, P Popovski, in *Proc. IEEE Global Telecommunications Conference*. Throughput and diversity gain of buffer-aided relaying, (2011), pp. 1–6
14. T Riihonen, R Wichman, S Werner, Evaluation of OFDM(A) relaying protocols: capacity analysis in infrastructure framework. IEEE Trans. Veh. Technol. **61**, 360–374 (2011)
15. N Zlatanov, A Ikhlef, T Islam, R Schober, Buffer-aided cooperative communications: opportunities and challenges. IEEE Commun. Mag. **52**, 146–153 (2014)
16. J Yang, H Hu, H Xi, L Hanzo, Online buffer fullness estimation aided adaptive media playout for video streaming. IEEE Trans. Multimed. **13**, 1141–1153 (2011)
17. J Yang, Y Ran, S Chen, W Li, L Hanzo, Online source rate control for adaptive video streaming over HSPA and LTE-style variable bitrate downlink channels. To appear in IEEE Trans. Veh. Technol., 1–1 (2015)
18. M Salem, A Adinoyi, M Rahman, H Yanikomeroglu, D Falconer, Y Kim, Fairness-aware radio resource management in downlink OFDMA cellular relay networks. IEEE Trans. Wirel. Commun. **9**, 1628–1639 (2010)
19. M Salem, A Adinoyi, H Yanikomeroglu, D Falconer, Fair resource allocation toward ubiquitous coverage in OFDMA-based cellular relay networks with asymmetric traffic. IEEE Trans. Veh. Technol. **60**, 2280–2292 (2011)
20. J Hajipour, A Mohamed, VCM Leung, in *Proc. International Conference on Wireless and Mobile Communications*. Dynamic distributed resource allocation in relay assisted OFDMA networks (IARIA, 2012). www.thinkmind.org
21. H Ju, B Liang, J Li, X Yang, Dynamic joint resource optimization for LTE-Advanced relay networks. IEEE Trans. Wirel. Commun. **12**, 5668–5678 (2013)
22. MJ Neely, *Stochastic Network Optimization with Application to Communication and Queueing Systems*. (Morgan & Claypool, San Rafael, 2010)
23. X Qiu, K Chawla, On the performance of adaptive modulation in cellular systems. IEEE Trans. Commun. **47**, 884–895 (1999)
24. 3GPP TS 36.300 V13.0.0 (2015-06). Evolved universal terrestrial radio access (E-UTRA) and evolved universal terrestrial radio access network (E-UTRAN); overall description; stage 2. (3GPP, Valbonne, France). http://www.3gpp.org/dynareport/36300.htm
25. S Supittayapornpong, MJ Neely, in *Proc. IEEE International Conference on Computer Communications*. Achieving utility-delay-reliability tradeoff in stochastic network optimization with finite buffers (IEEE, New York, USA, 2015), pp. 1–9
26. H Hu, H Yanikomeroglu, DD Falconer, S Periyalwar, in *Proc. IEEE Global Telecommunications Conference*. Range extension without capacity penalty in cellular networks with digital fixed relays (IEEE, New York, USA, 2004), pp. 3053–3057
27. DP Bertsekas, *Nonlinear Programming*, 2nd edn. (Athena Scientific, Nashua, 1999)
28. MC Jeruchim, P Balaban, KS Shanmugan, *Simulation of Communication Systems: Modeling, Methodology and Techniques*, 2nd edn. (Kluwer Academic, Dordrecht, 2000)
29. 3GPP TR 25.996 V7.0.0 (2007-06), Spatial channel model for multiple input multiple output (MIMO) simulations. (3GPP, Valbonne, France). http://www.3gpp.org/DynaReport/25996.htm
30. L Vandenberghe, EE236C. Class lecture, topic: optimization methods for large-scale systems: subgradient method. http://www.seas.ucla.edu/~vandenbe/236C/lectures/sgmethod.pdf