Hindawi Publishing Corporation
Mobile Information Systems
Volume 2016, Article ID 6306580, 23 pages
http://dx.doi.org/10.1155/2016/6306580

Research Article

Joint Bandwidth and Power Allocation for LTE-Based Cognitive Radio Network Based on Buffer Occupancy

Alia Asheralieva and Yoshikazu Miyanaga

Laboratory of Information Communication Networks, Hokkaido University, Kita 14, Nishi 9, Kita-ku, Sapporo, Hokkaido 060-0814, Japan

Correspondence should be addressed to Alia Asheralieva; aasheralieva@gmail.com

Received 17 August 2015; Accepted 23 November 2015

Academic Editor: Yuh-Shyan Chen

We investigate the problem of resource allocation in a cognitive long-term evolution (LTE) network, where the available bandwidth resources are shared among the primary (licensed) users (PUs) and secondary (unlicensed) users (SUs). Under such spectrum sharing conditions, the transmission of the SUs should have minimal impact on quality of service (QoS) and operating conditions of the PUs. To achieve this goal, we propose to assign the network resources based on the buffer sizes of the PUs and SUs in the uplink (UL) and downlink (DL) directions. To ensure that the QoS requirements of the PUs are satisfied, we enforce some upper bound on the size of their buffers considering two network usage scenarios. In the first scenario, PUs pay full price for accessing the spectrum and get full QoS protection; the SUs access the network for free and are served on a best-effort basis. In the second scenario, PUs pay less in exchange for sharing the bandwidth and get the reduced QoS guarantees; SUs pay some price for their access without any QoS guarantees. Performance of the algorithms proposed in the paper is evaluated using simulations in OPNET environment. The algorithms show superior performance when compared with other relevant techniques.

1. Introduction

The traditional fixed spectrum allocation policy has been characterized by a very ineffective spectrum utilization resulting in an artificial scarcity of the network resources [1], which stimulated a surge of interest to an alternative spectrum usage concept known as cognitive radio (CR) [2]. In a CR network (CRN), the available bandwidth resources can be shared among the PUs (paying some price for accessing spectrum) and the SUs (who can get the wireless access for free). Needless to say under such spectrum sharing conditions the transmission of the SUs should have minimal impact on QoS and operating conditions of the PUs [2].

Among existing wireless standards considered for deployment in CRNs, LTE is considered to be the most favourable due to such appealing features as spectrum flexibility, fast adaptation to time-varying channel conditions, high spectral efficiency, and robustness against interference [3]. A detailed description of the LTE radio interface can be found, for instance, in [4]. In short, LTE is based on the universal terrestrial radio access (UTRA) and a high-speed downlink packet access (HSDPA). In the DL, LTE uses an orthogonal frequency division multiple access (OFDMA), which has high spectral efficiency and robustness against the interference. A single carrier frequency division multiple access (SC-FDMA) is applied in the UL direction, due to its lower (compared to OFDM) peak-to-average power ratio (PAPR) [5]. The numerology of LTE includes a subcarrier spacing of 15 kHz, support for scalable bandwidth of up to 20 MHz, and a resource allocation granularity of 180 kHz × 1 ms (called a resource block). Available transmission resources are distributed among the users by the medium access control (MAC) schedulers located in enhanced NodeBs (eNBs). Depending on the implementation, the scheduling can be done based on queuing delay, instantaneous channel conditions, fairness, and so forth [6, 7].

Due to existence of two types of users (primary and secondary) with different QoS requirements, the problems of dynamic spectrum access (DSA) and resource allocation for CRNs are more complex than those considered in traditional

wireless networks. The majority of the works on resource allocation in LTE-based CRNs focus on various lower layer techniques for spectrum sensing and spectrum mobility (see, e.g., [8–10]). These techniques are very effective in identifying and reducing the interference in the physical channels but do not improve the overall user-perceived QoS, which is mainly expressed in terms of the packet end-to-end delay and loss for the network users. Consequently, the results of these works are applicable only in combination with the techniques designed for MAC and higher layers [11].

The QoS protection for the PUs and the admissibility of the SUs have been studied using the theoretical analysis of user behaviour. For instance, in [12] the authors propose a statistical traffic control for the LTE-based CRNs. To satisfy the timing constraints of all packets belonging to different streams (with diverse QoS characteristics and requirements) and achieve the statistical QoS guarantees, the authors deploy admission control and coordinated transmissions of the constant-bit-rate (CBR) and the variable-bit-rate (VBR) streams. Dynamic channel selection for autonomous wireless users transmitting delay-sensitive multimedia applications over a CRN has been studied in [13]. Unlike prior works where the application-layer requirements of the users have not been considered, here the rate and delay requirements of heterogeneous multimedia users are taken into account. The authors propose a novel priority based virtual queue interface to efficiently manage the available spectrum resources. This interface is used to (i) determine the required information exchanges, (ii) evaluate the expected delays experienced by various priority traffics, (iii) design a dynamic strategy learning (DSL) algorithm deployed at each user that exploits the expected delay, and (iv) adapt to the channel selection strategies to maximize the user's utility function.

The authors of [5] present a delay-power control (DPC) exploiting the trade-off between the transmission delay and transmission power in wireless networks. In a proposed resource allocation procedure, each wireless link autonomously updates its power based on the interference observed at its receiver (without any crosslink communication). The DPC scheme has been proved to converge to a unique equilibrium point and contrasted to the well-known Foschini-Miljanic (FM) formulation. Some theoretical underpinnings of DPC and their practical implications for efficient protocol design have also been established. In [14], the problem of allocating the network resources (channels and transmission power) in a multihop CRN is modeled as a multicommodity flow problem with the dynamic link capacity resulting from dynamic resource allocation. Based on such queue-balancing, the authors propose a distributed scheme for optimal resource allocation without exchanging the spectrum dynamics between remote nodes. Considering the power masks, each node makes resource allocation decisions based on current or past local information from neighboring nodes to satisfy the throughput requirement of each flow and maintain the network stability.

In this paper, we suggest an alternative strategy for resource allocation in a LTE-based CRN and propose to assign the network resources (bandwidth and transmission power) to the UL and DL of LTE system, based on buffer size of user equipment (UE) pieces. Note that, in many previously proposed algorithms (e.g., [8–13]), the bandwidth and transmission power are assigned without considering the buffer occupancy of UE pieces, which may lead to a rather unfair resource allocation (when users with lower demands are allocated larger bandwidth than the users with higher demands).

To ensure that the QoS requirements of the PUs are satisfied, we put the constraints on the sizes of their buffers. Two network usage scenarios are considered. In the first scenario, the PUs pay full price for accessing the spectrum and get the full QoS protection, whereas the SUs access the network for free and are served on a best-effort basis. In the second scenario, the PUs pay less in exchange for sharing the bandwidth and get the reduced QoS guarantees; SUs pay some price for their usage without any QoS guarantees. The proposed resource allocation algorithm is derived based on a discrete spectrum assumption, with the spectrum resources counted in terms of LTE resource blocks (RBs). Note that a continuous spectrum assumption (used in all past works) is not applicable for a practical LTE realization, since the number of RBs comprising the bandwidth is relatively small (6, 10, 20, 50, and 100 RBs corresponding to 1.4, 3, 5, 10, and 100 MHz widebands, resp., [15]). Unlike existing research contributions, the transmission rates of SUs and DUs are calculated using the modified Shannon expression which accounts for the adaptive modulation and coding (AMC) used in LTE.

The rest of the paper is organized as follows. In Section 2 we describe the network model and formulate the resource allocation problems in two network usage scenarios. In Section 3 we provide the solution methodology and discuss the implementation of a presented resource allocation approach in a real LTE system. In Section 4 we outline a simulation model and evaluate performance of the proposed resource allocation algorithms in two network usage scenarios. The paper is finalized in Conclusion.

2. Problem Statement

2.1. Network Model, Assumptions, and Notation. In this work, the problem of joint power and bandwidth allocation for the LTE-based CRN is investigated for both the UL and the DL directions. Similarly, the discussion through the rest of the paper is applicable (if not stated otherwise) to either direction.

Consider a basic LTE-based CRN architecture which consists of one eNB providing wireless access to n PUs numbered $\mathrm{PU}_1, \ldots, \mathrm{PU}_n$, and m SUs numbered $\mathrm{SU}_1, \ldots, \mathrm{SU}_m$. Two spectrum usage scenarios are considered in this paper. In the first scenario, PUs are the licensed network users who pay some price for accessing the spectrum and therefore must be provided with the certain guaranteed QoS levels; SUs are the unlicensed network users, who can access the spectrum for free, and therefore they are served on a best-effort basis. In the second scenario, the PUs pay less in exchange for sharing the spectrum; the SUs have to pay some price for the shared spectrum to compensate for the income losses of a service provider.

The considered network operates on a slotted-time basis with the time axis partitioned into mutually disjoint time intervals (slots) $\{T_s, (t+1)T_s\}, t = 0, 1, 2, \ldots$, with T_s denoting the slot length and t being the slot index. The number of active PUs and SUs can be tracked using an LTE random access channel (RACH) procedure [15], which is used for initial access to the network (i.e., for originating, terminating, or registration calls).

In the UL direction, the user-generated traffic (in bits per slot or bps) is enqueued in the buffers of UE pieces and then transmitted to the eNB using a standard packet scheduling procedure [15]. In this procedure, the information about the amount of data (in bps) enqueued in the buffers of UE pieces is constantly transmitted to the eNB, so that the eNB "knows" the exact amount of data generated by the users at any slot t. This information is then used by the eNB to allocate the UL transmission resources to UE pieces. In the DL direction, the transmission resources are allocated based on the amount of data transmitted by eNB to the users.

In LTE system, the transmission resources are allocated to the users in terms of RBs. Each user can be allocated only the integer number of RBs, and these RBs should not be adjacent to each other [7, 15]. Another important feature of both the SC-FDMA (applied in the UL direction) and OFDMA (applied in the DL direction) is the orthogonality of resource allocation, which allows achieving minimal level of cochannel interference between different transmitter-receiver pairs located within one cell (i.e., when no frequency reuse is considered).

It is assumed that the bandwidth of the eNB is fixed and equal to B RBs. For any i in $\mathbf{I} = \{1, \ldots, n\}$ and any j in $\mathbf{J} = \{1, \ldots, m\}$, consider the following:

(i) The channel between the eNB and PU_i is characterized by the noise and interference coefficient $0 < h_{Pi} \leq 1$, which is known to the eNB and PU_i. The channel between the eNB and SU_j is characterized by the noise and interference coefficient $0 < h_{Sj} \leq 1$, which is known to the eNB and SU_j. The values of h_{Pi} and h_{Sj} in the UL and DL directions can be obtained from the channel state information (CSI) through the use of the LTE reference signals (RSs) [16, 17].

(ii) The size of the buffers and the arrival rates (in bits) of PU_i and SU_j can be observed at any slot t in the UL and DL directions.

The following notations are used throughout the paper:

(i) $b_{Pi}(t)$ and $b_{Sj}(t)$ denote the number of RBs allocated at slot t to PU_i and SU_j, respectively;

(ii) $p_{Pi}(t)$ and $p_{Sj}(t)$ denote the transmission power allocated at slot t to PU_i and SU_j, respectively;

(iii) $a_{Pi}(t)$ and $a_{Sj}(t)$ denote the arrival rate (in bps) at slot t in PU_i and SU_j, respectively;

(iv) $q_{Pi}(t)$ and $q_{Sj}(t)$ denote the buffer size (in bits) at slot t of PU_i and SU_j, respectively;

(v) $r_{Pi}(t)$ and $r_{Sj}(t)$ denote the transmission rate (in bps) at slot t in the channels of PU_i and SU_j, respectively;

(vi) P_{eNB}, P_{Pi}, and P_{Sj} denote the maximal transmission power of the eNB, PU_i, and SU_j, respectively.

2.2. Scenario 1. The objective of this paper is to devise a sustainable algorithm for joint bandwidth and transmission power allocation, based on two goals:

(1) *QoS Protection for the PUs*. PUs should maintain their target QoS, irrespective of how many SUs enter the network and how much load they are generating.

(2) *Admissibility of the SUs*. SUs should be able to utilize the spare capacity of the eNB (if left) to maximize their QoS.

These goals, however, cannot be achieved without providing the quantified measures for the user-perceived QoS. Note that because of the orthogonality of the RB allocation, the need for interference mitigation in a considered CRN is eliminated. In this case, the use of such physical layer characteristics as signal-to-noise ratio (SNR) is not well-reasoned. More practical QoS measures can be obtained from the higher layers network information. For instance, such metrics as round-trip latency and loss are traditionally used for performance evaluation in wireless networks [5]. Unfortunately, in LTE system the direct estimation of delay and loss is rather complex. For instance, the end-to-end latency consists of various delay components, including transmission and queuing delays, propagation and processing delays, the UL delay due to scheduling, and delay due to hybrid automatic repeat request (HARQ) [18]. The accurate analysis of these delay components requires knowledge of many system parameters, which may be not available during resource allocation.

Consequently, in this paper we argue in favour of using the buffer size of UE pieces as a QoS measure. The buffer size of PUs and SUs can be easily estimated from the known parameters $q_{Pi}(t)$, $a_{Pi}(t)$ and $q_{Sj}(t)$, $a_{Sj}(t)$ using well-known Lindley's equation [19]:

$$q_{Pi}(t+1) = \left\lceil q_{Pi}(t) + a_{Pi}(t) - r_{Pi}(t) \right\rceil^{+}, \quad \forall i \in \mathbf{I}; \tag{1a}$$

$$q_{Sj}(t+1) = \left\lceil q_{Sj}(t) + a_{Sj}(t) - r_{Sj}(t) \right\rceil^{+}, \quad \forall j \in \mathbf{J}, \tag{1b}$$

where $\lceil x \rceil^{+} = \max\{0, x\}$, whereas the transmission rates $r_{Pi}(t)$, $r_{Sj}(t)$ depend on the unknown bandwidth and transmission power allocation vectors given by (relationship between the transmission rate, transmission power, and the bandwidth will be established in the next section)

$$\begin{aligned} \mathbf{p}_P &= \left[p_{P1}(t), \ldots, p_{Pn}(t)\right]^T, \\ \mathbf{b}_P &= \left[b_{P1}(t), \ldots, b_{Pn}(t)\right]^T; \end{aligned} \tag{2a}$$

$$\begin{aligned} \mathbf{p}_S &= \left[p_{S1}(t), \ldots, p_{Sm}(t)\right]^T, \\ \mathbf{b}_S &= \left[b_{S1}(t), \ldots, b_{Sm}(t)\right]^T. \end{aligned} \tag{2b}$$

Note that the average round-trip latency of any primary UE can be retained below some upper bound level by

restricting its buffer size to be less than a certain target buffer size. The only information which is necessary in this case is the desired upper bound on delay and the average arrival rate in PUs during some interval $T > 0$. Then, we can find the target buffer size by deploying Little's formula [20]

$$D_i \cdot \overline{a_{Pi}}(t) = L_i, \quad \forall i \in \mathbf{I}, \tag{3a}$$

where D_i is the upper bound on delay for PU_i; L_i is the target buffer size for PU_i; $\overline{a_{Pi}}(t)$ is the average arrival in PU_i during time interval $[t - T + 1, t]$ given by

$$\overline{a_{Pi}}(t) = \sum_{\tau=t-T+1}^{t} a_{Pi}(t), \quad \forall i \in \mathbf{I}. \tag{3b}$$

To maintain the desired upper bound on the end-to-end delay, the buffer size of PUs should be constrained as follows:

$$q_{Pi}(t+1) \le L_i, \quad \forall i \in \mathbf{I}, \tag{4}$$

or equivalently using (1a)

$$q_{Pi}(t) + a_{Pi}(t) - L_i \le r_{Pi}(t) \le q_{Pi}(t) + a_{Pi}(t), \quad \forall i \in \mathbf{I}. \tag{5}$$

If added to an optimization problem, constraint (5) can guarantee the QoS protection for PUs. Except some (very rare) cases when constraint (5) is not feasible, the buffer size will always stay below the target level irrespective of how many SUs are in the network and how much load they are generating. The setting of the parameters L_i, as well as the cases when constraint (5) is unfeasible, is discussed in Section 3.

Note that, at any slot t, the number of RBs allocated for data transmissions of PUs and SUs in the UL and DL directions should not exceed B RBs. That is,

$$\sum_{i\in\mathbf{I}} b_{Pi}(t) + \sum_{j\in\mathbf{J}} b_{Sj}(t) \le B, \tag{6a}$$

where

$$b_{Pi}(t) \in \mathbf{Z}^+, \quad b_{Sj}(t) \in \mathbf{Z}^+, \quad \forall i \in \mathbf{I},\ \forall j \in \mathbf{J}, \tag{6b}$$

where $\mathbf{Z}^+$ represents the set of all nonnegative integers.

Additionally, the transmission power allocated to the users at any slot t should be nonnegative and cannot exceed the maximal transmission power levels. Hence,

$$p_{Pi}(t) \le P_{Pi}, \quad p_{Sj}(t) \le P_{Sj}, \quad \forall i \in \mathbf{I},\ \forall j \in \mathbf{J}, \tag{7a}$$

for the UL direction, and

$$\sum_{i\in\mathbf{I}} p_{Pi}(t) + \sum_{j\in\mathbf{J}} p_{Sj}(t) \le P_{\mathrm{eNB}}, \tag{7b}$$

for the DL direction, where

$$p_{Pi}(t) \ge 0, \quad p_{Sj}(t) \ge 0, \quad \forall i \in \mathbf{I},\ \forall j \in \mathbf{J}. \tag{7c}$$

We are now ready to show how to attain the second goal of resource allocation. As in case with QoS protection of PUs, we use the buffer sizes of SUs as a measure of the SUs' QoS. We say that the admissibility of SUs in the network is preserved if they are allocated the bandwidth that is not used by the PUs. This can be done, for instance, by minimizing the sum (or the logarithmic sum) of the buffer sizes of SUs or minimizing the maximal buffer size of SUs subject to the constraints listed above.

For the UL direction, this gives us the following problem (to simplify notation, we skip the index t below):

$$\text{minimize} \quad \max_{j\in\mathbf{J}} \left\lceil q_{Sj} + a_{Sj} - r_{Sj} \right\rceil^+ \tag{8a}$$

$$\text{subject to:} \quad q_{Pi} + a_{Pi} - L_i \le r_{Pi} \le q_{Pi} + a_{Pi}, \quad \forall i \in \mathbf{I} \tag{8b}$$

$$\sum_{i\in\mathbf{I}} b_{Pi} + \sum_{j\in\mathbf{J}} b_{Sj} \le B \tag{8c}$$

$$b_{Pi} \in \mathbf{Z}^+, \quad b_{Sj} \in \mathbf{Z}^+, \quad \forall i \in \mathbf{I},\ \forall j \in \mathbf{J}, \tag{8d}$$

$$p_{Pi} \le P_{Pi}, \quad p_{Sj} \le P_{Sj}, \quad \forall i \in \mathbf{I},\ \forall j \in \mathbf{J}, \tag{8e}$$

$$p_{Pi} \ge 0, \quad p_{Sj} \ge 0, \quad \forall i \in \mathbf{I},\ \forall j \in \mathbf{J}. \tag{8f}$$

For the DL direction, constraint (8e) should be replaced by

$$\sum_{i\in\mathbf{I}} p_{Pi} + \sum_{j\in\mathbf{J}} p_{Sj} \le P_{\mathrm{eNB}}. \tag{8g}$$

In above formulation, each PU is guaranteed a certain target QoS level, and the optimal allocations for PUs are such that

$$q_{Pi} + a_{Pi} - r_{Pi} = L_i, \quad \forall i \in \mathbf{I}. \tag{9}$$

The spare bandwidth (if left) is distributed among the SUs to minimize the size of their buffer, which means that the two goals of resource allocation are achieved. However, this formulation works well only if the aggregated traffic demand of the SUs is greater than that of the PUs. Otherwise, there might be a situation when the optimal solution for SUs is such that

$$\max_{j\in\mathbf{J}} \left\lceil q_{Sj} + a_{Sj} - r_{Sj} \right\rceil^+ \le L_i, \quad \forall i \in \mathbf{I}, \tag{10}$$

which is apparently not fair, since the PUs have larger buffer sizes than the SUs, and therefore they will experience longer delays than the SUs.

To avoid such situation, we propose to modify problem (8a)–(8g) by replacing its objective (8a) with

$$\begin{aligned} \text{minimize} \quad & \max_{i\in\mathbf{I}} \left\lceil q_{Pi} + a_{Pi} - r_{Pi} \right\rceil^{+} \\ & + \max_{j\in\mathbf{J}} \left\lceil q_{Sj} + a_{Sj} - r_{Sj} \right\rceil^{+}. \end{aligned} \tag{11}$$

Objective (11) guarantees that PUs are served with maximal QoS, even if the aggregated traffic demand of SUs is greater than that of the PUs.

2.3. Scenario 2. We now consider a different scenario in which the PUs will benefit (by paying less in exchange) from sharing their bandwidth with the SUs. Let us assume that, in the network without price benefits (i.e., in Scenario 1), the price for accessing the licensed spectrum of the PUs equals 1. In this case, all the PUs get the full QoS protection (i.e., guaranteed target buffer size L_i and the average end-to-end latency D_i). Suppose that each PU can define the portion of allocated spectrum that it is willing to share with the SUs. The delay in turn is related to $\lceil q_{Pi} + a_{Pi} - r_{Pi} \rceil^{+}$.

Let us denote by α_i, $0 \le \alpha_i < 1$ the price for the portion of spectrum that PU_i is willing to share with the SUs. So, α_i represents the willingness of PU_i to trade off its QoS for money. Subsequently, the guaranteed QoS level of PU_i will reduce, and instead of the QoS protection constraint (8b) we will have

$$r_{Pi} \ge q_{Pi} + a_{Pi} - (1 + \alpha_i) L_i, \quad \forall i \in \mathbf{I}. \tag{12}$$

Now, to compensate for the income losses of a service provider, the SUs will have to pay the price for the shared spectrum. Let β denote the price that each SU will pay for accessing the shared spectrum. The amount that SU_j pays in this case is βr_{Sj}, whereas the revenue of the service provider is

$$\sum_{j\in\mathbf{J}} \beta r_{Sj} - \sum_{i\in\mathbf{I}} \alpha_i \left\lceil q_{Pi} + a_{Pi} - r_{Pi} \right\rceil^{+}. \tag{13}$$

Assuming that a service provider does not incur loss by allowing the secondary users to use the spectrum, we have to incorporate the constraint

$$\sum_{j\in\mathbf{J}} \beta r_{Sj} - \sum_{i\in\mathbf{I}} \alpha_i \left\lceil q_{Pi} + a_{Pi} - r_{Pi} \right\rceil^{+} \ge \nu, \tag{14}$$

where ν is the profit that a service provider would like to make by allowing the SUs into the network.

Note that, in such formulation, the SUs get no QoS guarantee, and therefore we have to minimize the price that the SUs pay for their usage. Thus, the optimization problem is

$$\begin{aligned} \text{minimize} \quad & \beta && \text{(15a)} \\ \text{subject to:} \quad & q_{Pi} + a_{Pi} - (1 + \alpha_i) L_i \le r_{Pi} \\ & \le q_{Pi} + a_{Pi}, && \text{(15b)} \\ & \forall i \in \mathbf{I}, \\ & \sum_{i\in\mathbf{I}} b_{Pi} + \sum_{j\in\mathbf{J}} b_{Sj} \le B && \text{(15c)} \\ & b_{Pi} \in \mathbf{Z}^{+}, \\ & b_{Sj} \in \mathbf{Z}^{+}, && \text{(15d)} \\ & \forall i \in \mathbf{I}, \ \forall j \in \mathbf{J}, \\ & p_{Pi} \le P_{Pi}, \\ & p_{Sj} \le P_{Sj}, && \text{(15e)} \\ & \forall i \in \mathbf{I}, \ \forall j \in \mathbf{J}, \\ & p_{Pi} \ge 0, \\ & p_{Sj} \ge 0, && \text{(15f)} \\ & \forall i \in \mathbf{I}, \ \forall j \in \mathbf{J}, \\ & \sum_{j\in\mathbf{J}} \beta r_{Sj} - \sum_{i\in\mathbf{I}} \alpha_i \left\lceil q_{Pi} + a_{Pi} - r_{Pi} \right\rceil^{+} \ge \nu, && \text{(15g)} \end{aligned}$$

for the UL direction. For the DL direction constraint (15e) is replaced by constraint (8g). The value of β obtained by solving (15a)–(15g) is then broadcasted by a service provider to the SUs that will have to pay a determined price.

We note, in passing, that there are many other ways in which the pricing problem could be addressed. For instance, the numbers α_i may be negotiated between the PUs and a service provider. One can also think of a scenario where different SUs may be able to express their willingness to pay via bidding the prices β_j. The options are numerous, but they are beyond the scope of this paper.

3. Implementation Issues

3.1. Rate as a Function of Bandwidth and Power. Note that the transmission rates in the UL and DL channels of a LTE system can be found using the modified Shannon equation [21], expressed as

$$\begin{aligned} r_{Pi}(\mathbf{b}_P, \mathbf{p}_P) &= \omega \psi_{Pi} b_{Pi} \log\left(1 + \xi_{Pi} \text{SNR}_{Pi}\right) \\ &= \omega \psi_{Pi} b_{Pi} \log\left(1 + \xi_{Pi} h_{Pi} p_{Pi}\right), \end{aligned} \tag{16a}$$

$$0 < \psi_{Pi}, \ \xi_{Pi} \le 1, \ \forall i \in \mathbf{I};$$

$$r_{Sj}(\mathbf{b}_S, \mathbf{p}_S) = \omega\psi_{Sj} b_{Sj} \log\left(1 + \xi_{Sj}\mathrm{SNR}_{Sj}\right)$$
$$= \omega\psi_{Sj} b_{Sj} \log\left(1 + \xi_{Sj} h_{Sj} p_{Sj}\right), \quad (16b)$$
$$0 < \psi_{Sj},\ \xi_{Sj} \le 1,\ \forall j \in \mathbf{J},$$

where SNR_{Pi} and SNR_{Sj} are the SNRs in the channels of PU_i and SU_j, respectively; ψ_{Pi} and ψ_{Sj} are the system bandwidth efficiencies for PU_i and SU_j, respectively; ξ_{Pi} and ξ_{Sj} are the SNR efficiencies for PU_i and SU_j, respectively; ω is the bandwidth of one RB ($\omega = 180$ kHz) [21].

In real LTE networks, the system bandwidth efficiency is strictly less than 1 due to the overheads on the link and system levels. The bandwidth efficiency is fully determined by the design and internal settings of the system and does not depend on the physical characteristics of the wireless channels between the user and the eNB [21, 22]. Hence, it is reasonable to assume that

$$\psi_{Pi} = \psi_{Sj} = \psi, \quad \forall i \in \mathbf{I},\ \forall j \in \mathbf{J}. \quad (17)$$

The SNR efficiency is mainly limited by the maximum efficiency of the supported modulation and coding scheme (MCS) [16]. In LTE, MCS is chosen using AMC to maximize the data rate by adjusting the transmission parameters to the current channel conditions. AMC is one of the realizations of a dynamic link adaptation. In AMC algorithm, the appropriate MCSs for packet transmissions are assigned periodically (within short fixed time interval usually equal to one slot) by the eNB based on the instantaneous channel conditions reported by the users. The higher MCS values are allocated to the high-quality channels to achieve higher transmission rate. The lower MCS values are assigned to the channels of poor quality to decrease the transmission rate and, consequently, ensure the transmission quality [16, 17].

The method for choosing MCS can be expressed as follows. Based on the instantaneous radio channel conditions, the SNR is calculated for the DL wireless channels between the eNB and the users. A supported MCS index is chosen using expression [16, 17]

$$\mathrm{MCS} = \begin{cases} \mathrm{MCS}_1, & \mathrm{SNR} < \gamma_1, \\ \mathrm{MCS}_2, & \gamma_1 \le \mathrm{SNR} < \gamma_2, \\ \vdots & \vdots \\ \mathrm{MCS}_l, & \gamma_{l-1} \le \mathrm{SNR} < \gamma_l, \end{cases} \quad (18)$$

where $\gamma_1 < \gamma_2 < \cdots < \gamma_l$ are the SNR thresholds corresponding to −10 dB bit error ratio (BER) given by the Additive White Gaussian Noise (AWGN) curves for each MCS (the standard AWGN curves can be found in [17]). The LTE standard allows $l = 15$ MCS levels (description of MCS levels, their code rate, and efficiency is shown in Table 1) [17].

Table 1: Allowed MCS values in LTE standard [17].

MCS index	Modulation	Code rate	Efficiency
0	No transmission		
1	QPSK	78	0.1523
2	QPSK	120	0.2344
3	QPSK	193	0.3770
4	QPSK	308	0.6016
5	QPSK	449	0.8770
6	QPSK	602	1.1758
7	16 QAM	378	1.4766
8	16 QAM	490	1.9141
9	16 QAM	616	2.4063
10	64 QAM	466	2.7305
11	64 QAM	567	3.3223
12	64 QAM	666	3.9023
13	64 QAM	772	4.5234
14	64 QAM	873	5.1152
15	64 QAM	948	5.5547

Based on the above, the SNR efficiency depends on the SNR of the DL wireless channel and therefore can be represented by a function

$$g(\mathrm{SNR}) = \begin{cases} \xi_1, & \mathrm{SNR} < \gamma_1, \\ \xi_2, & \gamma_1 \le \mathrm{SNR} < \gamma_2, \\ \vdots & \vdots \\ \xi_l, & \gamma_{l-1} \le \mathrm{SNR} < \gamma_l, \end{cases} \quad (19)$$

where $0 < \xi_1 < \xi_2 < \cdots < \xi_l$ are the values of the SNR efficiency corresponding to the supported MCS.

Using (17) and (19), the transmission rates of the users can be expressed as

$$r_{Pi}(\mathbf{b}_P, \mathbf{p}_P) = \omega\psi b_{Pi} \log\left(1 + g\left(\mathrm{SNR}_{Pi}\right) \cdot \mathrm{SNR}_{Pi}\right)$$
$$= \omega\psi b_{Pi} \log\left(1 + g\left(p_{Pi}\right) \cdot h_{Pi} p_{Pi}\right), \quad (20a)$$
$$\forall i \in \mathbf{I};$$
$$r_{Sj}(\mathbf{b}_S, \mathbf{p}_S) = \omega\psi b_{Sj} \log\left(1 + g\left(\mathrm{SNR}_{Sj}\right) \cdot \mathrm{SNR}_{Sj}\right)$$
$$= \omega\psi b_{Sj} \log\left(1 + g\left(p_{Sj}\right) h_{Sj} p_{Sj}\right), \quad (20b)$$
$$\forall j \in \mathbf{J}.$$

3.2. Solution Methodology. In this subsection, we describe a solution methodology for the resource allocation problem in Scenario 1 in the UL direction (with the objective given by (11) and the constraints defined in (8b)–(8f)). A resource allocation problem for the DL direction as well as problem (15a)–(15g) for Scenario 2 can be solved analogously.

Note that, in Scenario 1, a resource allocation problem in the UL direction is equivalent to

$$\text{minimize} \quad \Omega + \Phi \tag{21a}$$

$$\text{subject to:} \quad \sum_{i\in\mathbf{I}} b_{Pi} + \sum_{j\in\mathbf{J}} b_{Sj} \le B \tag{21b}$$

$$b_{Pi} \in \mathbf{Z}^{+},$$

$$b_{Sj} \in \mathbf{Z}^{+}, \tag{21c}$$

$$\forall i \in \mathbf{I},\ \forall j \in \mathbf{J}$$

$$p_{Pi} \le P_{Pi},$$

$$p_{Sj} \le P_{Sj}, \tag{21d}$$

$$\forall i \in \mathbf{I},\ \forall j \in \mathbf{J}$$

$$p_{Pi} \ge 0,$$

$$p_{Sj} \ge 0, \tag{21e}$$

$$\forall i \in \mathbf{I},\ \forall j \in \mathbf{J}$$

$$\Omega \ge 0,$$

$$\Phi \ge 0, \tag{21f}$$

$$f_i^1(\mathbf{b}_P, \mathbf{p}_P) = r_{Pi}(\mathbf{b}_P, \mathbf{p}_P) - (q_{Pi} + a_{Pi}) \le 0, \tag{21g}$$

$$\forall i \in \mathbf{I},$$

$$f_i^2(\mathbf{b}_P, \mathbf{p}_P) = (q_{Pi} + a_{Pi} - L_i) - r_{Pi}(\mathbf{b}_P, \mathbf{p}_P) \le 0, \tag{21h}$$

$$\forall i \in \mathbf{I},$$

$$f_i^3(\mathbf{b}_P, \mathbf{p}_P) = (q_{Pi} + a_{Pi}) - r_{Pi}(\mathbf{b}_P, \mathbf{p}_P) - \Omega \le 0, \tag{21i}$$

$$\forall i \in \mathbf{I},$$

$$f_j^4(\mathbf{b}_S, \mathbf{p}_S) = \left(q_{Sj} + a_{Sj}\right) - r_{Sj}(\mathbf{b}_S, \mathbf{p}_S) - \Phi \le 0, \tag{21j}$$

$$\forall j \in \mathbf{J}.$$

In above problem, some of the optimization variables (particularly the components of $\mathbf{b}_P$ and $\mathbf{b}_S$) can take only integer values, whereas the other variables (the components of $\mathbf{p}_P$ and $\mathbf{p}_S$) are real-valued ones. In addition, constraints (21g)–(21j) are represented by the nonsmooth nonlinear functions of ($\mathbf{b}_P$, $\mathbf{p}_P$) and ($\mathbf{b}_P$, $\mathbf{p}_P$). Hence, (21a)–(21j) is a nonlinear mixed integer programming (MINLP) problem, which is Nondeterministic Polynomial-time (NP) hard. For immediate proof of NP-hardness, note that MINLP includes mixed integer linear programming (MILP) problem, which is NP-hard [23].

Before applying any MINLP method for solving (21a)–(21j), the nonsmooth function $g(\cdot)$ in (19), included in the expressions of $r_{Pi}(\mathbf{b}_P, \mathbf{p}_P)$ and $r_{Sj}(\mathbf{b}_P, \mathbf{p}_P)$, should be replaced by its smooth approximation. To construct a smooth approximation of $g(x)$, note that this function is equivalent to the sum of the shifted and scaled versions of a Heaviside step function $H(x)$ [24]. That is,

$$g(x) = \sum_{l=1}^{15} (\zeta_l - \zeta_{l-1}) H(x - \gamma_l),$$
$$H(x) = \begin{cases} 1, & \text{if } x \ge 0, \\ 0, & \text{otherwise,} \end{cases} \tag{22}$$

where $\zeta_0 = 0$.

Recall that a smooth approximation for a step function $H(x)$ is given by a logistic sigmoid function [25]:

$$\widehat{H}_q(x) = \frac{1}{1 + e^{-2qx}}, \tag{23}$$

where $q > 0$; x is in range of real numbers from $-\infty$ to $+\infty$. If we take $H(0) = 1/2$, then larger q corresponds to a closer transition to $H(x)$; that is,

$$\lim_{q\to\infty} \widehat{H}_q(x) = H(x). \tag{24}$$

The above holds, because, for $x < 0$, we have

$$e^{-2qx} \longrightarrow +\infty,$$
$$\widehat{H}_q(x) \approx H(x) = 1; \tag{25}$$

for $x > 0$,

$$e^{-2qx} \longrightarrow 0,$$
$$\widehat{H}_q(x) \approx H(x) = 1; \tag{26}$$

for $x = 0$,

$$e^{-2qx} = 1,$$
$$\widehat{H}_q(x) = H(x) = \frac{1}{2}. \tag{27}$$

Consequently, an approximation for a shifted Heaviside function is represented by a shifted logistic function

$$\widehat{H}_q(x - \gamma_l) = \frac{1}{1 + e^{-2q(x-\gamma_l)}}, \tag{28}$$

defined for $q > 0$, with real x in range from $-\infty$ to $+\infty$.

Based on (28), we can construct a smooth approximation for $g(x)$ as

$$\widehat{g}_q(x) = \sum_{l=1}^{15} (\zeta_l - \zeta_{l-1}) \widehat{H}_q(x - \gamma_l) = \sum_{l=1}^{15} \frac{\zeta_l - \zeta_{l-1}}{1 + e^{2q(x-\gamma_l)}}. \quad (29)$$

Then, it is rather straightforward to verify that

$$\lim_{q\to\infty} \widehat{g}_q(x) = g(x), \quad (30)$$

and the approximations for $r_{Pi}(\mathbf{b}_P, \mathbf{p}_P)$ and $r_{Sj}(\mathbf{b}_P, \mathbf{p}_P)$ will take the form

$$\widehat{r}_{Pi}(\mathbf{b}_P, \mathbf{p}_P) = \omega\psi b_{Pi} \log\left(1 + \widehat{g}(p_{Pi}) \cdot h_{Pi} p_{Pi}\right), \quad \forall i \in \mathbf{I}; \quad (31a)$$

$$\widehat{r}_{Sj}(\mathbf{b}_S, \mathbf{p}_S) = \omega\psi b_{Sj} \log\left(1 + \widehat{g}\left(p_{Sj}\right) h_{Sj} p_{Sj}\right), \quad \forall j \in \mathbf{J}. \quad (31b)$$

With the approximations given by (31a) and (31b), problem (21a)–(21j) can be solved using a sequential optimization approach as

$$\left(\mathbf{b}_P, \mathbf{p}_P, \mathbf{b}_S, \mathbf{p}_S\right) = \arg\min \quad (\Omega + \Phi) \quad (32a)$$

$$\text{subject to:} \quad \sum_{i\in\mathbf{I}} b_{Pi} + \sum_{j\in\mathbf{J}} b_{Sj} \le B \quad (32b)$$

$$b_{Pi} \in \mathbf{Z}^+, \quad b_{Sj} \in \mathbf{Z}^+, \quad \forall i \in \mathbf{I},\ \forall j \in \mathbf{J}, \quad (32c)$$

$$p_{Pi} \le P_{Pi}, \quad p_{Sj} \le P_{Sj}, \quad \forall i \in \mathbf{I},\ \forall j \in \mathbf{J}, \quad (32d)$$

$$p_{Pi} \ge 0, \quad p_{Sj} \ge 0, \quad \forall i \in \mathbf{I},\ \forall j \in \mathbf{J}, \quad (32e)$$

$$\Omega \ge 0, \quad \Phi \ge 0, \quad (32f)$$

$$\widehat{f}_i^1(\mathbf{b}_P, \mathbf{p}_P) = \widehat{r}_{Pi}(\mathbf{b}_P, \mathbf{p}_P) - \left(q_{Pi} + a_{Pi}\right) \le 0, \quad \forall i \in \mathbf{I}, \quad (32g)$$

$$\widehat{f}_i^2(\mathbf{b}_P, \mathbf{p}_P) = \left(q_{Pi} + a_{Pi} - L_i\right) - \widehat{r}_{Pi}(\mathbf{b}_P, \mathbf{p}_P) \le 0, \quad \forall i \in \mathbf{I}, \quad (32h)$$

$$\widehat{f}_i^3(\mathbf{b}_P, \mathbf{p}_P) = \left(q_{Pi} + a_{Pi}\right) - \widehat{r}_{Pi}(\mathbf{b}_P, \mathbf{p}_P) - \Omega \le 0, \quad \forall i \in \mathbf{I}, \quad (32i)$$

$$\widehat{f}_j^4(\mathbf{b}_S, \mathbf{p}_S) = \left(q_{Sj} + a_{Sj}\right) - \widehat{r}_{Sj}(\mathbf{b}_S, \mathbf{p}_S) - \Phi \le 0, \quad \forall j \in \mathbf{J}. \quad (32j)$$

The above problem is smooth MINLP, which can be solved using some fast and effective MINLP method. Note that most of the MINLP techniques involve construction of the following relaxations to the considered problem: the nonlinear programming (NLP) relaxation (original problem without integer restrictions) and the MILP relaxation (original problem where nonlinearities are replaced by supporting hyperplanes). In our case, a smooth MILP relaxation to (32a)–(32j) in a given point $(\mathbf{b}_P^0, \mathbf{p}_P^0, \mathbf{b}_S^0, \mathbf{p}_S^0)$ is given by

$$\left(\mathbf{b}_P, \mathbf{p}_P, \mathbf{b}_S, \mathbf{p}_S\right) = \arg\min \quad (\Omega + \Phi) \quad (33a)$$

$$\text{subject to:} \quad \sum_{i\in\mathbf{I}} b_{Pi} + \sum_{j\in\mathbf{J}} b_{Sj} \le B \quad (33b)$$

$$b_{Pi} \in \mathbf{Z}^+, \quad b_{Sj} \in \mathbf{Z}^+, \quad \forall i \in \mathbf{I},\ \forall j \in \mathbf{J}, \quad (33c)$$

$$
\begin{aligned}
& p_{Pi} \le P_{Pi}, \\
& p_{Sj} \le P_{Sj}, \\
& \forall i \in \mathbf{I},\ \forall j \in \mathbf{J},
\end{aligned} \tag{33d}
$$

$$
\begin{aligned}
& p_{Pi} \ge 0, \\
& p_{Sj} \ge 0, \\
& \forall i \in \mathbf{I},\ \forall j \in \mathbf{J},
\end{aligned} \tag{33e}
$$

$$
\begin{aligned}
& \Omega \ge 0, \\
& \Phi \ge 0,
\end{aligned} \tag{33f}
$$

$$
\widehat{f}_i^{\,1}\left(\mathbf{b}_P^0, \mathbf{p}_P^0\right) + \nabla \widehat{f}_i^{\,1}\left(\mathbf{b}_P^0, \mathbf{p}_P^0\right)^T \begin{bmatrix} \mathbf{b}_P - \mathbf{b}_P^0 \\ \mathbf{p}_P - \mathbf{p}_P^0 \end{bmatrix} \le 0, \tag{33g}
$$

$$
\forall i \in \mathbf{I},
$$

$$
\widehat{f}_i^{\,2}\left(\mathbf{b}_P^0, \mathbf{p}_P^0\right) + \nabla \widehat{f}_i^{\,2}\left(\mathbf{b}_P^0, \mathbf{p}_P^0\right)^T \begin{bmatrix} \mathbf{b}_P - \mathbf{b}_P^0 \\ \mathbf{p}_P - \mathbf{p}_P^0 \end{bmatrix} \le 0, \tag{33h}
$$

$$
\forall i \in \mathbf{I},
$$

$$
\widehat{f}_i^{\,3}\left(\mathbf{b}_P^0, \mathbf{p}_P^0\right) + \nabla \widehat{f}_i^{\,3}\left(\mathbf{b}_P^0, \mathbf{p}_P^0\right)^T \begin{bmatrix} \mathbf{b}_P - \mathbf{b}_P^0 \\ \mathbf{p}_P - \mathbf{p}_P^0 \end{bmatrix} \le 0, \tag{33i}
$$

$$
\forall i \in \mathbf{I},
$$

$$
\begin{aligned}
& \widehat{f}_j^{\,4}\left(\mathbf{b}_S^0, \mathbf{p}_S^0\right) + \nabla \widehat{f}_j^{\,4}\left(\mathbf{b}_S^0, \mathbf{p}_S^0\right)^T \begin{bmatrix} \mathbf{b}_S - \mathbf{b}_S^0 \\ \mathbf{p}_S - \mathbf{p}_S^0 \end{bmatrix} \\
& \le 0, \\
& \forall j \in \mathbf{J}.
\end{aligned} \tag{33j}
$$

The NLP relaxation to (32a)–(32j) is

$$
\left(\mathbf{b}_P, \mathbf{p}_P, \mathbf{b}_S, \mathbf{p}_S\right) = \arg\min \quad (\Omega + \Phi) \tag{34a}
$$

$$
\text{subject to:} \quad \sum_{i \in \mathbf{I}} b_{Pi} + \sum_{j \in \mathbf{J}} b_{Sj} \le B \tag{34b}
$$

$$
\begin{aligned}
& b_{Pi} \ge 0, \\
& b_{Sj} \ge 0, \\
& \forall i \in \mathbf{I},\ \forall j \in \mathbf{J},
\end{aligned} \tag{34c}
$$

$$
\begin{aligned}
& p_{Pi} \le P_{Pi}, \\
& p_{Sj} \le P_{Sj}, \\
& \forall i \in \mathbf{I},\ \forall j \in \mathbf{J},
\end{aligned} \tag{34d}
$$

$$
\begin{aligned}
& p_{Pi} \ge 0, \\
& p_{Sj} \ge 0, \\
& \forall i \in \mathbf{I},\ \forall j \in \mathbf{J},
\end{aligned} \tag{34e}
$$

$$
\begin{aligned}
& \Omega \ge 0, \\
& \Phi \ge 0,
\end{aligned} \tag{34f}
$$

$$
\begin{aligned}
& \widehat{f}_i^{\,1}\left(\mathbf{b}_P, \mathbf{p}_P\right) \le 0, \\
& \forall i \in \mathbf{I},
\end{aligned} \tag{34g}
$$

$$
\begin{aligned}
& \widehat{f}_i^{\,2}\left(\mathbf{b}_P, \mathbf{p}_P\right) \le 0, \\
& \forall i \in \mathbf{I},
\end{aligned} \tag{34h}
$$

$$
\begin{aligned}
& \widehat{f}_i^{\,3}\left(\mathbf{b}_P, \mathbf{p}_P\right) \le 0, \\
& \forall i \in \mathbf{I},
\end{aligned} \tag{34i}
$$

$$
\begin{aligned}
& \widehat{f}_j^{\,4}\left(\mathbf{b}_S, \mathbf{p}_S\right) \le 0, \\
& \forall j \in \mathbf{J}.
\end{aligned} \tag{34j}
$$

Note that the MINLP problems can be solved using either deterministic technique or an approximation. A typical exact

method for solving MINLPs is a well-known branch-and-bound algorithm and its various modifications [26]; examples of heuristic approaches include local branching [27], large neighborhood search [28], and feasibility pump [29], to name a few. Since we are interested in a reasonably simple and fast algorithm, it is more convenient to use heuristics for solving the problem. In this paper, a Feasibility Pump (FP) heuristic [29, 30] is applied to solve (32a)–(32j). FP is perhaps the most simple and most effective method for producing more and better solutions in a shorter average running time. For the problems with nonbinary integer variables, the complexity of FP is exponential in size of a problem [31].

A fundamental idea of FP algorithm is to decompose the problem into two parts: integer feasibility and constraint feasibility. The former is achieved by rounding (solving a NLP relaxation to the original problem) and the latter by projection (solving a MILP relaxation). Consequently, two sequences of points are generated. The first sequence $\{(\overline{\mathbf{b}_P, \mathbf{p}_P, \mathbf{b}_S, \mathbf{p}_S})_k\}_{k=1}^K$, $K = 1, 2, \ldots$, contains the integral points that may violate the nonlinear constraints. The second sequence $\{(\underline{\mathbf{b}_P, \mathbf{p}_P, \mathbf{b}_S, \mathbf{p}_S})_k\}_{k=1}^K$ contains the points which are feasible for a continuous relaxation to the original problem but might not be integral.

More specifically, with input $(\underline{\mathbf{b}_P, \mathbf{p}_P, \mathbf{b}_S, \mathbf{p}_S})_1$ being a solution to the NLP relaxation given by (34a)–(34j), the algorithm generates two sequences by solving the following problems for $k = 1, \ldots, K$:

$$\left(\overline{\mathbf{b}_P, \mathbf{p}_P, \mathbf{b}_S, \mathbf{p}_S}\right)_k = \arg\min \quad \left\|(\mathbf{b}_P, \mathbf{p}_P, \mathbf{b}_S, \mathbf{p}_S) - \left(\underline{\mathbf{b}_P, \mathbf{p}_P, \mathbf{b}_S, \mathbf{p}_S}\right)_k\right\|_1 \tag{35a}$$

$$\text{subject to:} \quad \sum_{i \in \mathbf{I}} b_{Pi} + \sum_{j \in \mathbf{J}} b_{Sj} \leq B \tag{35b}$$

$$b_{Pi} \in \mathbf{Z}^+,$$

$$b_{Sj} \in \mathbf{Z}^+, \tag{35c}$$

$$\forall i \in \mathbf{I}, \ \forall j \in \mathbf{J},$$

$$p_{Pi} \leq P_{Pi},$$

$$p_{Sj} \leq P_{Sj}, \tag{35d}$$

$$\forall i \in \mathbf{I}, \ \forall j \in \mathbf{J},$$

$$p_{Pi} \geq 0,$$

$$p_{Sj} \geq 0, \tag{35e}$$

$$\forall i \in \mathbf{I}, \ \forall j \in \mathbf{J},$$

$$\Omega \geq 0,$$

$$\Phi \geq 0, \tag{35f}$$

$$\widehat{f}_i^1\left(\mathbf{b}_P^0, \mathbf{p}_P^0\right) + \nabla \widehat{f}_i^1\left(\mathbf{b}_P^0, \mathbf{p}_P^0\right)^T \begin{bmatrix} \mathbf{b}_P - \mathbf{b}_P^0 \\ \mathbf{p}_P - \mathbf{p}_P^0 \end{bmatrix} \leq 0, \tag{35g}$$

$$\forall i \in \mathbf{I},$$

$$\widehat{f}_i^2\left(\mathbf{b}_P^0, \mathbf{p}_P^0\right) + \nabla \widehat{f}_i^2\left(\mathbf{b}_P^0, \mathbf{p}_P^0\right)^T \begin{bmatrix} \mathbf{b}_P - \mathbf{b}_P^0 \\ \mathbf{p}_P - \mathbf{p}_P^0 \end{bmatrix} \leq 0, \tag{35h}$$

$$\forall i \in \mathbf{I},$$

$$\widehat{f}_i^3\left(\mathbf{b}_P^0, \mathbf{p}_P^0\right) + \nabla \widehat{f}_i^3\left(\mathbf{b}_P^0, \mathbf{p}_P^0\right)^T \begin{bmatrix} \mathbf{b}_P - \mathbf{b}_P^0 \\ \mathbf{p}_P - \mathbf{p}_P^0 \end{bmatrix} \leq 0, \tag{35i}$$

$$\forall i \in \mathbf{I},$$

$$\widehat{f}_j^4\left(\mathbf{b}_S^0, \mathbf{p}_S^0\right) + \nabla \widehat{f}_j^4\left(\mathbf{b}_S^0, \mathbf{p}_S^0\right)^T \begin{bmatrix} \mathbf{b}_S - \mathbf{b}_S^0 \\ \mathbf{p}_S - \mathbf{p}_S^0 \end{bmatrix} \leq 0, \quad \forall j \in \mathbf{J}; \tag{35j}$$

$$\left(\underline{\mathbf{b}_P, \mathbf{p}_P, \mathbf{b}_S, \mathbf{p}_S}\right)_{k+1} = \arg\min \quad \left\|(\mathbf{b}_P, \mathbf{p}_P, \mathbf{b}_S, \mathbf{p}_S) - \left(\overline{\mathbf{b}_P, \mathbf{p}_P, \mathbf{b}_S, \mathbf{p}_S}\right)_k\right\|_2 \tag{36a}$$

$$\text{subject to:} \quad \sum_{i\in\mathbf{I}} b_{Pi} + \sum_{j\in\mathbf{J}} b_{Sj} \leq B \tag{36b}$$

$$b_{Pi} \geq 0, \quad b_{Sj} \geq 0, \quad \forall i \in \mathbf{I},\ \forall j \in \mathbf{J}, \tag{36c}$$

$$p_{Pi} \leq P_{Pi}, \quad p_{Sj} \leq P_{Sj}, \quad \forall i \in \mathbf{I},\ \forall j \in \mathbf{J}, \tag{36d}$$

$$p_{Pi} \geq 0, \quad p_{Sj} \geq 0, \quad \forall i \in \mathbf{I},\ \forall j \in \mathbf{J}, \tag{36e}$$

$$\Omega \geq 0, \quad \Phi \geq 0, \tag{36f}$$

$$\widehat{f}_i^1\left(\mathbf{b}_P, \mathbf{p}_P\right) \leq 0, \quad \forall i \in \mathbf{I}, \tag{36g}$$

$$\widehat{f}_i^2\left(\mathbf{b}_P, \mathbf{p}_P\right) \leq 0, \quad \forall i \in \mathbf{I}, \tag{36h}$$

$$\widehat{f}_i^3\left(\mathbf{b}_P, \mathbf{p}_P\right) \leq 0, \quad \forall i \in \mathbf{I}, \tag{36i}$$

$$\widehat{f}_j^4\left(\mathbf{b}_S, \mathbf{p}_S\right) \leq 0, \quad \forall j \in \mathbf{J}, \tag{36j}$$

where $\|\cdot\|_1$ and $\|\cdot\|_2$ are l_1 norm and l_2 norm, respectively. The rounding step is carried out by solving (35a)–(35j), whereas the projection is the solution to (36a)–(36j). A suggested FP algorithm alternates between the rounding and projection steps until $(\overline{\mathbf{b}_P, \mathbf{p}_P, \mathbf{b}_S, \mathbf{p}_S})_k = (\underline{\mathbf{b}_P, \mathbf{p}_P, \mathbf{b}_S, \mathbf{p}_S})_k$ (which implies feasibility) or the number of iterations k has reached the predefined limit K. The workflow of a corresponding FP algorithm is illustrated in the following part.

FP Algorithm for Non-Convex MINLP

(0) INITIALIZATION: input K; set $k := 1$;

solve (34a)–(34j) to obtain $(\underline{\mathbf{b}_P, \mathbf{p}_P, \mathbf{b}_S, \mathbf{p}_S})_k$;

(1) WHILE ($k < K$) do:

{

(2) ROUNDING: solve (35a)–(35j) to obtain $(\overline{\mathbf{b}_P, \mathbf{p}_P, \mathbf{b}_S, \mathbf{p}_S})_k$;

(3) IF $((\overline{\mathbf{b}_P, \mathbf{p}_P, \mathbf{b}_S, \mathbf{p}_S})_k = (\overline{\mathbf{b}_P, \mathbf{p}_P, \mathbf{b}_S, \mathbf{p}_S})_k)$ THEN goto step (6);

(4) PROJECTION: solve (36a)–(36j) to obtain $(\underline{\mathbf{b}_P, \mathbf{p}_P, \mathbf{b}_S, \mathbf{p}_S})_{k+1}$;

(5) Set $k := k + 1$;

}

(6) OUTPUT: solution $(\mathbf{b}_P, \mathbf{p}_P, \mathbf{b}_S, \mathbf{p}_S)^* := (\overline{\mathbf{b}_P, \mathbf{p}_P, \mathbf{b}_S, \mathbf{p}_S})_k$.

Note that in order to retain the convergence of the algorithm, both problems (35a)–(35j) and (36a)–(36j) need to be solved exactly. Problem (36a)–(36j) (and consequently problem (34a)–(34j)) can be solved using any standard NLP method. In this paper, an interior point algorithm (described, e.g., in [32]) with polynomial complexity is applied to solve (36a)–(36j) and (34a)–(34j). MILP problem (35a)–(35j) is relatively simple and therefore can be solved efficiently for optimality by any technique from the family of the branch-and-bound methods (e.g., [26]).

3.3. Target Buffer Size for PUs and Feasibility Issues. Recall that L_i represents the target buffer size for PU_i, which indicates that the QoS requirements of PU_i are satisfied. By limiting the buffer size of PUs, we restrict the QoS of PUs (measured in terms of their buffer size) to be higher (or at least not smaller) than the required minimum. From this point of view, constraints (5) and (12) guarantee the QoS protection of PUs in Scenario 1 and Scenario 2, respectively. Apparently, one can always restrict the target buffer size of PUs to be equal to zero, so that after each subsequent data arrival the buffer of each PU is fully cleared. This approach represents a "greedy" policy which ensures that the PUs are always served with the best possible QoS but does not guarantee that some spare capacity will be left to serve the SUs.

A "less greedy" strategy would be to establish some lower bound on QoS for the PUs, which will provide an "opportunity" for SUs to be served. To establish this lower bound, recall that the growing delay and loss indicate that a buffer of a corresponding node is congested, whereas in an uncongested node the packet delay and loss are always minimal [33]. On the other hand, it has been reported in [33, 34] that in an uncongested node the buffer size $q(t)$ is increasing very slowly, that is, $q(t+1)/q(t) \approx 1$, whereas in a congested node the buffer size is growing very fast, so that $q(t+1)/q(t) \gg 1$. Hence, to serve the PUs with minimal delay and loss, it is enough to prevent the growth of their buffers; that is, for each PU set

$$L_i = q_{Pi}(t) + \varepsilon, \quad \forall i \in \mathbf{I}, \tag{37a}$$

where $\varepsilon \geq 0$ is some small value, such that

$$\varepsilon \ll q_{Pi}(t), \quad \forall i \in \mathbf{I}. \tag{37b}$$

More sophisticated approach to find the lower QoS bound can be deployed if information about the past buffer size is available at any time slot t. In this case, one can set the target buffer size based on past T observations $q(t), \ldots, q(t-T+1)$ as

$$L_i = \min_{\tau \in \mathbf{T}} q_{Pi}(t-\tau+1) + \varepsilon, \quad \forall i \in \mathbf{I}, \tag{38a}$$

or

$$L_i = \frac{1}{T} \sum_{\tau \in \mathbf{T}} q_{Pi}(t-\tau+1) + \varepsilon, \quad \forall i \in \mathbf{I}, \tag{38b}$$

where $\mathbf{T} = \{1, \ldots, T\}$.

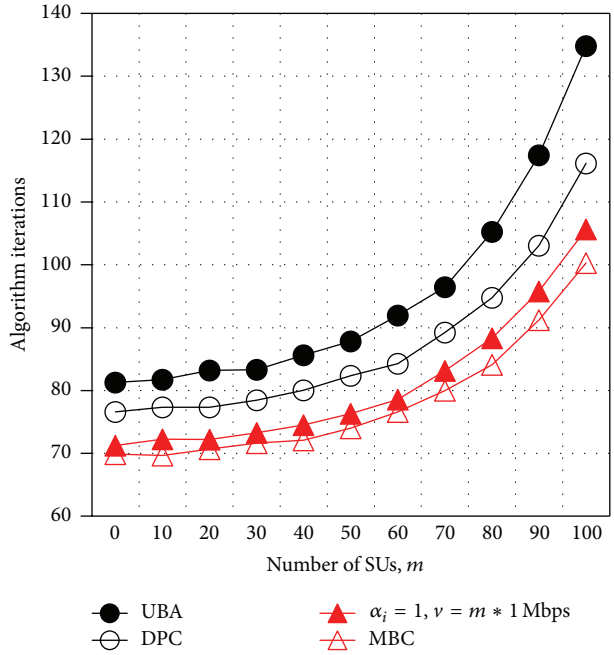

Figure 1: Average number of iterations before convergence as a function of number of SUs in simulations with $n = 100$ PUs and average SNR = 0 dB.

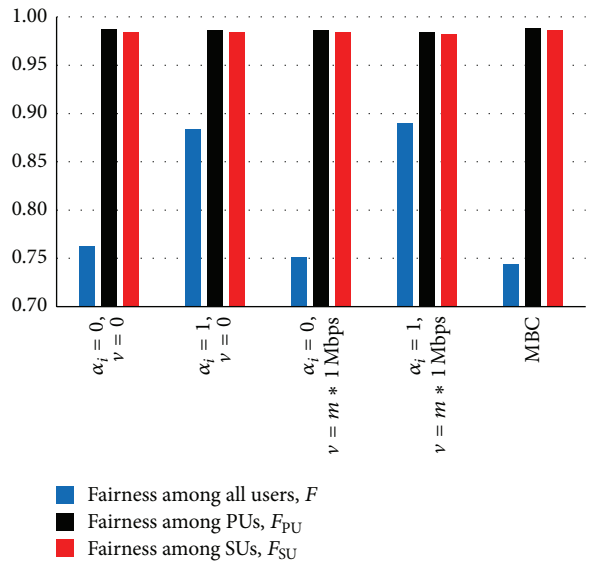

Figure 2: Jain's index of fairness F_{SU}, F_{PU}, and F for the PUs, SUs, and all users, in simulations with $n = 100$ PUs, $m = 100$ SUs, and average SNR = 0 dB.

In some (very rare) cases (when the service capacity of the eNB is not enough to guarantee that the buffer sizes in all PUs do not exceed the target buffer size), constraints (5) and (12) in Scenarios 1 and 2 may be not feasible. In the following, we show how to identify such situation and what should be the optimal resource allocation strategy in this case.

Note that the service capacity of LTE system greatly depends on physical characteristics (e.g., SINR and MCS) of

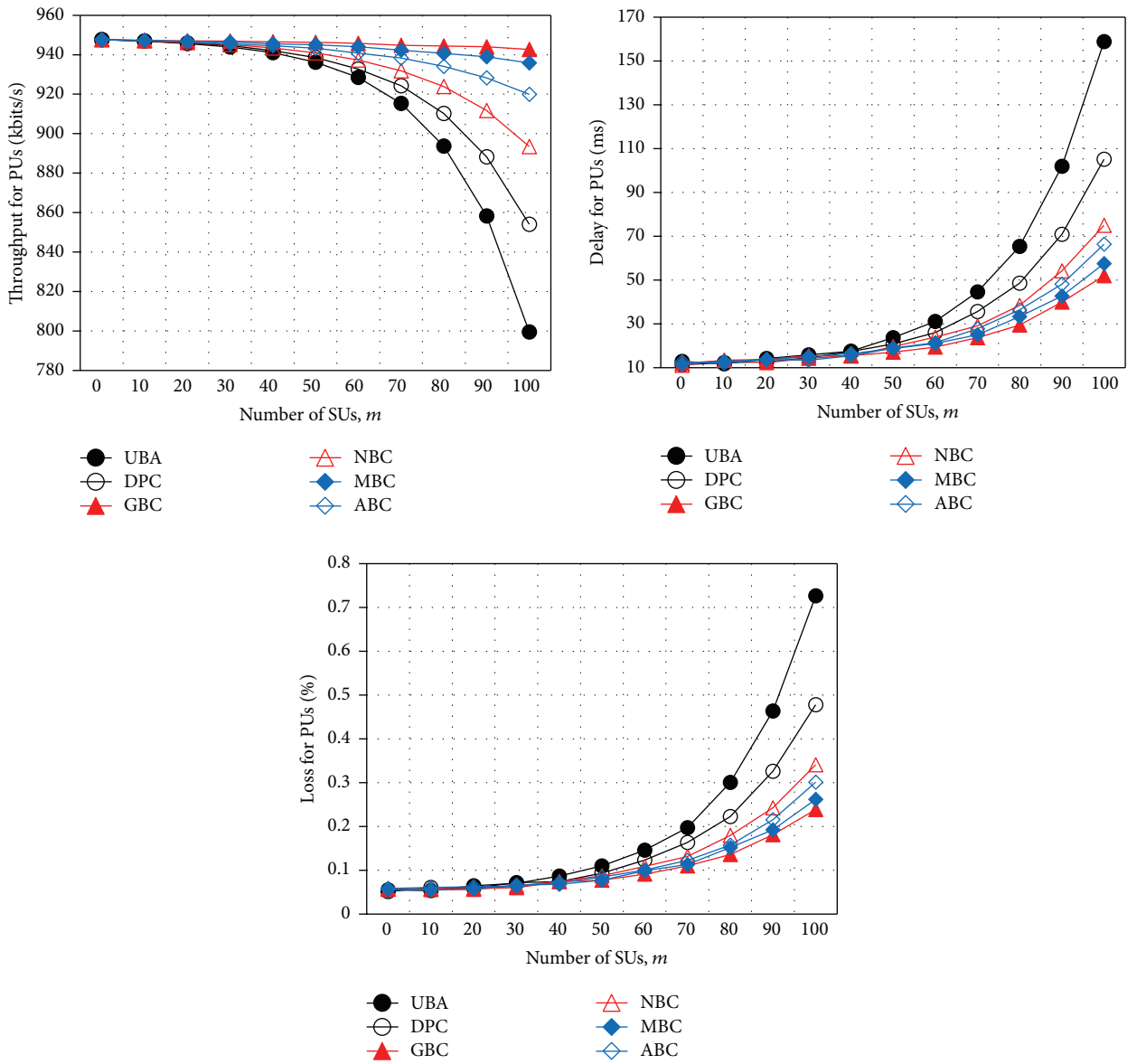

FIGURE 3: Mean throughput, packet delay, and loss for PUs as a function of number of SUs in simulations with $n = 100$ PUs and average SNR = 0 dB.

the wireless channels between the eNB and the users. Hence, at the moment preceding the bandwidth/power allocation, it is rather difficult to estimate the service capacity of a system. Nevertheless, it is possible to find the average service capacity and to determine the upper and lower bounds on service capacity of the eNB in the UL and DL directions, such that

$$R_{\min} \le \sum_{i \in \mathbf{I}} r_{Pi}(t) \approx R_{\text{ave}} \le R_{\max}, \tag{39}$$

where $R_{\min}$, R_{ave}, and $R_{\max}$ are the lower bound, upper bound, and average service capacity of the eNB, respectively.

In our case, the upper bound on service capacity of the eNB is given by

$$R_{\max} = \omega \psi B \log\left(1 + g\left(\text{SNR}_{\max}\right) \cdot \text{SNR}_{\max}\right), \tag{40a}$$

where $\text{SNR}_{\max}$ is the maximal possible SNR value.

The lower bound on service capacity of eNB is

$$R_{\min} = \omega \psi B \log\left(1 + g\left(\text{SNR}_{\min}\right) \cdot \text{SNR}_{\min}\right), \tag{40b}$$

where $\text{SNR}_{\min}$ is the minimal possible SNR value.

The average service capacity of the eNB equals

$$R_{\text{ave}} = \omega \psi B \log\left(1 + g\left(\text{SNR}_{\text{ave}}\right) \cdot \text{SNR}_{\text{ave}}\right), \tag{40c}$$

where SNR_{ave} is the average SNR value.

If SNR information is available, then the values $\text{SNR}_{\max}$, $\text{SNR}_{\min}$, and SNR_{ave} can be set as

$$\text{SNR}_{\max} = \max_{i \in \mathbf{I}} \text{SNR}_i; \tag{41a}$$

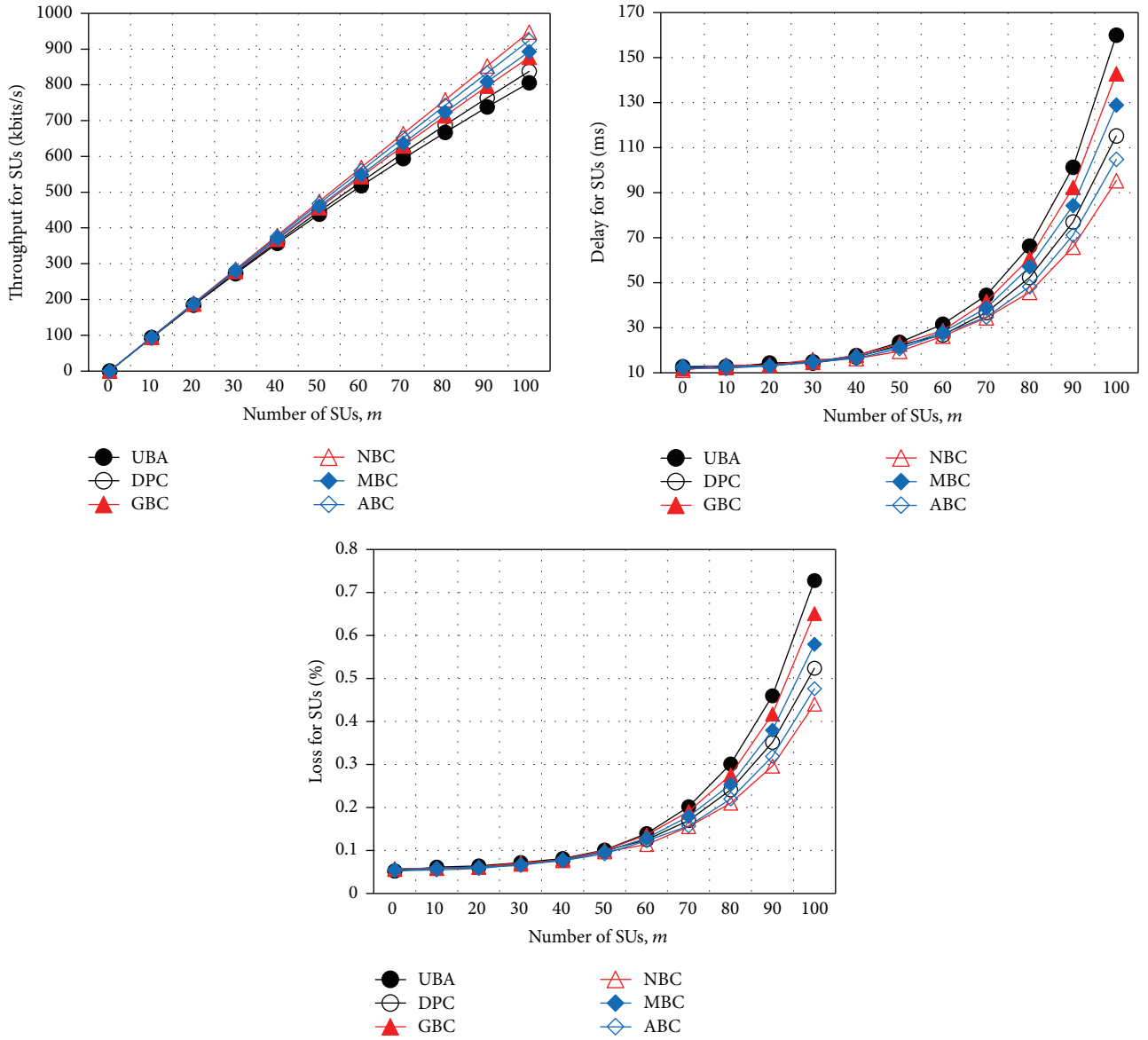

FIGURE 4: Mean throughput, packet delay, and loss for SUs as a function of number of SUs in simulations with $n = 100$ PUs and average SNR = 0 dB.

$$\mathrm{SNR}_{\min} = \min_{i \in \mathbf{I}} \mathrm{SNR}_i; \tag{41b}$$

$$\mathrm{SNR}_{\mathrm{ave}} = \frac{1}{n} \sum_{i \in \mathbf{I}} \mathrm{SNR}_i. \tag{41c}$$

Otherwise (i.e., if SNR information is not available), the values are set arbitrarily.

Now we can identify the situation when constraints (5) and (12) cannot be satisfied. For this, we should perform the following test. For Scenario 1, if

$$R_{\min} \geq \sum_{i \in \mathbf{I}} \left(q_{Pi}(t) + a_{Pi}(t) - L_i \right), \tag{42}$$

then constraint (5) is feasible. Otherwise, condition (5) cannot be satisfied. Hence, all the SUs receive zero allocations, and we assign the bandwidth/power resources only to the PUs by solving the following optimization problem:

$$\text{minimize} \quad \max_{i \in \mathbf{I}} \left\lceil q_{Pi} + a_{Pi} - r_{Pi}\left(\mathbf{b}_P, \mathbf{p}_P\right) \right\rceil^+ \tag{43a}$$

$$\text{subject to:} \quad r_{Pi}\left(\mathbf{b}_P, \mathbf{p}_P\right) \leq q_{Pi} + a_{Pi}, \quad \forall i \in \mathbf{I} \tag{43b}$$

$$\sum_{i \in \mathbf{I}} b_{Pi} \leq B \tag{43c}$$

$$b_{Pi} \in \mathbf{Z}^+, \quad \forall i \in \mathbf{I} \tag{43d}$$

$$p_{Pi} \leq P_{Pi}, \quad \forall i \in \mathbf{I} \tag{43e}$$

$$p_{Pi} \geq 0, \quad \forall i \in \mathbf{I} \tag{43f}$$

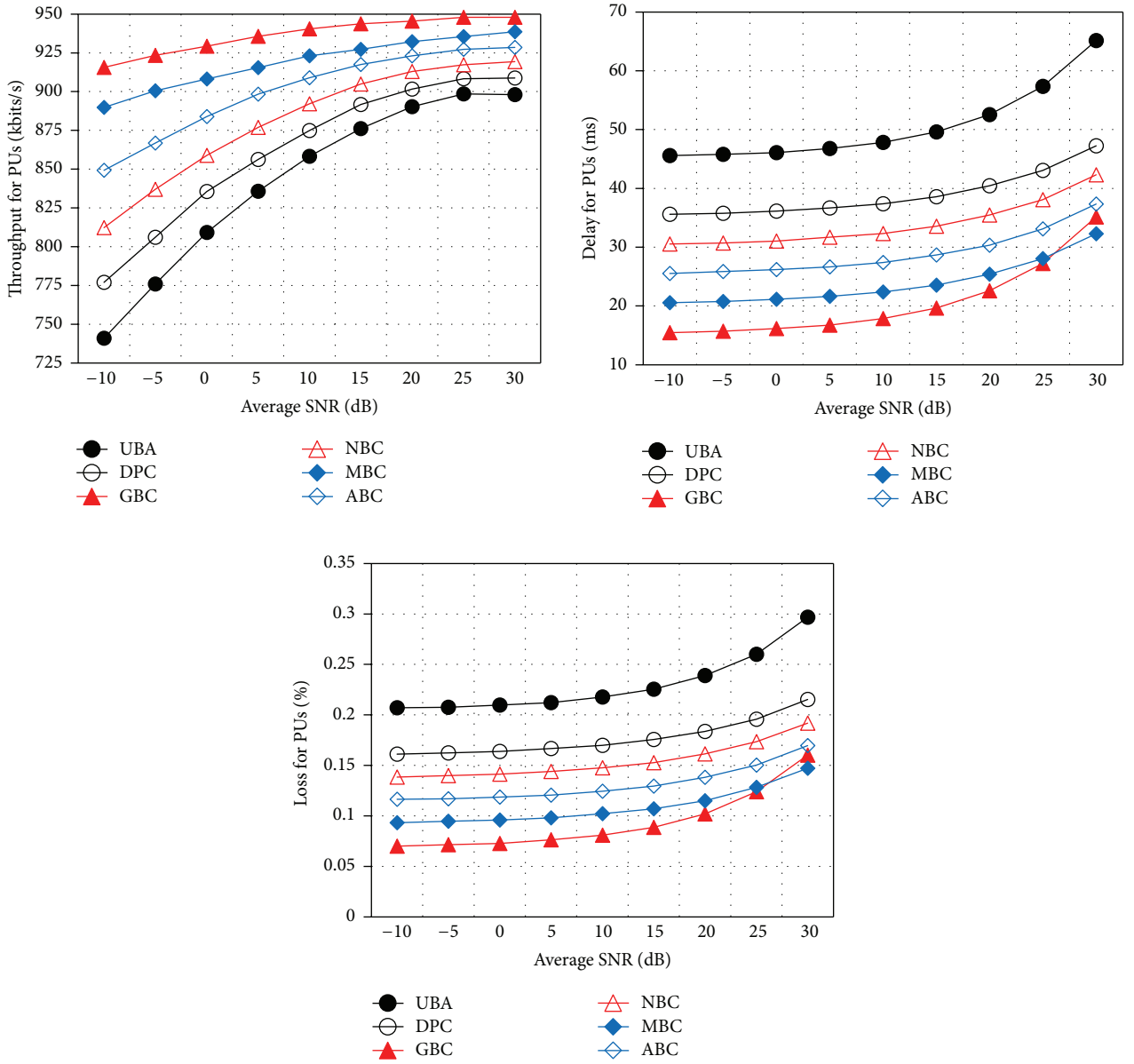

FIGURE 5: Mean throughput, packet delay, and loss for PUs as a function of SNR in simulations with $n = 100$ PUs and average SNR = 0 dB.

for the UL direction. For the DL directions, constraint (43e) should be replaced by

$$\sum_{i \in \mathbf{I}} p_{Pi} \leq P_{\text{eNB}}. \tag{43g}$$

For Scenario 2, if

$$R_{\min} \geq \sum_{i \in \mathbf{I}} \left(q_{Pi}(t) + a_{Pi}(t) - (1 + \alpha_i) L_i \right), \tag{44}$$

then constraint (12) is feasible. Otherwise, inequality (12) cannot be satisfied, and the bandwidth/power resources are assigned only to the PUs (according to (43a)–(43g)), whereas the SUs receive zero allocations.

4. Performance Evaluation

4.1. Simulation Model. A simulation model of the network consisting of one eNB, n PUs, and m SUs randomly located in a service area with 3 km radius has been developed upon a standard LTE platform using the OPNET package [35]. The bandwidth of the eNB equals $B = 50$ RBs (which is equivalent to 10 MHz). The slot duration in the algorithm equals $T_s =$ 1 ms. The maximal number of iterations in FP algorithm is set $K = 200$. The maximal transmission power of the eNB equals $P_{\text{eNB}} = 46$ dBm and the maximal transmission power of each PU/SU equals $P_{P1} = \cdots = P_{Pn} = P_{S1} = \cdots = P_{Sm} = 23$ dBm. Other simulation settings of the model are listed in Table 2 (all the parameters are based on LTE specifications [36]).

Table 2: Simulation parameters of the model.

Parameter	Value
Radio network model	
Pass loss	$L = 40\log_{10}R + 30\log_{10}f + 49$, R: distance (km), f: carrier frequency (Hz)
Shadow fading	Log-normal shadow fading with a standard deviation of 10/12 dB for outdoor/indoor users
Penetration loss	The average building penetration loss is 12 dB with a standard deviation of 8 dB
Multipath fading	Spatial Channel Model (SCM), Suburban macro
UE velocity	0 km/s
Transmitter/receiver antenna gain	10 dBi (pedestrian), 2 dBi (indoor)
Receiver antenna gain	10 dBi (pedestrian), 2 dBi (indoor)
Receiver noise figure	5 dB
Thermal noise density	−174 dBm/Hz
Cable/connector/combiner losses	2 dB
Physical profile	
Operation mode	FDD
Cyclic prefix type	Normal (7 symbols per slot)
Evolved packet core (EPC) bearer definitions	348 kbit/s (nonguaranteed bit rate)
Subcarrier spacing	15 kHz
Admission control parameters	
Physical downlink control channel (PDCCH) symbols per subframe	3
UL loading factor	1
DL loading factor	1
Inactive bearer timeout	20 sec
Buffer status report parameters	
Periodic timer	5 subframes
Retransmission timer	2560 subframes
L1/L2 control parameters	
Reserved size	2 RBs
Cyclic shifts	6
Starting RB preserver for Format 1 messages	0
Allocation periodicity	5 subframes
Random access (RA) parameters	
Number of preambles	64
Preamble format	Format 0 (1-subframe long)
Number of RA resources per frame	4
Preamble retransmission limit	5 subframes
RA response timer	5 subframes
Contention resolution timer	40 subframes
HARQ parameters	
Maximal number of retransmissions	3 (UL and DL)
HARQ retransmission timer	8 subframes (UL and DL)
Maximal number of HARQ processes	8 per UE (UL and DL)

A radio model of the network corresponds to the requirements of ITU-T Recommendation M.1225.

The algorithms proposed in the paper are denoted as follows:

(i) GBC (greedy buffer Control), where the target buffer size equals 0 for all the PUs;

(ii) NBC (nongreedy buffer control) with the target buffer size calculated according to (37a) and (37b);

(iii) MBC (Minimal Buffer Control) with $T = 10$ slots and the target buffer size given by (38a);

(iv) ABC (Average Buffer Control) with $T = 10$ slots and the target buffer size equaling (38b).

Two previously proposed schemes used to benchmark the performance of the algorithms proposed in this paper are

(i) UBA (Utility Based Allocation) presented in [11], where the network resources (bandwidth and transmission power) are allocated to the UL channels to maximize the total transmission rate of the users subject to the capacity and transmission power constraints,

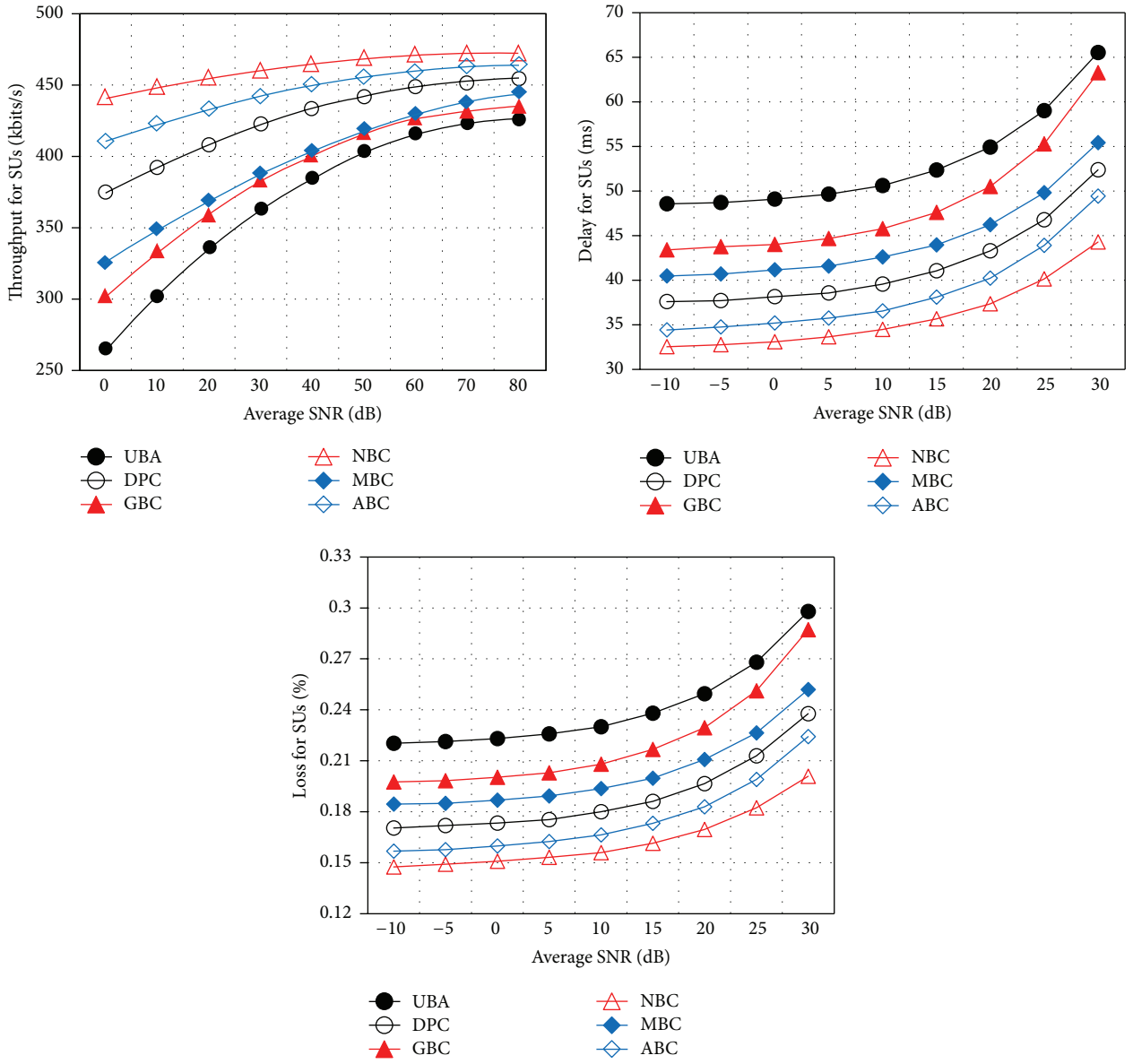

FIGURE 6: Mean throughput, packet delay, and loss for SUs as a function of SNR in simulations with $n = 100$ PUs and average SNR = 0 dB.

(ii) DPC (Delay Power Control) derived in [5], which is used to balance the transmission delay against transmission power in wireless channels between the eNB and the users.

We also gather simulation results for a proposed resource allocation approach in Scenario 2, with different settings of the parameters α_i and v, and benchmark its performance with the performance of the algorithm in Scenario 1.

All algorithms are simulated using identical settings. The data traffic in simulations is represented by voice over Internet Protocol (VoIP), video, and Hypertext Transfer Protocol (HTTP) applications mixed in proportion 2 : 3 : 5 using the models defined in [37].

4.2. Complexity and Fairness Issues. Let us first evaluate the complexity and the fairness of resource allocation in different algorithms. Figure 1 shows the average number of iterations before convergence for UBA, DPC, and the proposed algorithms in Scenario 1 (denoted as MBC) and Scenario 2, gathered in simulations with $n = 100$ PUs and average SNR = 0 dB. Note that the target buffer size for the PUs is calculated according to (38a) with $T = 10$ slots (in Scenarios 1 and 2) and $\alpha_i = 1$, $v = m * 1$ Mbps (for Scenario 2). Results demonstrate that the complexity of FP method deployed for resource allocation in Scenarios 1 and 2 is exponential in size of the problem (which corresponds to the findings reported in [31]) and is comparable to the complexities of other techniques (UBA and DPC).

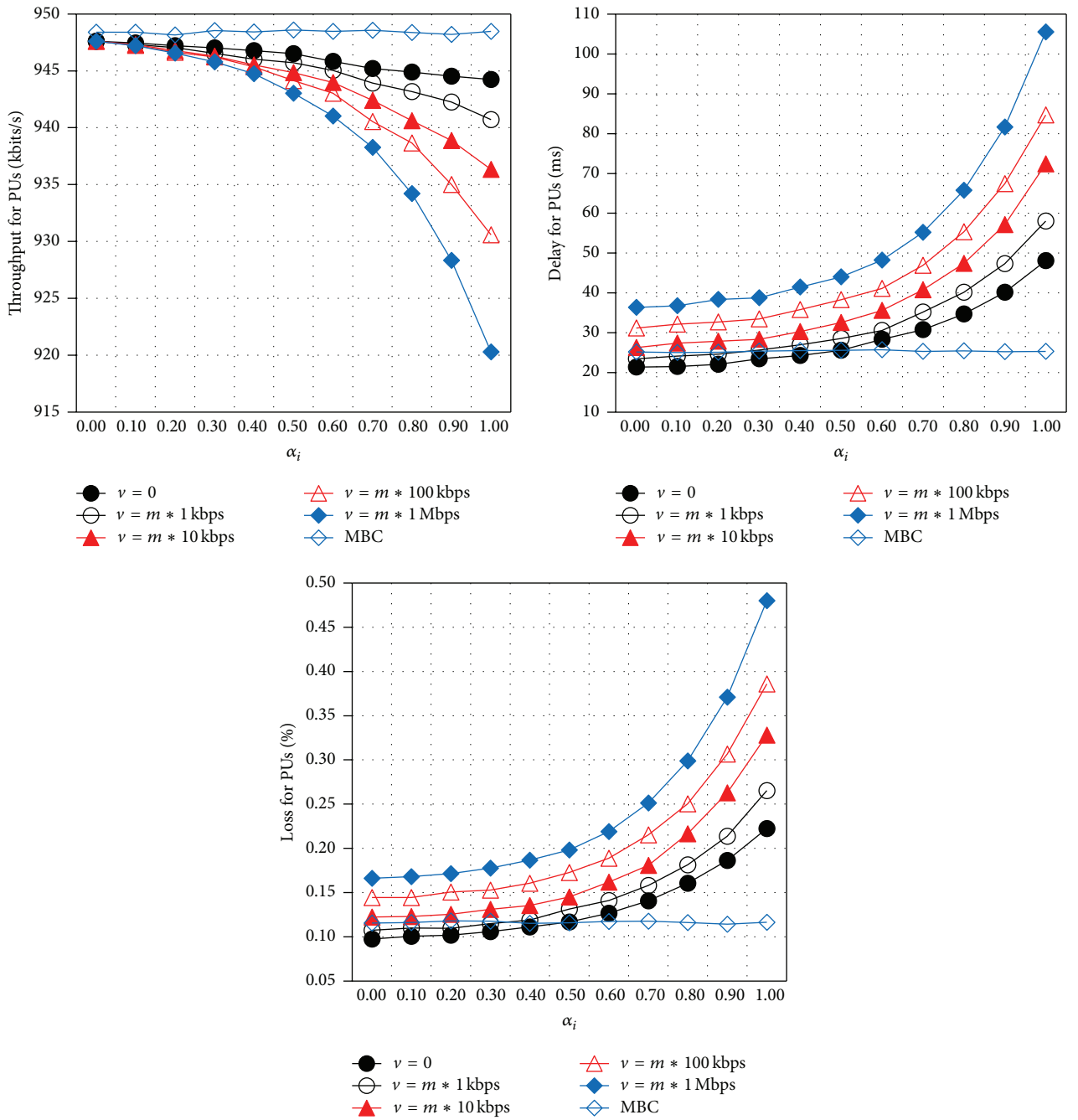

FIGURE 7: Mean throughput, packet delay, and loss for the PUs as a function of the parameter α_i with $n = 100$ PUs, $m = 50$ SUs, and average SNR = 0 dB.

Figure 2 shows the fairness of different methods calculated according to Jain's equation [38] as

$$F_{\mathrm{PU}} = \frac{\left(\sum_{i\in\mathbf{I}} \bar{r}_{Pi}\right)^2}{n \cdot \sum_{i\in\mathbf{I}} \bar{r}_{Pi}^2},$$

$$F_{\mathrm{SU}} = \frac{\left(\sum_{j\in\mathbf{J}} \bar{r}_{Sj}\right)^2}{m \cdot \sum_{j\in\mathbf{J}} \bar{r}_{Sj}^2},$$

$$F = \frac{\left(\sum_{i\in\mathbf{I}} \bar{r}_{Pi} + \sum_{j\in\mathbf{J}} \bar{r}_{Sj}\right)^2}{(n+m) \cdot \left(\sum_{i\in\mathbf{I}} \bar{r}_{Pi}^2 + \sum_{j\in\mathbf{J}} \bar{r}_{Sj}^2\right)}, \tag{45}$$

where F_{PU}, F_{SU}, and F stand for Jain's index of fairness for the PUs, SUs, and all of the users, respectively; $\bar{r}_{Pi}, \bar{r}_{Sj}$ are the mean throughputs for PU_i and SU_j, respectively. Results have been gathered for $n = 100$ PUs, $m = 100$ SUs, average

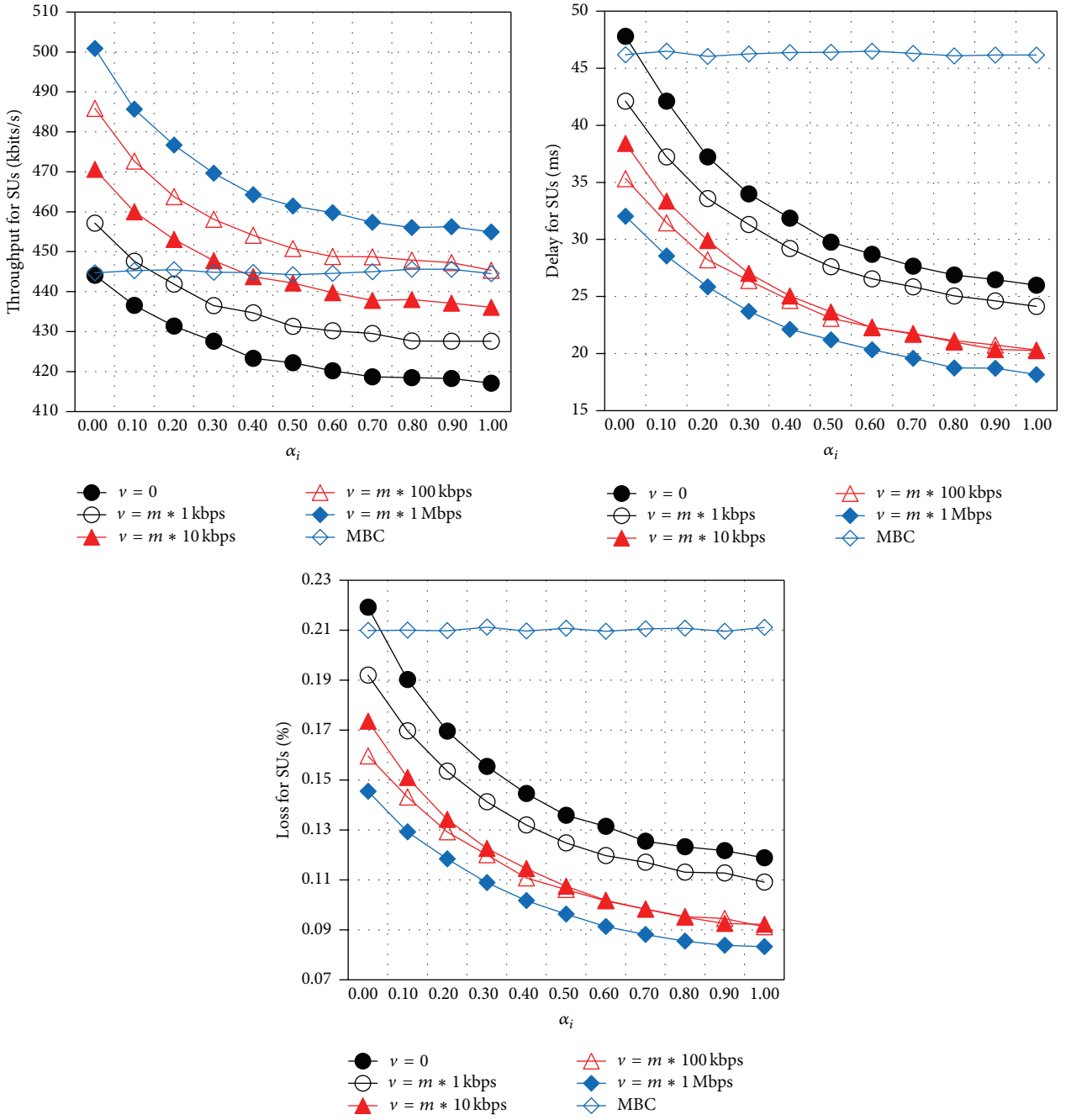

FIGURE 8: Mean throughput, packet delay, and loss for the SUs as a function of the parameter α_i with $n = 100$ PUs, $m = 50$ SUs, and average SNR = 0 dB.

SNR = 0 dB, and target buffer size for the PUs calculated according to (38a) (Scenarios 1 and 2) with different settings of α_i and v in Scenario 2.

It follows from these graphs that the fairness index among all of the users (denoted as F) is higher in Scenario 2 with $\alpha_i = 1$ and lower in Scenario 1 and Scenario 2 with $\alpha_i = 0$. Such results are rather expectable; in Scenario 1 and Scenario 2 with $\alpha_i = 0$, the SUs are served for free on a best-effort basis, whereas the PUs enjoy full QoS guarantees. Hence, the network resources are shared nonequally between the PUs and the SUs, with most of the RBs allocated to the PUs and the rest of the bandwidth distributed among the SUs. On the other hand, the fairness index for the PUs and SUs (i.e., F_{PU} and F_{SU}) is rather high in both scenarios with different settings of α_i and v, which indicates that the users of the same type are treated equally during resource allocation.

4.3. Simulation Results in Scenario 1. Let us evaluate the performance of a resource allocation algorithm in Scenario 1 in terms of mean throughput, packet end-to-end delay, and

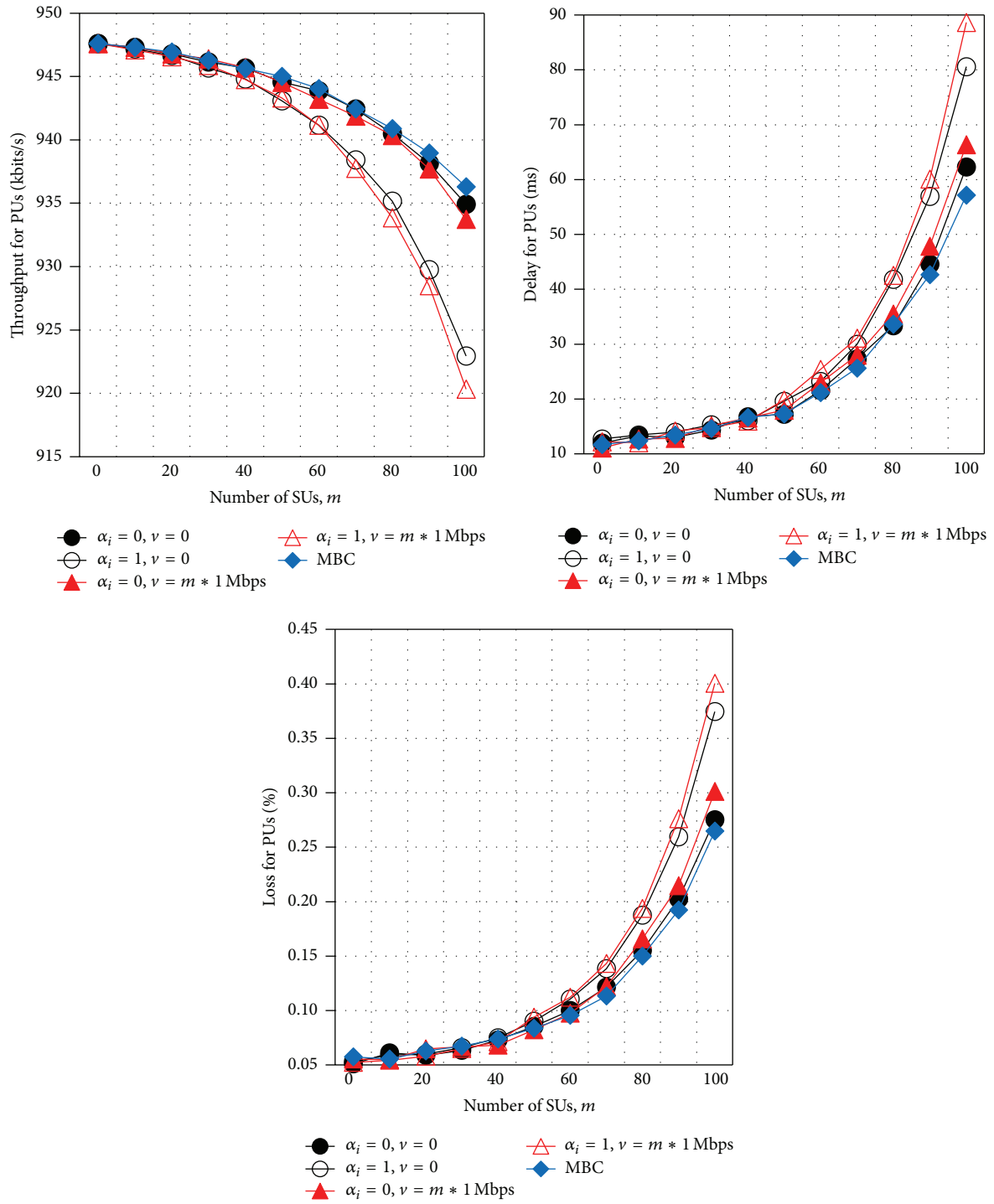

FIGURE 9: Mean throughput, packet delay, and loss for the PUs as a function of number of SUs with $n = 100$ PUs, $m = 50$ SUs, and average SNR = 0 dB.

loss for different user types. Figures 3–6 show results obtained in simulations with $n = 100$ PUs and average SNR = 0 dB. Results demonstrate that a UBA scheme has the largest packet delay and loss and the lowest throughput for both the PUs and the SUs. A poor performance of this scheme is explained as follows:

(i) QoS protection of PUs is not considered in UBA. Hence, the throughput, delay, and loss are the same for PUs and SUs.

(ii) Buffer sizes of individual users are not considered in resource allocation, and therefore the QoS for

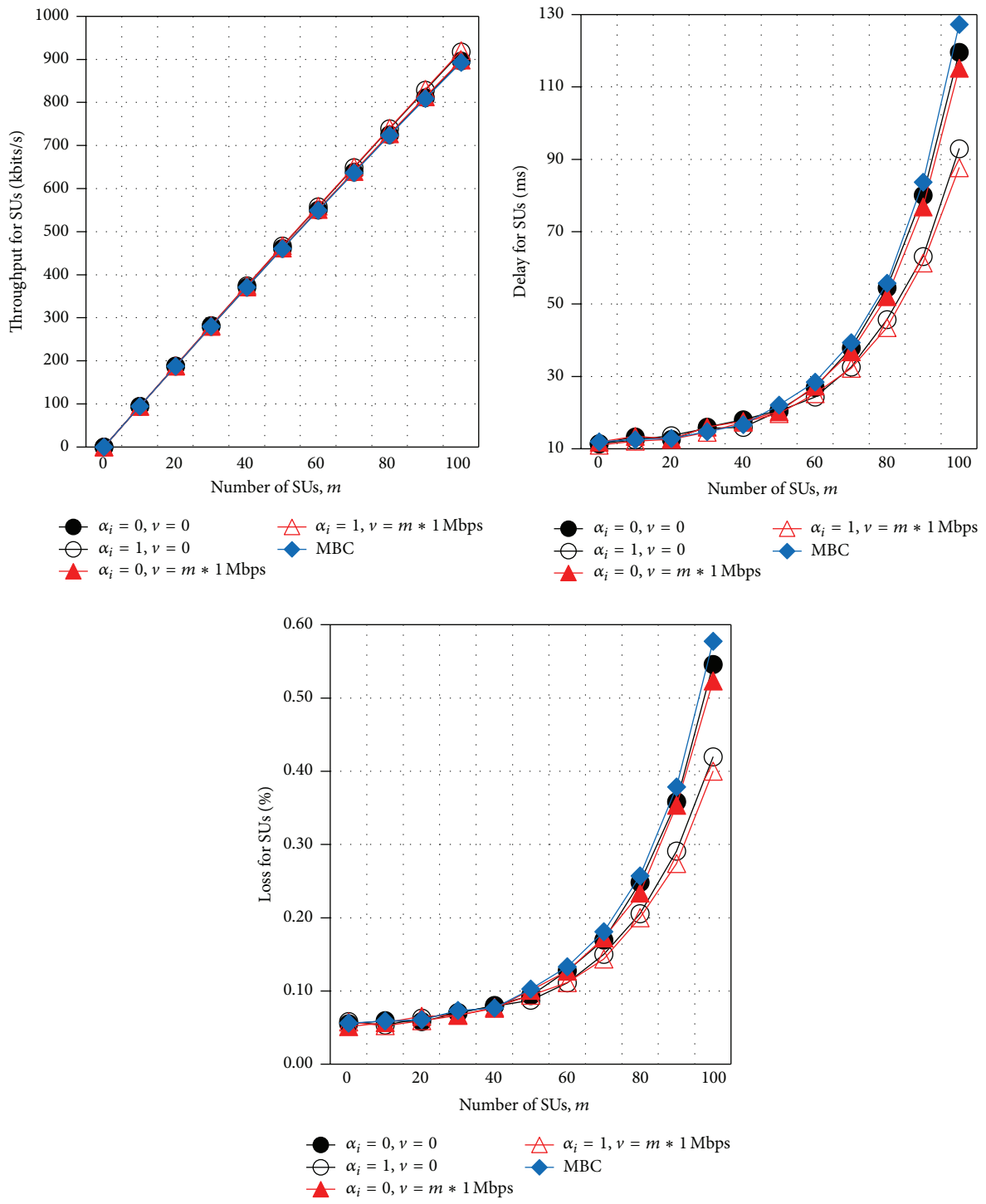

FIGURE 10: Mean throughput, packet delay, and loss for the SUs as a function of number of SUs with $n = 100$ PUs, $m = 50$ SUs, and average SNR = 0 dB.

different users varies significantly (since different users have different traffic demands).

(iii) Although widely used, the choice of transmission rate as an individual user utility is not very rational: some users may have very low demand and small buffer size and therefore there is no need to maximize their transmission rate and spend the network resources.

Performance of the DPC scheme is a little better than that of UBA. Note that DPC takes into account transmission delay, which is balanced against the transmission power in wireless

channels, and therefore performance of DPC for PUs and SUs is better than performance of UBA. However, the queuing delay related to users' buffer size and the QoS requirements of PUs are not considered, and therefore the delay and loss for PUs are much greater in UBA than those in GBC, NMC, MBC, and ABC.

GBC, NBC, MBC, and ABC show very consistent performance for the users. All of these techniques take into account the users' buffer size, which makes it possible to minimize the average delay and loss in the network for both the PUs and the SUs. Additionally, the QoS for the PUs is guaranteed by restricting the buffer size of each PU to stay below the target buffer size level, which ensures that the packet delays for PUs do not exceed the average target delays.

The best performances for the PUs and the SUs are achieved by GBC and NBC, respectively. Such results provide good demonstration of trade-off between delay and loss for the SUs and the target buffer size of the PUs. In particular, a "greedy" strategy shows good throughput, delay, and loss results for the PUs but poor performance for the SUs. On the other hand, it is possible to achieve some compromise between holding the QoS guarantees of the PUs and providing a reasonable performance for the SUs by using the less strict target buffer size settings (e.g., NBC, ABC, and MBC).

4.4. Simulations Results in Scenario 2. We now present simulation results in Scenario 2 with different settings of the parameters α_i and v and the target buffer size of the PUs calculated according to (38a) for $T = 10$ and benchmark these results with the performance of the MBC algorithm in Scenario 1. Figures 7 and 8 show mean throughput, packet end-to-end delay, and loss in simulations with $n = 100$ PUs, $m = 50$ SUs, and average SNR = 0 dB.

Results demonstrate that QoS of the PUs and the SUs depends to a great extent on the settings of α_i and v. The parameter α_i has negative impact on service performance for the PUs and positive impact on performance for the SUs. Note that the target buffer size of the PUs increases with α_i, which leads to decreased throughput, and increased delay, and loss for these users.

We also observe that a service provider's revenue v has negative impact on QoS of the SUs and positive impact on performance of the PUs. Such results are due to the more strict constraints on the network usage for different types of users. Consequently, the cases when the target QoS of PUs cannot be satisfied (and problem (15a)–(15g) is unfeasible) happen more frequently, and the QoS for these users degrades. On the other hand, the SUs utilize more bandwidth in response to the revenue constraints, and therefore their QoS improves.

4.5. Comparison between Scenario 1 and Scenario 2. Results below show performance of the algorithms in Scenario 1 and Scenario 2 with the target buffer size for the PUs calculated according to (38a) with $T = 10$ slots (in Scenarios 1 and 2) and different settings of the parameters α_i and v in Scenario 2. Figures 9 and 10 show mean throughput, packet end-to-end delay, and loss in simulations with $n = 100$ PUs, $m = 50$ SUs, and average SNR = 0 dB.

Results demonstrate that the QoS of the PUs is maximal in Scenario 1 and Scenario 2 with $\alpha_i = 0$, because in this case the PUs are provided with full QoS guarantees, whereas the SUs are served on a best-effort basis. On the other hand, the service performance for the SUs is better in Scenario 2 with $\alpha_i = 1$, since the target buffer size for the PUs increased, leading to degraded throughput and increased delay and loss for these users and improved performance for the SUs that pay for the shared resources.

5. Conclusion

The problem considered in this paper relates to the problem of transmission power and bandwidth allocation in a cognitive LTE network where the bandwidth is shared among the PUs and SUs. To guarantee that the QoS requirements of the PUs are satisfied, we enforce some upper bound on their buffer size. Unlike previously proposed resource allocation techniques, our algorithm is derived based on realistic assumptions, such as discrete spectrum resources and consideration of AMC deployed in LTE systems. Performance of the algorithm is evaluated using simulations in OPNET environment. Results show that the proposed resource allocation strategy performs significantly better than other relevant resource allocation techniques.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

References

[1] FCC Spectrum Policy Task Force, "Report of the spectrum efficiency working group," Tech. Rep., 2002.

[2] S. Haykin, "Cognitive radio: brain-empowered wireless communications," *IEEE Journal on Selected Areas in Communications*, vol. 23, no. 2, pp. 201–220, 2005.

[3] 3GPP, "Requirements for evolved UTRA (E-UTRA) and evolved UTRAN (E- UTRAN)," 3GPP TR 25.913, 2008.

[4] E. Dahlman, S. Parkvall, and J. Skod, *3G Evolution: HSPA and LTE for Mobile Broadband*, Academic Press, Oxford, UK, 2007.

[5] F. Baccelli, N. Bambos, and N. Gast, "Distributed delay-power control algorithms for bandwidth sharing in wireless networks," *IEEE/ACM Transactions on Networking*, vol. 19, no. 5, pp. 1458–1471, 2011.

[6] C. Y. Wong, R. S. Cheng, K. B. Letaief, and R. D. Murch, "Multiuser OFDM with adaptive subcarrier, bit, and power allocation," *IEEE Journal on Selected Areas in Communications*, vol. 17, no. 10, pp. 1747–1758, 1999.

[7] A. Larmo, M. Lindström, M. Meyer, G. Pelletier, J. Torsner, and H. Wiemann, "The LTE link-layer design," *IEEE Communications Magazine*, vol. 47, no. 4, pp. 52–59, 2009.

[8] V. Osa, C. Herranz, J. F. Monserrat, and X. Gelabert, "Implementing opportunistic spectrum access in LTE-advanced," *EURASIP Journal on Wireless Communications and Networking*, vol. 2012, article 99, 2012.

[9] A. Alizadeh, S. M.-S. Sadough, and S. A. Ghorashi, "Relay selection and resource allocation in LTE-advanced cognitive relay

networks," *International Journal on Communications Antenna and Propagation*, vol. 1, no. 4, pp. 303–310, 2011.

[10] Y. Chen, C. Cho, I. You, and H. Chao, "A cross-layer protocol of spectrum mobility and handover in cognitive LTE networks," *Simulation Modelling Practice and Theory*, vol. 19, no. 8, pp. 1723–1744, 2011.

[11] J. Acharya and R. D. Yates, "Dynamic spectrum allocation for uplink users with heterogeneous utilities," *IEEE Transactions on Wireless Communications*, vol. 8, no. 3, pp. 1405–1413, 2009.

[12] S. Y. Lien and K. C. Chen, "Statistical traffic control for cognitive radio empowered LTE-advanced with network MIMO," in *Proceedings of the IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS '11)*, pp. 80–84, Shanghai, China, April 2011.

[13] H.-P. Shiang and M. van der Schaar, "Queuing-based dynamic channel selection for heterogeneous multimedia applications over cognitive radio networks," *IEEE Transactions on Multimedia*, vol. 10, no. 5, pp. 896–909, 2008.

[14] W. Wang, K. G. Shin, and W. Wang, "Distributed resource allocation based on queue balancing in multihop cognitive radio networks," *IEEE/ACM Transactions on Networking*, vol. 20, no. 3, pp. 837–850, 2012.

[15] E-UTRA, "MAC protocol specification," 3GPP TS 36.321. (Release 8), 2007.

[16] 3GPP, "Physical channels and modulation," 3GPP TS 36.211, 2011.

[17] 3GPP, "Multiplexing and channel coding," 3GPP TS 36.212, 2008.

[18] H. Holma and A. Toskala, *LTE for UMTS: Evolution to LTE-Advanced*, John Wiley and Sons, 2011.

[19] D. V. Lindley, "The theory of queues with a single server," *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 48, no. 2, pp. 277–289, 1952.

[20] J. D. C. Little, "A proof for the queuing formula: $L = \lambda W$," *Operations Research*, vol. 9, no. 3, pp. 383–387, 1961.

[21] P. Mogensen, W. Na, I. Z. Kováes et al., "LTE capacity compared to the shannon bound," in *Proceedings of the IEEE 65th Vehicular Technology Conference (VTC-Spring '07)*, pp. 1234–1238, Dublin, Ireland, April 2007.

[22] P. Vieira, M. P. Queluz, and A. Rodrigues, "MIMO antenna array impact on channel capacity for a realistic macro-cellular urban environment," in *Proceedings of the IEEE 68th Vehicular Technology Conference (VTC-Fall '08)*, pp. 1–5, Calgary, Canada, September 2008.

[23] R. Kannan and C. Monma, "On the computational complexity of integer programming problems," in *Optimization and Operations Research*, vol. 157 of *Lecture Notes in Economics and Mathematical Systems*, pp. 161–172, Springer, 1978.

[24] R. Bracewell, "Heaviside's unit step function, H(x)," in *The Fourier Transform and Its Applications*, pp. 61–65, McGraw-Hill, New York, NY, USA, 3rd edition, 2000.

[25] M. Ó. Searcóid, *Metric Spaces*, Springer Undergraduate Mathematics Series, Springer, Berlin, Germany, 2006.

[26] S. Leyffer, "Integrating SQP and branch-and-bound for mixed integer nonlinear programming," *Computational Optimization and Applications*, vol. 18, no. 3, pp. 295–309, 2001.

[27] M. Fischetti and A. Lodi, "Local branching," *Mathematical Programming*, vol. 98, no. 1–3, pp. 23–47, 2003.

[28] A. Lodi, "The heuristic (dark) side of MIP solvers," *Hybrid Metaheuristics*, vol. 434, pp. 273–284, 2013.

[29] C. D'Ambrosio, A. Frangioni, L. Liberti, and A. Lodi, "A storm of feasibility pumps for nonconvex MINLP," *Mathematical Programming*, vol. 136, no. 2, pp. 375–402, 2012.

[30] T. Berthold, *Heuristic Algorithms in Global MINLP Solvers*, Verlag Dr. Hut, 2014.

[31] M. Fischetti and D. Salvagnin, "Feasibility pump 2.0," *Mathematical Programming Computation*, vol. 1, no. 2-3, pp. 201–222, 2009.

[32] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, 2004.

[33] V. Jacobson, "Congestion avoidance and control," *ACM SIGCOMM Computer Communication Review*, vol. 18, no. 4, pp. 314–329, 1988.

[34] J. Mo and J. Walrand, "Fair end-to-end window-based congestion control," *IEEE/ACM Transactions on Networking*, vol. 8, no. 5, pp. 556–567, 2000.

[35] OPNET website, http://www.opnet.com.

[36] E-UTRA and E-UTRAN, "Overall description," 3GPP TS 36.300. Stage 2 (Release 8), 2008.

[37] IPOGUE, Internet study 2007 and 2008/2009, research report, http://www.cs.ucsb.edu/~almeroth/classes/W10.290F/papers/ipoque-internet-study-08-09.pdf.

[38] R. Jain, W. Hawe, and D. Chiu, "A quantitative measure of fairness and discrimination for resource allocation in shared computer systems," Tech. Rep. DEC-TR-301, Eastern Research Laboratory, 1984.