Admission control and resource optimization for multiple-user OFDMA Cognitive Radio systems☆

Tariq Al-Khasib[1], Lutz Lampe*

Department of Electrical and Computer Engineering, University of British Columbia, Vancouver, BC, Canada

A R T I C L E   I N F O

A B S T R A C T

In Cognitive Radio (CR) systems, the fluctuating nature of the available frequency resource due to Primary Users (PUs) activity necessitates the introduction of admission and eviction measures at the CR system if a guaranteed Quality of Service (QoS) is required by Real Time (RT) Secondary Users (SUs). This problem has been recently addressed in the literature with simplified assumptions that might become unrealistic in practical system setups. In this paper, we tackle the problem of admission and eviction control of RT SUs in multiple-user Orthogonal Frequency Division Multiple Access (OFDMA) CR systems and propose new solutions that are practical and efficient at the same time. In particular, we propose three different ways to install a Resource Buffer Zone (RBZ) at the time of admission to limit future call drops resulting from fluctuating PU activity. We also study the effect of PU activity on the feasibility of the resource allocation problem and propose three different methods to resolve system outages once they occur. Numerical results obtained through Monte Carlo simulations demonstrate the efficacy of the proposed techniques.

© 2011 Elsevier GmbH. All rights reserved.

1. Introduction

As more services migrate towards the wireless domain to support user mobility, the need for more spectrum is higher than ever. Nevertheless, several measurement-based studies have indicated that the currently allocated spectrum is underutilized due to the static nature of the spectrum assignment. That is why dynamic spectrum allocation (a.k.a Cognitive Radio (CR)) has been recognized to be one of the best means to solve the rising spectrum scarcity problem [1,2]. In CR systems, a Secondary User (SU) is allowed to communicate over a specific frequency band as long as it guarantees no harmful interference to the spectrum owners (Primary Users (PUs)). This can be achieved by either fully prohibiting any transmission from SUs in the presence of any PU activity, or by limiting the transmission power at the SU transmitter such that the interference generated at the PU receiver does not exceed a predetermined level known as the *interference temperature* [3]. Either one of these techniques would require the SU transmitter to sense the band of interest to identify spectrum holes that represent a transmission opportunity for SUs [4,5].

Orthogonal Frequency Division Multiple Access (OFDMA) has been recognized to be one of the prime candidates to enable CR communications due to the flexible nature of carrier assignment. Carriers that are occupied by the PUs can be identified and excluded when transmitting to SU receivers. Thus, resource allocation techniques similar to the ones used in multiple-user OFDMA [6–8] can be used to allocate the available carriers to different SUs in an exclusive manner (i.e., each carrier is assigned to one SU) in the CR context. The solutions in [7,8] utilizing a Lagrange dual decomposition approach are particularly apt, as the resource allocation problem can be divided into several, per-carrier, sub-problems that can be solved efficiently in parallel. Furthermore, it was shown in [7,9] how the duality gap between the primal and dual problems approaches zero as the number of carriers increases. If spectrum overlap with the PUs is allowed by constraining the amount of interference experienced by the PUs due to the CR transmission, the techniques from [6–8] can be extended as shown in [10,11].

Looking back at the works in [6–8,10,11], we can see that only Best Effort (BE) users were considered with no guarantees on the rate achieved by each user. However, in most Real Time (RT) applications, a constant bit rate is necessary to achieve a satisfactory user experience. For this reason, extensions to include per-user rate constraints were studied in [12,13]. We note that introducing the new rate constraints to the problem of rate maximization could deem the problem infeasible due to the limited amount of power the system is allowed to utilize. The authors in [12,13] assumed that the problem is always feasible, although recognizing that this might not always be true. We argue that the possibility of overloading the

system at some point necessitates the introduction of admission and eviction control mechanisms that deal with the two fundamental tasks of (i) accepting/rejecting requests by new users in a way that limits the probability of calls being dropped during the lifetime of the user's session and (ii) dealing with the problem of infeasibility when it occurs due to changes in system resources dominated by unexpected PU activity on some of the previously available carriers. The authors in [14] recognized the possible infeasibility problem and proposed to fairly allocate the available resources to all the users when infeasibility occurs. This way, however, the rate constraint of almost every user in the system gets violated. Furthermore, [14] falls short of providing an admission control mechanism that limits such a disliked possibility.

Admission and eviction control for CR systems was the subject of several recent studies [15,16]. In [15], a fractional guard channel reservation scheme was used to control the balance between the blocking rate of newly arriving users and the dropping rate of the already admitted ones. The authors used a continuous time Markov chain model to describe the dynamics of the system, and by solving the global balance equation they were able to obtain the stationary probabilities of blocking and dropping a user. In [16], the authors proposed the use of a semi-Markov decision process to solve a profit maximization problem. By assigning a cost/revenue to each action of interest (admission, blocking, eviction, and completion), they were able to come up with an admission/eviction policy that maximizes the revenue of the system. In both [15] and [16], rate requirements by users were translated directly into bandwidth requirements. That is, there is a deterministic link between the amount of bandwidth currently available and the ability to admit or the necessity to evict a user at any point in time. While such an assumption was necessary for [15,16] to apply tools from Markov decision process theory, it ignores the current link conditions in the CR system, which limits the practicality of the resulting schemes.

In this paper, we tackle the problem of admission and eviction control in a CR environment while taking into account the important effect of the instantaneous channel conditions of users on the admission, resource optimization, and eviction dynamics of the system. We propose the use of a Resource Buffer Zone (RBZ) to limit the number of RT call drops due to the fluctuating nature of the available frequency resource. Unlike the RBZ described in [15], the buffer zone we install is only used at the time of admission of a new RT SU and can be freely utilized by the system afterwards to achieve the highest possible system throughput. We propose three different ways to set up such a buffer zone, two of which use statically allocated RBZs consisting of reserved power levels or frequency carriers that get held back at the time of admission. The third technique predicts the size of the RBZ based on current system dynamics. We also tackle the problem of user removal that is necessary to treat any infeasibility situation the system could face due to the dynamic nature of the available set of resources resulting from independent PU activity. We present three algorithms for user removal and discuss their advantages and disadvantages. Results show how the proposed admission control techniques can be efficiently used to reduce the drop rate of admitted RT SUs. The results also show how we can resolve any system infeasibility using user removal algorithms that have low complexity and high efficiency at the same time.

Organization: The rest of this paper is organized as follows. In Section 2 we present the system model under consideration. In Section 3, we deal with the problem of admission control and propose techniques to limit future user drops due to fluctuations in the available resources. In Section 4, the resource optimization problem is presented and the problem of infeasibility is identified and resolved using several proposed techniques. Finally, in Section 5, we present selected performance results and we conclude the paper
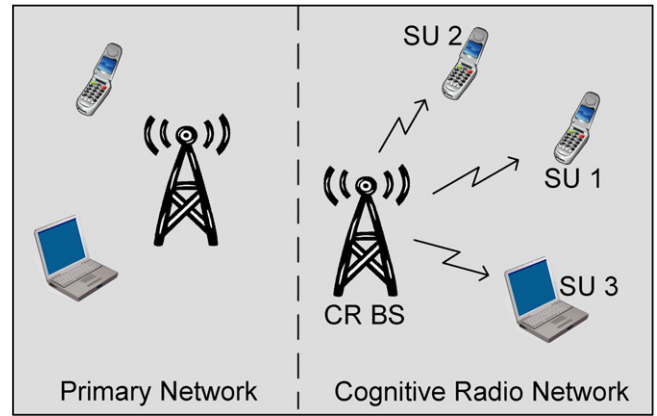

Fig. 1. Multiple-user downlink Cognitive Radio system.

in Section 6. Due to the large number of used abbreviations, a list of them is provided in Appendix A.

2. System model

We consider a multiple-user downlink CR system, as schematically illustrated in Fig. 1, that shares a frequency bandwidth with a primary system, which is the licensed owner of the spectrum. The CR system can utilize any portion of the bandwidth as long as it is not actively used by the primary system. The bandwidth of interest is B Hz and is divided at the CR Base Station (BS) into N OFDMA sub-bands.

2.1. Signaling model

OFDMA is utilized at the CR BS for signal transmission to multiple SUs in which an exclusive carrier allocation is assumed to prevent interference between different SUs. The received signal at the kth SU on the nth carrier can be written as

$$y_{k,n} = h_{k,n}x_{k,n} + v_{k,n}, \tag{1}$$

where $h_{k,n}$, $x_{k,n}$ and $v_{k,n}$ are the channel gain, transmitted signal and the zero mean unit variance i.i.d circularly symmetric complex Gaussian noise associated with the kth SU and nth carrier, respectively.

We define $p_{k,n}$ and $r_{k,n}$ as the transmit power and rate associated with the nth carrier and the kth SU, respectively. Tight bounds on $p_{k,n}$ as a function of the Bit Error Rate (BER) requirement, and carrier loading, $r_{k,n}$, are available in the literature for uncoded M-ary Quadrature Amplitude Modulation (QAM) signals transmitted over Additive White Gaussian Noise (AWGN) channels, as in (1). For example, in [17,18] $p_{k,n}$ is tightly approximated by

$$p_{k,n} \approx \beta_{k,n}\left(2^{r_{k,n}} - 1\right), \tag{2}$$

where $\beta_{k,n} > 0$ is a function of $\left|h_{k,n}\right|$ and the BER requirement. We observe that the approximation of $p_{k,n}$ is convex,[2] increasing in $r_{k,n}$, and that $p_{k,n} = 0$ when $r_{k,n} = 0$. The convexity of the approximation in (2) is of great importance as it facilitates the use of efficient convex optimization techniques that are necessary to solve the feasibility and resource optimization problems discussed in Sections 3 and 4, respectively.

---

[2] For real-valued rates, $r_{k,n}$, convexity of $p_{k,n}$ is well defined. When $r_{k,n}$ belongs to a discrete set of rates, convexity of the discontinuous staircase function means that the lines connecting consecutive corners of the staircase constitute a convex continuous function [8].

2.2. PU activity model

On the primary system's side, the same bandwidth of interest, B Hz, can support up to L PUs with a bandwidth of B/L Hz each. It is widely accepted that the activity on each one of the PU bands is independent and can be modeled using a two state Markov Chain (MC) [14,19,20]. The independence assumption is plausible since traffic and channels experienced by different PUs will generally be independent. Once a PU band is busy, it becomes free in the next time slot with probability $p_{b \to f}$. Otherwise, the PU band remains busy with probability $1 - p_{b \to f}$. On the other hand, once the PU band is in the free state, it becomes busy or stays free in the next time slot with probabilities $p_{f \to b}$ and $1 - p_{f \to b}$, respectively. It can be easily shown that this model is equivalent to a PU activity model that has independent exponentially distributed busy and free times similar to the one adopted in [16].

The CR BS continuously senses the spectrum of interest and only those carriers that carry no primary communications can be utilized for secondary transmission. At any time instance t, the set of free carriers is $\mathcal{N}_{\text{free}}(t)$. The cardinality of the set of free carriers, $N_{\text{free}}(t)$, evolves according to a fully connected MC in which each state represents the number of free PU bands that experience no active PU transmission in the time slot of interest. Thus, while in state $i \in 0, \dots, L$ at time t, the MC indicates the availability of $N_{\text{free}}(t) = iN/L$ free carriers that can be utilized for CR transmission. The state transition matrix $\boldsymbol{P}$ can be obtained using [14]

$$
\begin{aligned}
p_{i,j} &= \sum_{\ell=0}^{L} \binom{i}{\ell} p_{f \to b}^{\ell} (1 - p_{f \to b})^{i-\ell} \\
&\quad \times \binom{L-i}{\ell - i + j} p_{b \to f}^{\ell - i + j} (1 - p_{b \to f})^{L-j-\ell},
\end{aligned}
\tag{3}
$$

where $p_{i,j}$ is the element from the ith row and jth column of $\boldsymbol{P}$ that represents the transition probability from the ith state to the jth state.

2.3. SU arrival process

We consider two types of SUs based on their rate requirements. Non Real-Time (NRT) SUs have no rate requirements and are considered for BE service. The second set of users is the RT users that have a minimum rate requirement that is necessary to achieve a satisfactory user experience. We assume that the SU arrival to the CR system follows a Poisson process with rates λ_{RT} and λ_{NRT} for RT and NRT traffic, respectively. We also assume that the call duration of each RT and NRT user, k, is exponentially distributed with an average of $1/\mu_k$. These assumptions are common in the literature for data and voice traffic over wireless links and are not far from reality (e.g. [15,16]).

Upon their arrival, NRT SUs are always admitted and they receive as much resources as the system allows in a best effort manner. On the other hand, RT SUs are admitted only when there are enough system resources, including power and frequency carriers, to support their minimum rate requirement, assuming that they have higher priority than NRT SUs. Specifically, NRT SUs receive little or no resources when the system goes through a high load of RT traffic.

Without loss of generality, we assume that all system events, including SU arrival, SU departure, and PU activity changes are discrete time events that are synchronized based on a unified system clock. This is a practical assumption since transmission decisions are often made on a frame-by-frame basis.

3. Admission control

Upon the arrival of a new RT SU, an admission control mechanism is invoked to decide on whether to admit or reject the new request. The decision is made based on the availability of enough resources and the ability to maintain these resources over the course of the connection's lifetime. The two main goals of any admission control system are to i) minimize the number of users that get dropped before their actual session end time and to ii) minimize the number of users being blocked for lack of resources.

The set of carriers that are free from any PU activity, $\mathcal{N}_{\text{free}}(t)$, continuously changes according to the MC model discussed in Section 2.2. Thus, an already admitted RT SU can possibly lose service if the available set of free carriers can no longer support all admitted RT SUs without violating the maximum transmit power threshold ($P_{\max}$). The user drop rate can be reduced by installing an RBZ consisting of unallocated system resources to deal with any expected variations over the life-time of the user's session. Obviously, the larger the RBZ is the lower the drop rate. On the other hand, a too large RBZ would underutilize the system and prevent users from being admitted while resources are setting idle.

The concept of an RBZ has been used extensively in the cellular networks context [21,22]. In cellular networks, dropping an ongoing call is perceived to be more annoying than blocking a new call request. Thus, some system channels are reserved to serve calls handing over from neighboring cells. Similarly, dropping an already admitted RT user is more inconvenient than blocking a new admission request. Thus, we propose two types of admission control policies based on a static RBZ that is composed of portions of one of the available resources: frequency carriers or transmission power. In the third proposed admission control technique, we use a pragmatic approach towards estimating the size of the RBZ, in terms of number of frequency carriers, dynamically such that a pre-defined level of protection against RT SU call drops is achieved.

3.1. Fixed power threshold

In this scheme, a fixed portion of the available transmission power will be held back when a user is admitted. In other words, upon the arrival of a new RT SU with a minimum rate constraint, the following optimization problem is solved (for clarity, we drop the time index t)

$$
\min \sum_{k \in \mathcal{K}_{\text{RT}}^{+}} \sum_{n \in \mathcal{N}_{\text{free}}} p_{k,n}
\tag{4a}
$$

$$
\text{s.t.} \sum_{n \in \mathcal{N}_{\text{free}}} r_{k,n} \geq R_k, \quad \forall k \in \mathcal{K}_{\text{RT}}^{+},
\tag{4b}
$$

$$
\text{if } p_{\hat{k},n} \neq 0 \text{ then } p_{k,n} = 0, \ \forall k \neq \hat{k} \in \mathcal{K}_{\text{RT}}^{+}, \forall n \in \mathcal{N}_{\text{free}},
\tag{4c}
$$

where $\mathcal{K}_{\text{RT}}^{+}$ is the set of the already admitted RT SUs plus the newly arriving SU seeking admission and R_k is the minimum rate requirement of the kth SU. Assuming that the sum power constraint at the CR BS is $P_{\max}$, the new user is admitted to the system if and only if the total amount of power required to fulfill the rate requirements of all RT SUs is lower than a predetermined threshold. That is, if

$$
\sum_{k \in \mathcal{K}_{\text{RT}}^{+}} \sum_{n \in \mathcal{N}_{\text{free}}} p_{k,n} > \alpha P_{\max},
\tag{5}
$$

the new user is blocked for lack of resources. The factor $\alpha \in [0, 1]$ limits the maximum allowable power level at the BS at admission time. This way, the system is allowed to "breathe", i.e., even if the set of available resources change, up to a certain limit, after the user becomes admitted, the call is not dropped and the required rate is maintained. The lower α is, the higher the probability the call is

maintained for the whole call duration without interruption. On the other hand, a low α increases the probability of rejecting new requests from RT users.

3.2. Fixed carrier release

The other available resource dimension that can be employed to implement an RBZ in OFDMA systems is frequency. We then propose, in the second scheme, to forge an RBZ by voluntarily surrendering a fixed number, C, of the available carriers for admission purposes only. The carriers to surrender are picked randomly and the system is then checked for feasibility as in (4), but with a reduced set of available carriers $\hat{\mathcal{N}}_{\text{free}}$. If the problem turns feasible, i.e., $\sum_{k \in \mathcal{K}_{\text{RT}}^+} \sum_{n \in \hat{\mathcal{N}}_{\text{free}}} p_{k,n} \leq P_{\max}$, the new user is admitted and the full list of available carriers $\mathcal{N}_{\text{free}}$ is used for resource allocation.

3.3. Predictive carrier release

The previous two protection schemes are static in the sense that the size of the RBZ is fixed and needs to be chosen beforehand. This requires the availability of actual or simulated system measurements for the system administrator to be able to set the corresponding system parameters, α and C, accordingly. Alternatively, in this scheme, we try to predict the size of the RBZ, in term of number of carriers C, dynamically such that the system achieves a certain level of protection against call drops of RT SUs.

Towards that, we assume that a user drop happens mainly due to a decline in the number of carriers available for secondary transmission after the user was admitted. Although a user drop could also happen when the total number of available carriers remains the same or even increases if the set of available carriers is different from the one at admission time, we ignore this possibility since we assume that (i) the number of available carriers is usually large (64–2048), (ii) the channels are frequency selective across carriers, i.e., there is sufficient frequency diversity such that it is less likely that all available frequency carriers go into a deep fade simultaneously, (iii) the diversity created by the presence of multiple users reduces the effect of carrier shuffling due to PU activity, (iv) and most importantly, unlike [15,16], the resource allocation process, presented in Section 4 below, is dynamic and will produce a new carrier allocation map that achieves the best possible performance.

In general, a user that is admissible at time T_0 is only admitted if it would still be admissible at all time slots, $T_0 < t \leq T_0 + T_{\max}$, in the foreseen future with some probability p_{protect}. Thus, based on the argument presented above, we need to determine by how many carriers the number of free carriers at time T_0, $N_{\text{free}}(T_0)$, is expected to drop within a time interval of size $T_{\max}$ given an accuracy of prediction that is greater or equal to p_{protect}. The RT SU dropping rate is directly related to the prediction accuracy, p_{protect}, only when the load of the system is high enough to keep the system full at most of the times. It is at that system environment when the existence of an RBZ reflects directly on preventing a user drop. In particular, the need for an admission control is manifested and only takes effect when the network experiences high system loads. On the other hand, at lower system loads, the installation of an RBZ does not affect the operation of the system nor does it affects the efficiency of the resource allocation. Thus, for the rest of this section, we will refer to the probability $1 - p_{\text{protect}}$ as the dropping risk probability (p_{risk}).

Our goal is to find, up to a dropping risk probability (p_{risk}), what is the maximum number of PU bands ($C_{\max}$) that are expected to be lost to PUs in future time slots until the first SU departure happens. Limiting the time horizon of the prediction algorithm to the time the first SU departure is expected to happen is intuitively valid since the departure of an SU releases some of the system resources that

add directly to the protection of the already admitted users. But, we note that other choices for the time horizon are possible. Without loss of generality, we assume that $T_0 = 0$ to simplify the notation in the following.

From the properties of the exponential distribution, the time until the first departure amongst K_{RT} RT SUs each with exponentially distributed call duration with mean $1/\mu_k$, is also exponentially distributed with mean $1/\mu = 1/\sum_{k=1}^{K_{\text{RT}}} \mu_k$. Thus, the probability that the first SU departs after time T is

$$p_1(T) = Pr(\text{First SU departs after } T \text{ time slots}) = \exp(-\mu T). \quad (6)$$

On the other hand, the probability that the total number of free PU bands drops by C bands by time T is,

$$
\begin{aligned}
p_2(C, T) &= Pr(X_t \leq X_0 - C \text{ for any } t \leq T | X_0) \\
&= \sum_{t=1}^{T} Pr(X_t \leq X_0 - C, X_{t-1,\dots,1} > X_0 - C | X_0) \\
&= \sum_{t=1}^{T} \sum_{j=0}^{X_0 - C} Pr(X_t = X_0 - C - j, X_{t-1,\dots,1} > X_0 - C | X_0),
\end{aligned}
\quad (7)
$$

where X_t is the number of free PU bands at time t such that $N_{\text{free}}(t) = X_t N/L$. Assuming that $\overline{\boldsymbol{P}}_{(X_0 - C)}$ is equal to the transition matrix $\boldsymbol{P}$ but with all the transition probabilities leading to states less than or equal to $X_0 - C$ set to zero, then $Pr(X_t = X_0 - C - j, X_{t-1,\dots,1} > X_0 - C | X_0)$ in (7) is equal to $a_{X_0, X_0 - C - j}(t, X_0 - C)$, where

$$\boldsymbol{A}(t, X_0 - C) = \overline{\boldsymbol{P}}_{(X_0 - C)}^{t-1} \boldsymbol{P}, \quad (8)$$

and $a_{i,j}(t, X_0 - C)$ represents the element of matrix $\boldsymbol{A}(t, X_0 - C)$ on the ith row and jth column.

The joint probability that both events happen, i.e., the probability that the total number of free PU bands drops by C bands by time T and the first SU departs after time T, is equal to the product of the two probabilities, i.e., $p_1(T)p_2(C, T)$, since both events are independent.

Define C_T as the minimum number of PU bands that makes $p_1(T)p_2(C, T) \leq p_{\text{risk}}$. Then we need to determine

$$C_{\max} = \max_{T \in \{T_0, \dots, T_{\max}\}} C_T. \quad (9)$$

The time horizon, $T_{\max}$, is the time at which $p_1(T_{\max}) \leq p_{\text{risk}}$. The pseudocode in Table 1-Algorithm I provides an efficient implementation of the procedure discussed above.

As an example, consider a CR system that has $N = 128$ carriers, out of which $N_{\text{free}}(T_0) = 100$ are free from any PU activity at time T_0. At the primary system side, the same band is divided amongst 128 PUs, i.e., $L = N$, that switch independently between the Free and Busy states according $p_{b \to f} = 0.2$ and $p_{f \to b} = 0.05$. Assume that the CR system has 10 already admitted RT SUs with independent call durations that are exponentially distributed with identical mean $1/\mu_{\text{RT}} = 30$ time slots each. For a dropping risk probability of $p_{\text{risk}} = 0.01$, the product $p_1 p_2$ from Algorithm I in Table 1 evolves with time T as shown in Table 2. It is clear from Table 2 that the maximum drop in the number of free carriers the system could face is $C_{\max} = 9$ carriers. Thus, the size of the RBZ that needs to be installed to achieve a dropping risk of at most 1% is 9 carriers. It should be noted that most figures in Table 2 are shown for illustrative purposes only. Algorithm I in Table 1 only captures the shaded numbers in Table 2, which are sufficient to obtain $C_{\max}$ in (9). The underlined figure on each row of Table 2 corresponds to C_T as defined above.

4. Resource optimization

After having investigated the problem of admission control in Section 3, we now turn our attention to the problems of resource

Table 1
Pseudocode of Algorithms I and II.

Algorithm I: find expected carrier drop	
1:	$T = 1$
2:	$C_{\max} = 0$
3:	$p_1 = \exp(-\mu T)$
4:	**if** $p_1 \leq p_{\text{risk}}$ **then**
5:	exit
6:	**end if**
7:	$p_2 = \Pr(X_t \leq X_0 - C_{\max} \text{ for any } t \leq T | X_0)$
8:	**if** $p_1 p_2 < p_{\text{risk}}$ **then**
9:	$T = T + 1$
10:	go to 3
11:	**end if**
12:	$C_{\max} = C_{\max} + 1$
13:	go to 7
Algorithm II: Optimal User Removal (OptRem)	
1:	// Initialize number of RT SUs to drop
2:	$d = 1$
3:	$\mathcal{S}$: set of all $\binom{K}{K-d}$ user combinations
4:	$\mathcal{F}$: set of feasible combinations in $\mathcal{S}$
5:	**if** $\mathcal{F}$ is empty **then**
6:	$d = d + 1$
7:	go to 3
8:	**end if**
9:	find combination in $\mathcal{F}$ with highest sum rate

optimization and eviction control of already admitted SUs. At any system event such as a new SU being admitted, an already admitted SU leaving the system, or a change to the system's resources, i.e., available carriers, due to PU activity, the following optimization problem is solved to efficiently and optimally allocate the available system resources

$$\max \sum_{k \in \mathcal{K}} \sum_{n \in \mathcal{N}_{\text{free}}} r_{k,n}, \tag{10a}$$

$$\text{s.t.} \sum_{n \in \mathcal{N}_{\text{free}}} r_{k,n} \geq R_k, \quad \forall k \in \mathcal{K}_{\text{RT}}, \tag{10b}$$

$$\sum_{k \in \mathcal{K}} \sum_{n \in \mathcal{N}_{\text{free}}} p_{k,n} \leq P_{\max}, \tag{10c}$$

$$\text{if } p_{\hat{k},n} \neq 0 \text{ then } p_{k,n} = 0, \quad \forall k \neq \hat{k} \in \mathcal{K}, \forall n \in \mathcal{N}_{\text{free}}, \tag{10d}$$

where $\mathcal{K}$ is the set of all, RT and NRT, admitted users.

This problem and the power minimization problem in (4) can be efficiently solved using a Lagrange dual decomposition similar to the one in [7,8,12,13]. In particular, each one of the rate constraints in (4b) and (10b) as well as the sum power constraint in (10c) gets associated with a Lagrange multiplier and the Lagrange dual problem is solved. It should be noted that the problem in (10) might

become infeasible if the set of available carriers becomes insufficient to provide the required data rates by the admitted users. The infeasibility problem could still be faced even when an admission control mechanism as proposed in Section 3 is used. Changes to the available system resources beyond those predicted can still occur forcing the system into an infeasible state. Thus, before attempting to solve the problem in (10), a feasibility check has to be performed to ensure that the system is still in a feasible state. The feasibility check involves solving a power minimization problem similar to that in (4). The system is declared feasible when the total amount of power required to fulfill the rate requirements of all RT SUs is lower than the maximum power constraint (i.e., $\sum_{k \in \mathcal{K}_{\text{RT}}^+} \sum_{n \in \mathcal{N}_{\text{free}}} p_{k,n} > P_{\max}$).

When the system can no longer support the promised rates, a decision needs to be made to bring the system into a feasible state again. One possible solution could be to drop the rate requirements for all users and then allocate the available resources to the already admitted users in a fair manner. This approach was investigated in [14]. Another approach could be to drop as few users as possible to bring the system back to a feasible state. We believe that the latter user removal approach is more appropriate, since dropping the rate requirement of all users could lead to serious QoS degradation for most of the users by violating the promised rate guarantees. A user removal approach would limit the unsatisfactory effects to a limited set of users while satisfying the needs of most of the already admitted users.

In the following sub-sections, we propose three different algorithms to resolve system infeasibilities through user removal. The first two approaches represent the two extreme alternatives in terms of performance complexity tradeoff. In the third algorithm, we devise an innovative approach that delivers performance advantages over the long run and still maintains a low computational complexity profile.

4.1. Optimal User Removal (OptRem)

We consider the Optimal Removal (OptRem) algorithm to be the one that (i) minimizes the number of users that need to be dropped to make the problem in (10) feasible, and (ii) if there exist multiple user dropping choices with equivalent number of users, choose the one that maximizes the overall achievable rates. The pseudocode for the OptRem algorithm is shown in Table 1-Algorithm II. It tests all user combinations that involve a single user removal for feasibility. If a single user removal yields no feasible combinations, i.e., set $\mathcal{F}$ (Line 4) is empty, all user combinations with two users removed are tested and so forth until a non empty set of feasible combinations is obtained (Lines 4–8). Once a non-empty set $\mathcal{F}$ is available, the algorithm picks the user combination that maximizes the system sum rate (Line 9).

Table 2
$p_1 p_2$ from Algorithm I in Table 1.

	$C = 0$	$C = 1$	$C = 2$	$C = 3$	$C = 4$	$C = 5$	$C = 6$	$C = 7$	$C = 8$	$C = 9$
$T = 1$	0.3461	0.2539	0.1729	0.1089	0.0633	0.0340	0.0168	0.0077		
$T = 2$	0.3113	0.2463	0.1834	0.1282	0.0841	0.0518	0.0301	0.0166	0.0086	
$T = 3$	0.2485	0.2041	0.1587	0.1165	0.0809	0.0531	0.0331	0.0196	0.0111	0.0060
$T = 4$	0.1909	0.1606	0.1283	0.0972	0.0698	0.0475	0.0308	0.0190	0.0113	0.0064
$T = 5$	0.1439	0.1232	0.1005	0.0778	0.0572	0.0399	0.0266	0.0169	0.0103	0.0060
$T = 6$	0.1073	0.0932	0.0772	0.0608	0.0455	0.0324	0.0219	0.0142	0.0088	0.0052
$T = 7$	0.0794	0.0698	0.0586	0.0468	0.0355	0.0257	0.0177	0.0116	0.0073	0.0044
$T = 8$	0.0584	0.0519	0.0441	0.0357	0.0274	0.0200	0.0140	0.0093	0.0059	0.0036
$T = 9$	0.0428	0.0384	0.0330	0.0270	0.0210	0.0155	0.0109	0.0073	0.0047	0.0029
$T = 10$	0.0313	0.0283	0.0245	0.0203	0.0159	0.0119	0.0084	0.0057	0.0037	0.0023
$T = 11$	0.0228	0.0208	0.0182	0.0151	0.0120	0.0090	0.0065	0.0044	0.0029	0.0018
$T = 12$	0.0166	0.0152	0.0134	0.0113	0.0090	0.0068	0.0049	0.0034	0.0022	0.0014
$T = 13$	0.0120	0.0111	0.0099	0.0084	0.0067	0.0051	0.0037	0.0026	0.0017	0.0011
$T = 14$	0.0087	0.0081	0.0072	0.0062	0.0050	0.0039	0.0028	0.0020	0.0013	0.0008

The OptRem algorithm is very complex since it explores up to $\sum_{d=1}^{D} \binom{K}{K-d}$ user combinations, D being the number of users actually dropped, before coming up with the list of users to be dropped. This complexity grows as the number of already admitted users increases and as the number of users to be removed at a time, D, increases. This algorithm is only used as a benchmark to compare to other algorithms and is not recommended for real systems given the high complexity associated with the exhaustive search.

4.2. Random User Removal (RandRem)

The OptRem algorithm, discussed above, can be classified as the most extreme solution in terms of computational complexity. Another, pragmatic and lower-complexity heuristic is the Random Removal (RandRem) algorithm. Once the system hits a state of infeasibility, this algorithm randomly picks one user for call termination and the system gets checked for feasibility again. If the system remains infeasible, another user is removed until a feasible set of users is achieved.

Compared to the OptRem algorithm above, RandRem can result in the removal of more users before bringing the system to a feasible state. The extra loss would only happen if a single user drop was not sufficient to alleviate the problem of infeasibility. This can be mostly observed when the PU dynamics are very fast or the range of user rate requirements is very large.

4.3. User removal based on Lagrange multipliers (LagRem)

The Lagrange multiplier associated with each rate constraint in the admission control problem in (4) or the resource allocation problem in (10) represents a measure of how difficult it was to meet the corresponding constraint. That is, the higher the Lagrange multiplier the bigger the effect of that constraint on the system's objective function [23]. Thus, dropping it would give the biggest possible room for other users to achieve their needs.

That being said, when the problem in (10) becomes infeasible, we propose to sequentially drop the user with the highest Lagrange multiplier until the system becomes feasible. We will refer to this algorithm as the LagRem algorithm. As we will see in Section 5, this method leads to desirable long term effects in contrast to the OptRem algorithm. To explain, the "optimality" of the OptRem algorithm introduced above is only guaranteed on a short term basis since the user dropping decision is based on the instantaneous conditions of the system. Such a decision might turn out to be sub-optimal over the long term dynamics of the system. Since the OptRem algorithm frees the least amount of resources necessary to bring the system back to a feasible state it creates a tight fit, in terms of number of users and achievable rates, that might cause more drops in the near future as the system progresses. Different from this, the LagRem algorithm is best suited for long term activity by design since it drops the user that gives the biggest room possible for other users to maintain their sessions. Furthermore, it enjoys a low computational complexity that is equal to that of the RandRem algorithm discussed above. This is true since the Lagrange multipliers come as a bi-product of solving the feasibility check discussed earlier using a Lagrange dual decomposition similar to that proposed in [7,8,12,13].

5. Simulation results

In this section, we present simulation results to evaluate the proposed admission control mechanisms from Section 3 and the user removal algorithms from Section 4. We adopt the approximation from (2) with $\beta_{k,n}$ according to [18] for the per-tone transmission
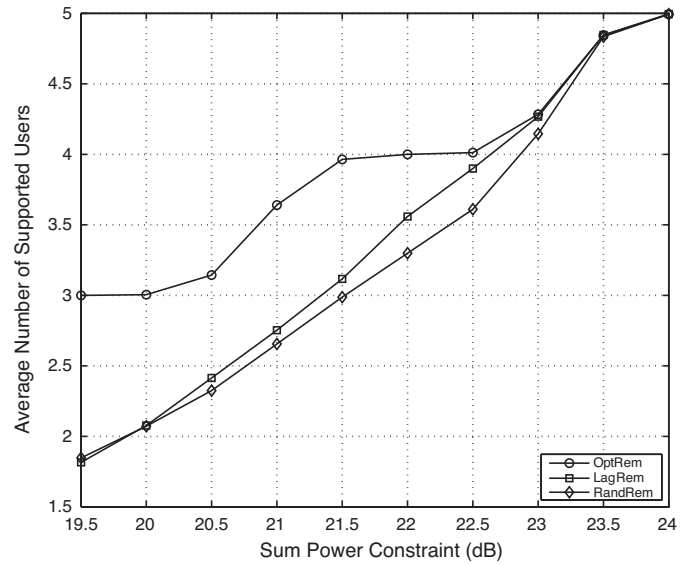

Fig. 2. Average of the maximum number of SUs, among 5 SUs, the system can maintain while achieving a feasible resource allocation given a sum power constraint: comparing three different user removal algorithms.

power, $p_{k,n}$, as a function of the required BER and the carrier loading, $r_{k,n}$. We assume that the rates $r_{k,n}$ can take real values[3] and the minimum BER requirement is 10^{-4}.

5.1. Performance of user removal algorithms

5.1.1. Short term performance of user removal algorithms

We start by evaluating the short term performance of the three user removal algorithms from Section 4. Towards that end, we assume that 5 SUs with minimum rate requirements of {5, 10, 15, 20, 25} bits/symbol were initially admitted based on an earlier channel realization that is of no significance. We then generated 10,000 i.i.d channel realizations and tested for system feasibility with the presence of all 5 users given a sum power constraint. If the problem turns infeasible given a specific channel realization, we invoke the three removal algorithms to achieve a feasible set of users. Fig. 2 shows the average number of supported users, and equivalently the system goodput, using the three removal algorithms for different levels of the sum power constraint. In this figure, we assume the presence of no primary system and the band is divided into $N = 64$ frequency carriers. The relatively low number of users and frequency carriers is inevitable due to the very high complexity of the OptRem algorithm.

The figure clearly shows the superiority of the OptRem algorithm and how it supports the highest number of users compared to the other two algorithms. However, the complexity of the OptRem algorithm is prohibitive, especially when the number of admitted users is high or the number of users to be removed is high. The figure also shows how the LagRem algorithm has a performance that is close to that of the OptRem algorithm when the removal of one user is sufficient to bring the system into a feasible state. It also performs better than the RandRem algorithm especially in the region that involves up to two user drops. This region is of great significance since one or two user removals are usually sufficient, in most realistic models, to bring the system to a feasible state.

---

[3] Without loss of generality, a finite set of discrete carrier loadings, $r_{k,n}$, can be easily accommodated, see e.g. [8] for an optimization problem similar to (4).
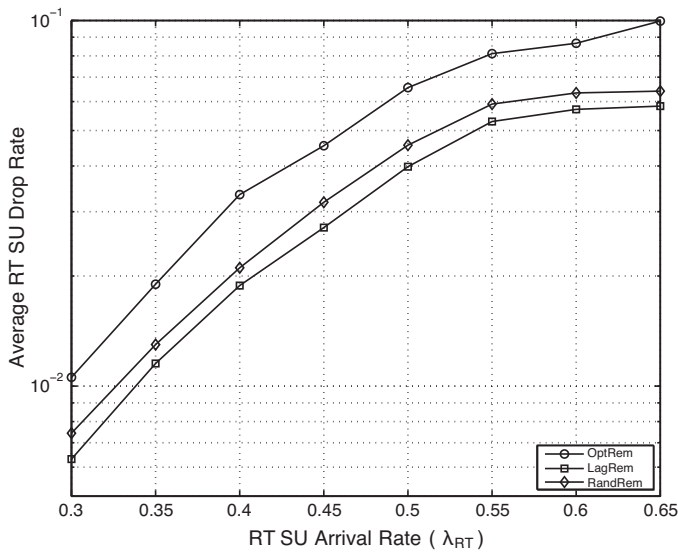
Fig. 3. Average RT SU drop rate versus a range of RT SU arrival rates (λ_{RT}): comparing three different user removal algorithms.

Fig. 4. Average RT SU drop rate versus a range of RT SU arrival rates (λ_{RT}) when an RBZ based on fixed power threshold is installed.

5.1.2. Long term performance of user removal algorithms

After confirming the superiority of the OptRem algorithm on short term basis, we now consider the performance of the three user removal algorithms in the long run. For this, and for the rest of this section, we consider the downlink of a CR system which shares the bandwidth with a primary system that has priority over the band of interest. This bandwidth is divided into $N = 128$ OFDMA narrowband carriers, and the channels between the CR BS and the different SUs are i.i.d circularly symmetric unit variance complex Gaussian across tones and SUs. We would like to emphasize that the assumed channel model is not critical to the performance of the proposed algorithms, and was chosen for simplicity. We further assume, for simplicity and ease of analysis, that all arriving RT SUs have the same rate requirement of $R = 20$ bits/symbol and the same average call duration of $1/\mu = 30$ time slots. The maximum power that the CR BS can utilize is $P_{max} = 30$ dB (note the power normalizations applied to channel gains and noise). The number of PUs occupying the same bandwidth is $L = 128$. Each one of the PU bands transitions between the busy and free states with probabilities $p_{f \to b} = 0.05$ and $p_{b \to f} = 0.2$.

In Fig. 3, we plot the average drop rate of admitted RT SUs versus their arrival rate to the system for the three different removal algorithms. Interestingly, we see that the OptRem algorithm achieves the worst performance, in terms of RT SU drop rate, compared to the other two algorithms. This is mainly due to the fact that the OptRem algorithm releases the least amount of system resources necessary to achieve a feasible solution to the resource allocation problem. Recall that the OptRem algorithm maintains the highest number of users possible and the highest sum rate possible amongst the admitted users set. This way, the algorithm tightly fits as many users and as much rate as possible making the system vulnerable to more service outages in the near future. In contrast to this, the LagRem algorithm is designed to remove users that have the biggest influence on the system performance and thus, creating the biggest room, resource wise, that accommodates future system dynamics better than other removal techniques. That is why the LagRem algorithm achieves the lowest drop rate in the long run. For the rest of the results presented in this section, we will assume that the LagRem algorithm is used to resolve any system infeasibility due to its low complexity and superior long term performance.
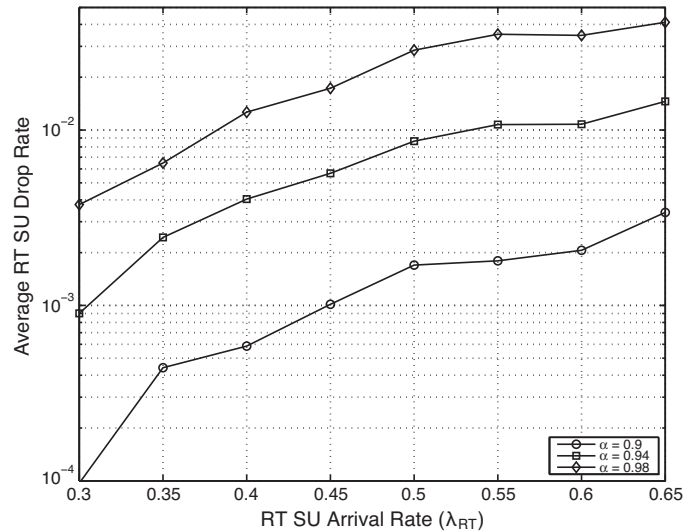
5.2. Performance of admission control algorithms

5.2.1. Admission control with static RBZs

After evaluating the performance of the different user removal algorithms, we move forward towards evaluating the performance of the three admission control mechanisms from Section 3. First, in Fig. 4, we present simulation results for the power threshold based RBZ technique from Section 3.1. Fig. 4 shows the average drop rate of RT SUs against their arrival rate, λ_{RT}, for different admission power thresholds, $\alpha = \{0.9, 0.94, 0.98\}$. The figure clearly shows how the drop rate of RT SUs decreases as more power gets held back when admitting new arrivals. The power gap between admission time and resource optimization time in current and future time slots protects admitted RT SUs against a level of resource fluctuation that is proportional to the size of the power gap.

The second alternative towards installing a fixed RBZ for user protection against call drops after admission is through the use of the fixed carrier release method from Section 3.2. Similar to Fig. 4, Fig. 5 plots the average RT SU drop rate for different arrival rates and different levels of carrier release C. It can be seen that the more carriers we voluntarily release while testing for admissibility, the more protection the admitted users get against call drops while in service due to PU activity.

In Figs. 4 and 5, the RBZ is static and is usually assigned before any system activity can begin. Thus, some measurements, actual or simulated, are necessary for these thresholds to be set if a specific level of RT SU drops is required. By looking at plots similar to the ones in Figs. 4 and 5, a system administrator seeking a drop rate that is below 1%, for example, would pick a fixed power threshold or a fixed number of carriers to be released according to the maximum anticipated system load.

To see the effect of the RBZ on the number of RT SUs that get blocked from service, Fig. 6 plots the average blocking rate of RT SUs versus the size of the RBZ, in terms of released carriers C, for different levels of RT SU arrival rates. As the size of the RBZ increases, more carriers are held back at the time of admission, which reduces the number of RT SUs the system can support. Thus, more RT SUs get denied admission for the sake of protecting already admitted RT SUs. Although the number of RT SUs that get admitted drops as the size of the RBZ increases, the number of dropped RT SUs decreases too. Thus, the number of RT SUs that successfully finish their sessions, referred to as "good" RT
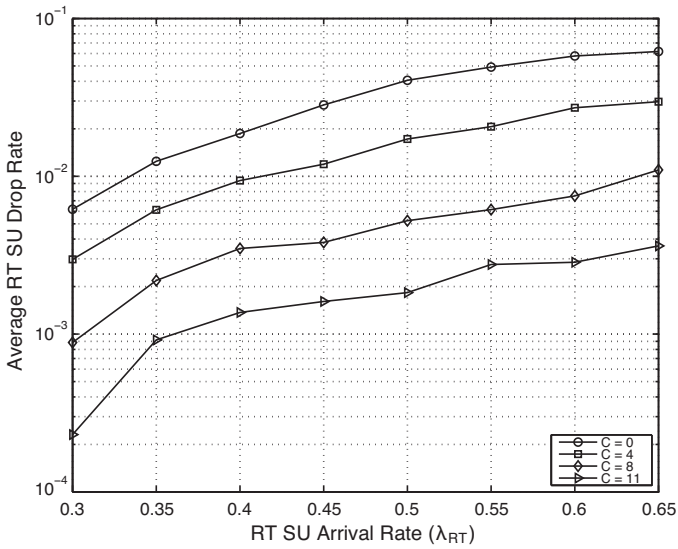
Fig. 5. Average RT SU drop rate versus a range of RT SU arrival rates (λ_{RT}) when an RBZ based on fixed carrier release is installed.

SUs, does not necessarily drop in the same manner. To illustrate this, Fig. 7 plots the average number of good RT SUs admitted to the system at any given time slot versus the size of the RBZ for different levels of RT SU arrival rates. At low RBZ sizes, the reduction in the drop rate is more significant than the reduction in the admission rate. Thus, the number of good users increases. As the RBZ size increases, the drop rate eventually approaches zero, and any increase in the RBZ size only prevents new RT SUs from being admitted causing a drop in the overall number of good RT SUs.

The system administrator could either choose the size of the RBZ, C, such that the overall system *goodput* is maximized regardless of the dropping and blocking rates associated with that C or, alternatively, aim at maintaining a minimum level of RT SU dropping rate and choose C that achieves the best overall system goodput given the dropping rate constraint. We believe that the latter approach is preferable since it provides a guarantee on the level of QoS as seen by the end user.
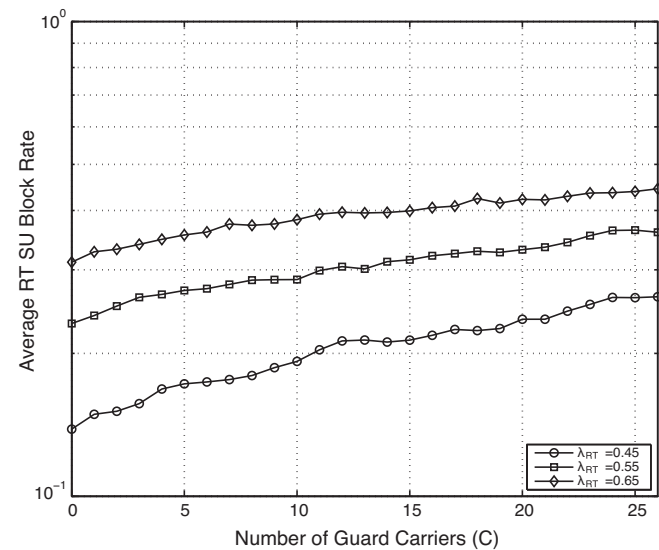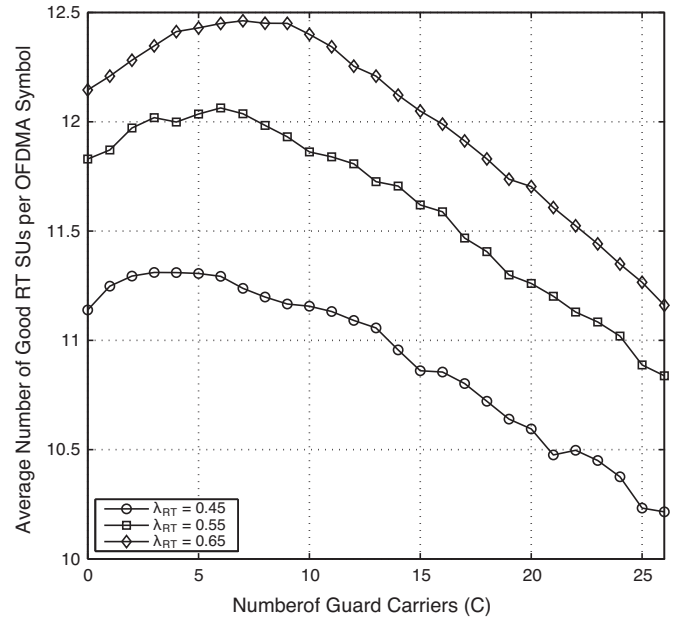

Fig. 7. Average number of "good" RT SUs admitted to the system at any given time slot versus size of the installed RBZ that is based on the fixed carrier release algorithm.

5.2.2. Admission control with predictive RBZ

Figs. 4 through 7 reveal the inherent tradeoff between the dropping rate and blocking rate and its implications on the overall system performance. To capture all of these effects, the availability of actual or simulated measurements is crucial to set the size of the RBZ accordingly. If these measurements were not available beforehand, protection based on predictive carrier release RBZ becomes the preferred option. In the predictive carrier release approach, a target maximum level of dropping rate, p_{risk}, is specified and the algorithm sets the size of the RBZ accordingly without compromising the overall system performance.

In Fig. 8, we plot the average RT SU drop rate against the RT SU arrival rate when the requested dropping risk probability was $p_{risk} = \{0.7\%, 1\%, 7\%\}$. Recall that the dropping risk probability, p_{risk}, represents the maximum level of call drops the system should encounter when the system load is high. The figure clearly
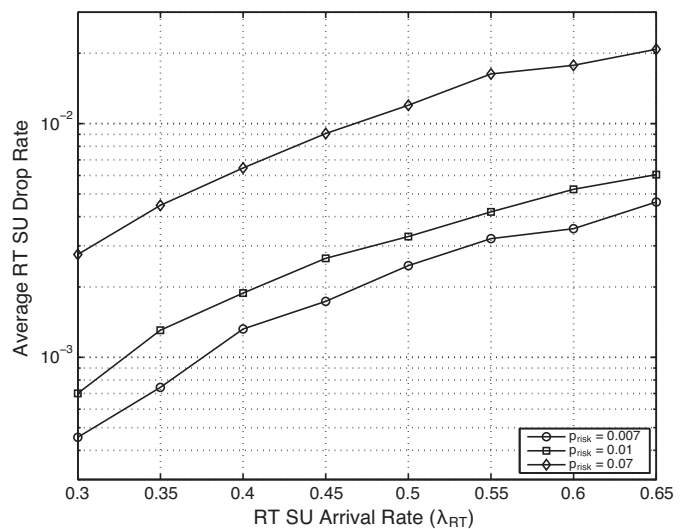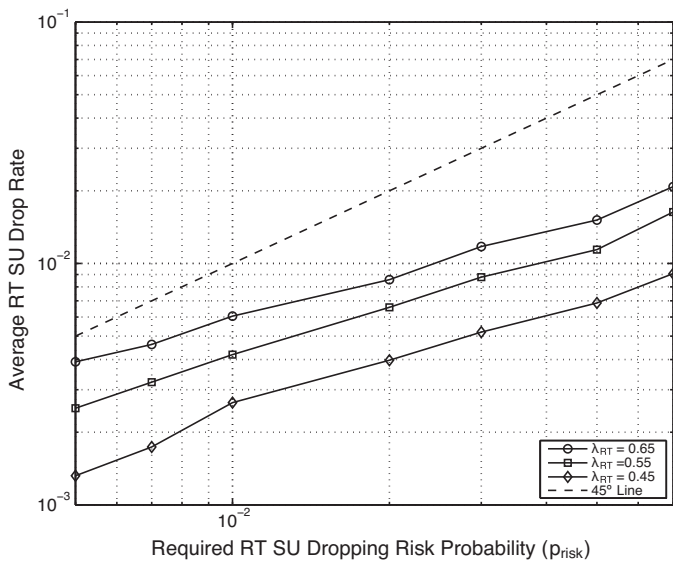

Fig. 6. Average RT SU blocking rate versus size of the installed RBZ that is based on the fixed carrier release algorithm.

Fig. 8. Average RT SU drop rate versus a range of RT SU arrival rates (λ_{RT}) when an RBZ based on predictive carrier release is installed.

Fig. 9. Average RT SU drop rate versus the required dropping risk probability (p_{risk}) for different levels of RT SU arrival rates (λ_{RT}) when an RBZ based on predictive carrier release is installed.

shows how the algorithm successfully protects against a level of RT SU dropping rate that is proportional to the requested p_{risk}. The figure also shows how the average RT SU drop rate approaches p_{risk} as the arrival rate of RT SUs increases. The same observations can be made when looking at Fig. 9, in which the average RT SU drop rate curves are all below the 45° line that represents the point at which the requested dropping risk probability matches the actual drop rate. Again, as the system load increases, the closer the performance curve moves towards the 45° line. We believe that adopting an admission control mechanism that is based on the predictive RBZ algorithm alongside the LagRem algorithm for infeasibility resolution constitutes a complete and versatile solution for the downlink of multiple-user OFDMA CR systems.

6. Conclusion

In this paper, we studied the problem of admission and eviction control of RT SUs in the downlink of multiple-user OFDMA CR networks. We proposed three different ways to protect against service outages by means of an RBZ that gets installed at the time of user admission. Two of the proposed methods are static and require the availability of actual or simulated system measurements to effectively choose the static RBZ size. The third method is dynamic and adapts the size of the RBZ according to the current conditions of the system by predicting resource availability in the future. We also devised three different methods to deal with the problem of system infeasibility that might occur due to PUs activity on the shared resource. We showed how the user removal method based on Lagrange multipliers gave the best performance on long term basis. In conclusion, admission and eviction control mechanisms are essential components of any viable CR system and the proposed techniques are excellent candidates to effectively realize these mechanisms.

Appendix A. List of acronyms

AWGN	Additive White Gaussian Noise
BE	Best Effort
BER	Bit Error Rate
BS	Base Station
CR	Cognitive Radio
MC	Markov Chain
NRT	Non Real-Time
OFDMA	Orthogonal Frequency Division Multiple Access
PU	Primary User
QAM	Quadrature Amplitude Modulation
QoS	Quality of Service
RBZ	Resource Buffer Zone
RT	Real Time
SU	Secondary User

References

[1] Staple G, Werbach K. The end of spectrum scarcity [spectrum allocation and utilization]. IEEE Spectr 2004;41(March (3)):48–52.

[2] Federal Communications Commission Spectrum Policy Task Force. Report of the Spectrum Efficiency Working Group, FCC, Tech Report; November 2002.

[3] Haykin S. Cognitive radio: brain-empowered wireless communications. IEEE J Select Areas Commun 2005;23(February (2)):201–20.

[4] Yucek T, Arslan H. A survey of spectrum sensing algorithms for cognitive radio applications. IEEE Commun Surv Tutorials 2009;11(March (1)):116–30.

[5] Akyildiz IF, Lee W-Y, Vuran MC, Mohanty S. Next generation/dynamic spectrum access/cognitive radio wireless networks: a survey. Comput Netw (Elsevier) 2006;50(September (13)):2127–59.

[6] Wong CY, Cheng R, Lataief K, Murch R. Multiuser OFDM with adaptive subcarrier, bit, and power allocation. IEEE J Select Areas Commun 1999;17(October (10)):1747–58.

[7] Seong K, Mohseni M, Cioffi J. Optimal resource allocation for OFDMA downlink systems. In: Proc. IEEE Int. Symp. Inform. Theory. July 2006. p. 1394–8.

[8] Wong I, Evans B. Optimal downlink OFDMA resource allocation with linear complexity to maximize ergodic rates. IEEE Trans Wireless Commun 2008;7(March (3)):962–71.

[9] Yu W, Lui R. Dual methods for nonconvex spectrum optimization of multicarrier systems. IEEE Trans Commun 2006;54(July (7)):1310–22.

[10] Zhang R, Liang Y-C. Exploiting multi-antennas for opportunistic spectrum sharing in cognitive radio networks. IEEE J Select Top Signal Process 2008;2(February (1)):88–102.

[11] Bansal G, Hossain M, Bhargava V. Optimal and suboptimal power allocation schemes for OFDM-based cognitive radio systems. IEEE Trans Wireless Commun 2008;7(November (11)):4710–8.

[12] Jiao W, Cai L, Tao M. Competitive scheduling for OFDMA systems with guaranteed transmission rate. Comput Commun 2009;32(February (3)):501–10.

[13] Tao M, Liang Y-C, Zhang F. Resource allocation for delay differentiated traffic in multiuser OFDM systems. IEEE Trans Wireless Commun 2008;7(June (6)):2190–201.

[14] Zhang Y, Leung C. Cross-layer resource allocation for mixed services in multiuser OFDM-based cognitive radio systems. IEEE Trans Veh Technol 2009;58(October (8)):4605–19.

[15] Diego Pacheco-Paramo VP, Martinez-Bauset J. Optimal admission control in cognitive radio networks. In: Proc. 4th Int. Conf. on Cognitive Radio Oriented Wireless Networks and Commun. June 2009.

[16] Kim H, Shin K. Optimal admission and eviction control of secondary users at cognitive radio hotspots. In: Proc. 6th Annual IEEE Commun. Society Conf. on Sensor, Mesh and Ad Hoc Commun. and Networks. June 2009. p. 1–9.

[17] Goldsmith A, Chua S-G. Variable-rate variable-power MQAM for fading channels. IEEE Trans Commun 1997;45(October (10)):1218–30.

[18] Proakis JG. Digital communications. 4th ed. New York: McGraw-Hill; 2001.

[19] Niyato D, Hossain E. In: Xiao Y, Hu F, editors. Medium access control protocols for dynamic spectrum access in cognitive radio networks: a survey, invited chapter in Cognitive Radio Networks. CRC Press; 2008.

[20] Su H, Zhang X. Cross-layer based opportunistic MAC protocols for QoS provisionings over cognitive radio wireless networks. IEEE J Select Areas Commun 2008;26(January (1)):118–29.

[21] Tekinay S, Jabbari B. Handover and channel assignment in mobile cellular networks. IEEE Commun Mag 1991;29(November (11)):42–6.

[22] Tripathi N, Reed J, VanLandinoham H. Handoff in cellular systems. IEEE Personal Commun Mag 1998;5(December (6)):26–37.

[23] Gange G, Marriott K, Stuckey PJ. Smooth linear approximation of non-overlap constraints. In: Proc. 5th int. conf. on diagrammatic representation and inference. Berlin, Heidelberg: Springer-Verlag; 2008. pp. 45–59.