Castellanos-Lopez *et al. EURASIP Journal on Wireless Communications and Networking* (2017) 2017:218
DOI 10.1186/s13638-017-1003-3

EURASIP Journal on Wireless Communications and Networking

RESEARCH **Open Access**

Analysis and performance evaluation of resource management mechanisms in heterogeneous traffic cognitive radio networks

S. Lirio Castellanos-Lopez[1*], Felipe A. Cruz-Pérez[2], Genaro Hernandez-Valdez[1] and Mario E. Rivero-Angeles[3]

Abstract

In this paper, the Erlang capacity achieved by the separate or joint use of several resource management mechanisms commonly considered in the literature (spectrum aggregation, spectrum adaptation, call buffering, channel reservation, selective interruption, and preemptive prioritization) to mitigate the effects of secondary call interruptions in cognitive radio networks (CRNs) is evaluated and compared. Heterogeneous traffic is considered, and service differentiation between real-time and elastic (data) traffic is done in terms of their different delay tolerance characteristics. The aim of our investigation is to identify the most relevant resource management mechanisms to improve the performance of the considered networks. As such, both the individual and joint effect of each resource management mechanisms on system performance are evaluated with the objective of comparing the gains in capacity achieved by each resource management mechanism studied in this work. For this purpose, the different resource management mechanisms studied are carefully combined and, for each resulting strategy, optimization of its configuration is presented to maximize the achievable Erlang capacity. For the performance evaluation of the considered strategies in heterogeneous traffic CRNs, a general teletraffic analysis is developed. Numerical results show that spectrum adaptation and call buffering are the mechanisms that best exploit the elasticity of delay-tolerant traffic in heterogeneous traffic CRNs and, therefore, most significantly improve system performance.

Keywords: Cognitive radio networks, Resource management, Performance analysis, Spectrum aggregation, Spectrum adaptation, Preemptive priority, Heterogeneous traffic, Elastic traffic, Call buffering, Transmission delay

1 Introduction

Cognitive radio technology has been proposed for performing dynamic spectrum sharing such that the scarce spectrum in wireless communication systems is utilized more efficiently [1]. Two types of users are defined in cognitive radio networks (CRNs): the licensed or primary users (PUs) who own the license of spectrum usage and the unlicensed or cognitive or secondary users (SUs) having opportunistic access to the licensed spectrum when the PUs are not occupying primary channels (a.k.a., white spaces or spectrum holes). However, when a PU decides to access a spectrum hole, all SUs using this primary channel must relinquish their transmission immediately. These unfinished secondary sessions may be either simply blocked or switched to another white space to continue their transmissions (this process is called spectrum handoff.) If no white spaces are available, the secondary session is dropped (forced to terminate.) Due to this unpredictable nature of spectrum hole availability in a secondary network, in the early years, CRNs were best suited for best effort services without any quality-of-service (QoS) guarantees. Nonetheless, with the ever increasing popularity of cost-efficient voice over the Internet protocol-based applications, mobile internet, social networks, and internet of things, CRNs are required to support real-time and interactive traffic (with most stringent QoS requirements) in addition to

---

* Correspondence: salicalo@correo.azc.uam.mx
[1]Electronics Department, UAM-A, Mexico City, Mexico
Full list of author information is available at the end of the article

best effort (elastic) type of traffic. In this work, it is assumed that secondary data users have a type of service where data integrity is much more important than delay. Hence, all the information of such users has to be relayed successfully to the destination. Then, these users cannot experience a forced termination since in this case data would be lost. Also, these users can experience long queuing delay without degrading their QoS. Services with these characteristics are for instance e-mail, SMS, ftp, ATM-related data, among others.

For these reasons, it is of paramount importance to employ and develop spectrum management strategies for QoS provisioning in CRNs. In this research direction, the most effective mechanism for guarantee QoS for real-time and interactive applications in CRNs is spectrum partitioning (a.k.a. coordinated CRNs) [2–8]. *Spectrum partitioning* means that the total spectrum band is divided into normal spectrum band (in which PUs can preempt SUs) and reserved spectrum band (dynamically leased from the primary network) for exclusive use of the secondary network [3, 5–8]. However, spectrum partitioning implies that the secondary network has to give some (may monetary) incentive to the primary network for the exclusive use of its licensed spectrum [6–8]. CRNs that perform spectrum partitioning are called coordinated CRNs (CCRNs) [6–9]. The use of spectrum partitioning is not considered in this work; however, for the interested reader, in our early work [7, 8], the performance of dynamic spectrum leasing strategies in CCRNs is investigated. Furthermore, the extension of the considered strategies in the paper for CCRNs is straightforward and subject of our current research work.

On the other hand, in CRNs with heterogeneous traffic, *spectrum adaptation* is perhaps the most important mechanism to provide QoS provisioning among the different classes of traffic [10]. *Spectrum adaptation* means that an ongoing secondary elastic session can adaptively adjust the number of assembled channels according to both availability of white spaces and activities of secondary users (i.e., arrivals, departures, interruptions.) In this context, *channel aggregation* (also known as *channel assembling*) means that two or more idle channels are combined together as one channel to provide higher data rate and, thus, reduce the total transmission time of secondary elastic traffic calls [10–24]. On the contrary, elastic traffic may flexibly adjust downward the number of assembled channels to provide, for instance, immediate access to secondary real-time traffic when needed [10–23].[1] Additionally, other relevant mechanisms are identified in the literature to overcome the impact of service interruption of SUs in CRNs [25–27]. Among these mechanisms are the following.

- *Spectrum handoff* process that, as explained above, allows the interrupted secondary calls to be switched to an idle channel, if one is available, to continue its service.
- *Call buffering* strategy that can be used to reduce the forced termination probability of preempted secondary calls. That is, if spectrum handoff is not possible (i.e., no white spaces are available), preempted secondary sessions may be queued into a buffer to wait for a spectrum hole [25, 28]. In this case, when a queued SU finds a new spectrum hole, it is allowed to continue transmitting its information. Although not very common employed in the related literature, a buffer may be used to queue new secondary sessions to improve blocking probability [29].
- *Preemptive priority* which is an effective strategy to provide QoS provisioning among the different classes of secondary traffic [30]. For instance, real-time (delay-sensitive) traffic calls can preempt delay-tolerant (elastic) traffic calls. Of course, those preempted elastic traffic calls may be queued into a buffer to reduce forced termination probability at the expense of increasing transmission delay. In this sense, and according to realistic situations for other types of applications (with limited tolerance to the transmission delay), some expiration mechanisms (reneging due to impatience or thrown away by the system after a certain time) of preempted sessions in the queue or in the system must be considered. However, in this paper, only elastic traffic with unlimited transmission delay tolerance is considered for secondary data users.
- *Double preemptive priority* refers to the strategy in which both primary and secondary voice users can interrupt a data session in case all resources are busy.
- *Channel reservation* is another prioritization mechanism proposed in the literature for QoS provisioning in CRNs [25–27]. In [26], for instance, it is proposed a channel reservation mechanism in which a certain number of white spaces are precluded to be used by new arriving secondary sessions (that is, those white spaces are reserved for spectrum handoff). Blocking access to new secondary sessions, even if there are enough white spaces, lessen the number of secondary sessions to be forcedly terminated. Thus, channel reservation allows the tradeoff between forced call termination and new call blocking probabilities according to the QoS requirements of the secondary traffic. To finely control the performance metrics, the number of reserved channels can take a real value. When a real number of channels is reserved, this strategy is

known as fractional channel reservation [31] and works as follows.

- *Selective interruption* means that upon the arrival of a PU, if a secondary call must be interrupted, the flexible resource allocation (FRA) strategy firstly chooses a data secondary call for being interrupted.

From the arguments just exposed, it is evident that in preemptive priority CRNs with spectrum adaptation and call buffering, transmission delay (defined as the total time that a SU session, that is preempted at least once, spends in the queue during its lifetime in the system) is one of the most important performance metrics for elastic traffic classes (this performance metric is also referred in the literature as file transfer delay [32]). Surprisingly, transmission delay has not been previously considered as performance metric [10–23, 28]. Moreover, previous studies that address CRNs with heterogeneous traffic and spectrum adaptation [10–23] have not considered the use of call buffering, which is a relevant mechanism to exploit the elasticity of best effort traffic in benefit of both real-time and elastic secondary sessions. Exception of this are [28, 33, 34]; however, in these works, neither Erlang capacity nor transmission delay is considered.

In this paper, the Erlang capacity achieved by different resource management mechanisms typically employed in the literature to mitigate the effects of call interruptions (i.e., spectrum adaptation, buffering for new or interrupted secondary data calls, channel reservation, selective interruption, and preemptive prioritization) is evaluated and compared when they are employed either separately or jointly in heterogeneous traffic CRNs. The evaluated Erlang capacity of the cognitive radio system refers to the maximum offered traffic load of secondary users for which the QoS requirements are guaranteed. In our previous work [35], the Erlang capacity of a CRN was analyzed but only VoIP traffic was considered. In [35], neither different service types nor the individual and joint effect of each strategy on system performance was studied. In the present paper, service differentiation between real-time and elastic traffic is considered according to their different delay tolerance characteristics. Also, several strategies that jointly employ different resource management techniques to take advantage of the flexibility and delay tolerance features of elastic traffic are analyzed and evaluated. The studied strategies include as special cases other adaptive assembling strategies that have been recently reported in the open literature [10, 14]. Additionally, a general teletraffic analysis for the performance analysis of the considered strategies in heterogeneous traffic CRNs is developed. A major characteristic of the proposed mathematical analysis requires less state variables than the one proposed in [10]. Hence, the proposed analysis methodology can be used in the context of future works in CRNs where the complexity is considerably high. Contrary to the previously published related works, non-homogenous bandwidth among primary and secondary channels is captured by our mathematical model, that is, the possibility that multiple secondary calls be preempted by the arrival of a PU is now considered. Finally, it is important to note that a similar work was developed in [36]; however, the main contribution of [36] is the formulation of two different spectrum leasing strategies for the performance improvement of CRNs in terms of Erlang capacity and cost per Erlang of capacity. In particular, the Erlang capacity as function of the maximum allowable number of simultaneously rented resources and the fraction of time that resources are rented to achieve such capacity were evaluated. For the strategies developed in [36], spectrum adaptation and spectrum leasing were used jointly with spectrum handoff, call buffering, and preemptive priority mechanism. As in this paper, two service types (i.e., real-time and elastic traffic differentiated according to their different delay tolerance characteristics) were considered in [36]. Contrary to [36], the aim of our investigation in this paper is to identify the most relevant resource management mechanisms (among those commonly considered in the literature: spectrum aggregation, spectrum adaptation, call buffering, channel reservation, selective interruption, and preemptive prioritization) to improve the performance of CRNs. To this end, we develop a new mathematical model to capture the (individual or joint) effect of these resource management mechanisms on system performance. System performance is evaluated in term of Erlang capacity, mean transmission delay of secondary elastic traffic, and forced termination probability of secondary elastic traffic. Building on this, we evaluate and quantify the impact of each individual resource management mechanism and also the impact of jointly employing these mechanisms on system performance. Hence, new numerical results and insights are obtained that were neither presented nor discussed in [36]

The rest of the paper is organized as follows. System model and general assumptions are presented in Section 2. This is followed, in Section 3, by a description of the different strategies studied in this paper. In Section 4, teletraffic analysis is developed for the performance evaluation of the considered spectrum allocation strategies. Numerical results are analyzed and discussed in Section 5, before conclusions are presented in Section 6.

2 System model

The considered opportunistic spectrum sharing wireless system consists of two types of radio users, the PUs of the spectrum and the SUs that opportunistically share the spectrum resources with the PUs in a coverage area. PUs are unaffected by any resource management and admission policies used in the cognitive network, as SUs are transparent to them. The system consists of M primary identical bands, and each primary band is divided into N identical sub-bands (channels). This asymmetry of the primary and secondary channels is based on the fact that PUs and SUs can support different applications. However, the mathematical analysis developed in Section 4 is valid for any integer value of $N \geq 1$. Thus, there are NM channels that are dynamically shared by the primary and secondary users. Two heterogeneous secondary traffic types are considered: real-time (voice) and elastic (data) traffic. Secondary voice users ($\mathrm{S_vUs}$) have absolute priority over secondary data users ($\mathrm{S_dUs}$). Non-homogeneity between the number of channels used by primary and secondary calls is considered. Each PU service occupies a fixed number of N channels. On the other hand, $\mathrm{S_vUs}$ occupy one channel, while all the secondary data calls have the same minimum ($b_{\min}$) and maximum ($b_{\max}$) number of channel requirements. In order to exploit elasticity of secondary data traffic, the number of assembled channels can be adjusted upward or downward in benefit of secondary user satisfaction. That is, data sessions can flexibly adjust upward the number of assembled channels (as long as they are not being used by primary nor secondary voice sessions) to achieve higher data rate and, thus, reduce its total transmission time. (Notice that, for real-time traffic, sufficient QoS is provided given a fixed number of radio resources, and its service time is not modified even if more channels are assembled, but may be improved from the final user point of view when more channels are assembled.) On the other hand, data sessions can also flexibly adjust downward the number of assembled channels to provide immediate service (when needed) to either new secondary (voice or data) calls or ongoing (just preempted) secondary voice calls.

On the other hand, it is considered that PUs, $\mathrm{S_vUs}$, and $\mathrm{S_dUs}$ arrive to the system according to a Poisson process with rates $\lambda^{(\mathrm{P})}$, $\lambda_{\mathrm{v}}^{(\mathrm{S})}$, and $\lambda_{\mathrm{d}}^{(\mathrm{S})}$, respectively. Also, the unencumbered service times for PUs, $\mathrm{S_vUs}$, and $\mathrm{S_dUs}$ (when $\mathrm{S_dUs}$ employ one channel), (represented, respectively, by the random variables $X^{(\mathrm{P})}$, $X_{\mathrm{v}}^{(\mathrm{S})}$, and $X_{\mathrm{d}}^{(\mathrm{S})}$) are considered independent and identically negative-exponentially distributed with mean values $1/\mu^{(\mathrm{P})}$, $1/\mu_{v}^{(\mathrm{S})}$, and $1/\mu_{\mathrm{d}}^{(\mathrm{S})}$, respectively. Note that the transmission time and departure rate of secondary data users depend on the number of channels that they employ. It is assumed that as secondary offered traffic load increases, the same proportion of service requests is maintained for each service considered. Thus, as the considered offered traffic changes, there is a dependence of the arrival rates of the SU voice and data flows; this dependence is capture through the parameter f_{v}. Parameter f_{v} represents the proportion of the total secondary offered traffic that corresponds to secondary voice users (delay-sensitive users). In Table 1, a list of the variables and acronyms used in the manuscript is provided.

3 Considered adaptive spectrum allocation strategies

In this section, the considered and/or studied adaptive allocation strategies are described. They are called strategies E1, E2, E3, E4, E5, E6, and E7. Table 2 presents the characteristics of each of the FRA strategies studied in this section. Strategy E1 does not employ any of the studied resource management mechanisms. Strategy E2 only employs spectrum adaptation. Strategy E3 employs all the studied resource management mechanisms but channel reservation.[2] Strategy E4 employs all the studied resource management mechanisms but data call buffering. Strategy E5 employs spectrum adaptation, selective interruption, and data call buffering. Strategy E6 only employs selective interruption. Strategy E7 employs spectrum adaptation and selective interruption. Note that strategies E1, E2, E5, E6, and E7 are special cases of strategy E3. Figure 1 shows the block diagram of operation of strategies E1, E2, E6, and E7. The block diagrams of operation of strategies E3, E4, and E5 are shown in Figs. 2, 3 and 4, respectively.

Strategies E1 to E7 are differentiated according to the different mechanisms used to mitigate the adverse effects of secondary call interruptions as well as to provide QoS in CRNs (i.e., spectrum handoff, data call buffering, spectrum adaptation, double preemptive priority, channel reservation, and selective interruption mechanisms). The study carried out in this work allows us to investigate the performance improvement due to the use of these resource management mechanisms commonly separately employed in the literature. All of the strategies studied in this work consider the use of spectrum handoff.

Strategy E1 is the most basic one; this strategy is studied in [10] and, in that work, it is known as the non-assembling (NA) strategy. Strategy E2 is also proposed in [10] to improve system performance. Strategy E2 is characterized by using spectrum adaptation. Also, under strategy E2, interrupted secondary calls are selected randomly among data and voice calls (that is, the selective interruption mechanism is not considered), no buffer is used for exploiting the elasticity of the data traffic, and neither channel reservation nor double preemptive mechanisms are considered. Strategy E3 represents the reference strategy, and it is also the most complete strategy in the sense that it employs most of the mechanisms to improve the QoS in CR networks. Strategies E4 and E5 represent two alternatives of our reference E3 strategy (for instance, from Table 2, it is observed that

Table 1 List of variables and acronyms

Variable/ acronym	Description
$\lambda^{(P)}$	Arrival rate of primary users.
$\lambda_{\mathrm{d}}^{(S)}$	Arrival rate of secondary data users.
$\lambda_{\mathrm{v}}^{(S)}$	Arrival rate of secondary voice users.
γ	Indicator of the resource management mechanism. If $\gamma = 1$ the selective interruption mechanism is not employed; otherwise, $\gamma = 0$.
$1/\mu^{(P)}$	Mean value of the unencumbered service time of primary users.
$1/\mu_{\mathrm{d}}^{(S)}$	Mean value of the unencumbered service time of secondary data users when employ one channel.
$1/\mu_{\mathrm{v}}^{(S)}$	Mean value of the unencumbered service time of secondary voice users.
Ω	Set of feasible states.
A_{c}	Number of available channels.
B_{avg}	Average number of resources used by each data session
b_{avg}	the average number of resources used by active $\mathrm{S_dUs}$
b_{max}	Maximum number of channel requirement of data users.
b_{min}	Minimum number of channel requirement of data users.
b_{v}	Number of channel requirement of voice users.
$b(k_2)$	Number of resources occupied by secondary data users.
$E\{L\}$	Average length of queued secondary data users.
$E\{W\}/E\left\{X_{\mathrm{d}}^{(S)}\right\}$	Normalized average transmission delay.
$\mathbf{e}_0$	Vector of value $\mathbf{e}_0 = [1, 0, 0]$.
$\mathbf{e}_1$	Vector of value $\mathbf{e}_0 = [0, 1, 0]$.
$\mathbf{e}_2$	Vector of value $\mathbf{e}_0 = [0, 0, 1]$.
f_{v}	Proportion of the total secondary offered traffic that corresponds to secondary voice users (delay-sensitive users).
FRA	Flexible resource allocation.
$\mathbf{k}$	Vector of state variables of the system.
k_0	Number of primary users in the system.
k_1	Number of secondary voice users being served in the secondary system.
k_2	Number of secondary data users (both buffered and being served).
κ	Number of data sessions using $b_{\mathrm{avg}} + 1$ resources.
κ_a	Maximum number of active data sessions in the system.
M	Primary bands.
N	Number of identical sub-bands in which a primary band is divided it to be used by secondary users.
NA	No channel assembling.
$P(\mathbf{k})$	State probability.
$P(\mathbf{k},\mathbf{X} = x,\mathbf{Y} = y)$	Probability that x ongoing voice secondary sessions and y ongoing data secondary sessions be interrupted in system state $\mathbf{k}$.
$P_{\mathrm{b}}^{(P)}$	New call blocking probability of primary users.
$P_{\mathrm{b\_d}}^{(S)}$	New call blocking probability of secondary data users.
$P_{\mathrm{b\_v}}^{(S)}$	New call blocking probability of secondary voice users.
$P_{\mathrm{ft\_v}}^{(S)}$	Forced call termination probability of secondary voice users.
PU	Primary user.
p_{q}	Probability of queueing in into the buffer a new secondary data session when the occupation of the buffer equals $\lfloor Q_{\mathrm{thr}} \rfloor$. $p_{\mathrm{q}} = Q_{\mathrm{thr}} - \lfloor Q_{\mathrm{thr}} \rfloor$.
p_{v}	Probability of admitting a new session request of a secondary user when the total number of idle sub-channels equals $\lfloor r_v \rfloor + 1$. $p_v = r_v - \lfloor r_v \rfloor$.
Q_{thr}	Buffer admission threshold for new data secondary sessions when there are not enough resources in the system to be served.
r_{v}	Number of reserved channels to prioritize secondary voice calls handoffs.

Table 1 List of variables and acronyms *(Continued)*

Variable/ acronym	Description
SU	Secondary user.
S_dU	Secondary data users.
S_vU	Secondary voice users.
$X^{(P)}$	Random variable that represents the unencumbered service time of primary users.
$X_d^{(S)}$	Random variable that represents the unencumbered service time of secondary data users.
$X_v^{(S)}$	Random variable that represents the unencumbered service time of secondary voice users.
x	Number of interrupted voice sessions.
x^i_{born}	Instant when the i^{th} S_dU arrives to the system.
x^i_{death}	Instant when the i^{th} S_dU successfully completes its service.
$x^{(S,i)}_{d\_ideal}$	Service time when the maximum amount of allowed resources b_{max} are used by a given S_dU.
y	Number of interrupted data sessions.

strategy E4 considers channel reservation for ongoing secondary calls and does not consider the use of a buffer for new data calls). To finely control the different performance metrics, strategy E4 considers fractional channel reservation. Thus, to reserve, on average, a real number r_v of channels, $\lfloor r_v \rfloor + 1$ channels has to be reserved with probability $p_v = r_v - \lfloor r_v \rfloor$ and $\lfloor r_v \rfloor$ channels with the complementary probability [31]. Finally, strategies E6 and E7 are variants of strategies E1 and E2, respectively (specifically, as it is shown in Table 2, these strategies include the use of selective interruption mechanism).

The aim of the considered resource management strategies is to take advantage of the flexibility and delay tolerance of elastic traffic. In order to protect ongoing SU (both voice and data) calls, the reference E3 strategy jointly employs spectrum handoff, call buffering, double preemptive priority for PUs and voice calls, and degradation/compensation mechanisms. For the sake of space, only detailed description of operation of the reference strategy E3 (shown in Fig. 2) is provided. Nevertheless, with the provided description of strategy E3 and the block diagrams shown in Figs. 1, 3, and 4, the understanding of the operation of the other strategies should be straightforward.

Table 2 Characteristics of the analyzed strategies

Used mechanisms	Spectrum allocation strategy						
	E1	E2	E3	E4	E5	E6	E7
Spectrum adaptation	✗	✓	✓	✓	✓	✗	✓
Selective interruption	✗	✗	✓	✓	✓	✓	✓
Buffer for interrupted data calls	✗	✗	✓	✓	✓	✗	✗
Double preemptive	✗	✗	✓	✓	✗	✗	✗
Buffer for new data calls	✗	✗	✓	✗	✗	✗	✗
Channel reservation	✗	✗	✗	✓	✗	✗	✗

In strategy E3, a common buffer is used to queue both new and interrupted S_dU sessions. As explained below, the number of resources assigned to data sessions is dynamically adjusted. In order to protect interrupted S_dUs by the preemptive priority mechanism due to the arrival of either PU or S_vU calls, a victim buffer is employed. In this way, the elasticity of data call traffic is exploited in benefit of QoS satisfaction of S_vUs at the price of increasing mean transmission delay of S_dU calls. Additionally, new data sessions for SUs can be buffered only when there are not enough resources in the system to be served and the number of queued data sessions is less than the threshold Q_{thr}. To finely control de performance metrics, this threshold is considered to have a real value [37]. The required buffer size to queue all interrupted data sessions is of finite number $\lfloor MN/b_{min} \rfloor + \lceil Q_{thr} \rceil$ of locations.[3] Also, data sessions from SUs remain in the system until service termination. For all the evaluated strategies, resume retransmission is considered and queued sessions are served in the order of arrival [38]. Notice that, because of non-homogenous bandwidth among primary and secondary channels, there is the possibility that multiple secondary calls be simultaneously preempted by the arrival of a PU.

Upon the arrival of a S_dU (S_vU) call, if the number A_c of available channels is larger than or equal to b_{min} (one), the call is accepted and the number of assembled channels assigned to it is given by $\min\{b_{max}, A_c\}$ (one). If not enough idle channels exist for a newly arrived SU call, instead of blocking it, a degradation procedure for ongoing S_dU calls take place to try allowing the newly arrived SU call to join the network. The degradation procedure applied for the data traffic calls is based on the equal resource sharing allocation (ERSA) strategy [39]. The degradation procedure works as follows: When it is necessary to degrade ongoing data calls, a cyclic process exists in which ongoing data calls with the

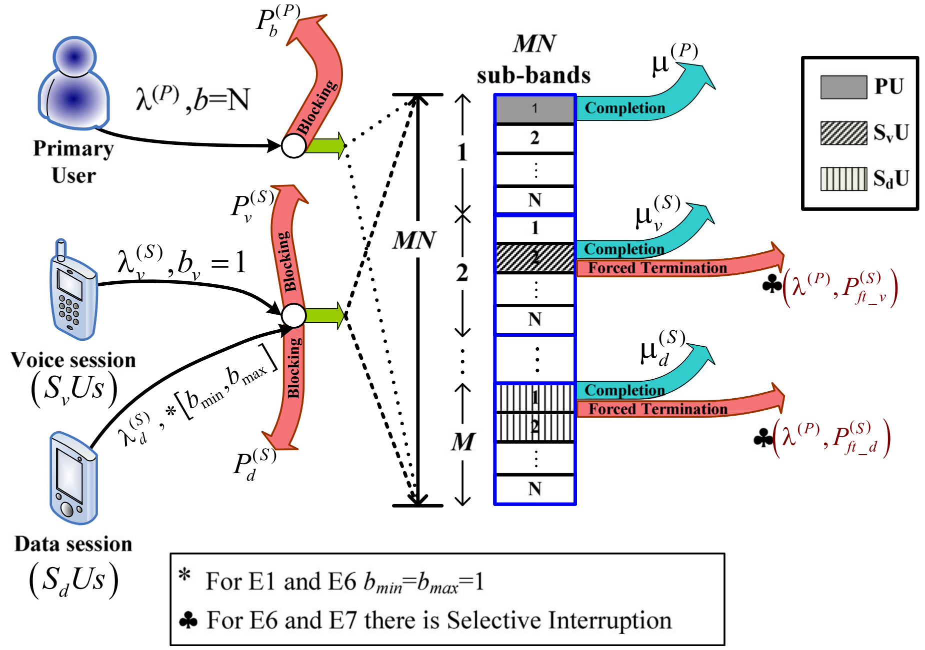

Fig. 1 Block diagram of strategies E1, E2, E6, and E7

largest amount of assembled channels are degraded first and calls with the least number of assembled channels are degraded last. Calls are degraded one channel at a time. This helps with distributing the resources within data traffic calls as equally as possible. The degradation process for an engaged call ends when the number of channels allocated to it decreases up to the minimum required ($b_{\min}$). When it is not possible to perform further spectrum degradation and the number of available channels plus the ones to be released by all ongoing S_dU calls is still not enough to attend the newly arrived SU call, it is queued in the buffer if the admission threshold (Q_{thr}) at the buffer is not exceeded; otherwise, it is blocked. On the other hand, if the new service request is for a voice call, a preemption priority mechanism is triggered. The preemption priority mechanism is used to further protect the access of S_vU call traffic. Under this mechanism, an ongoing data call is interrupted to allow

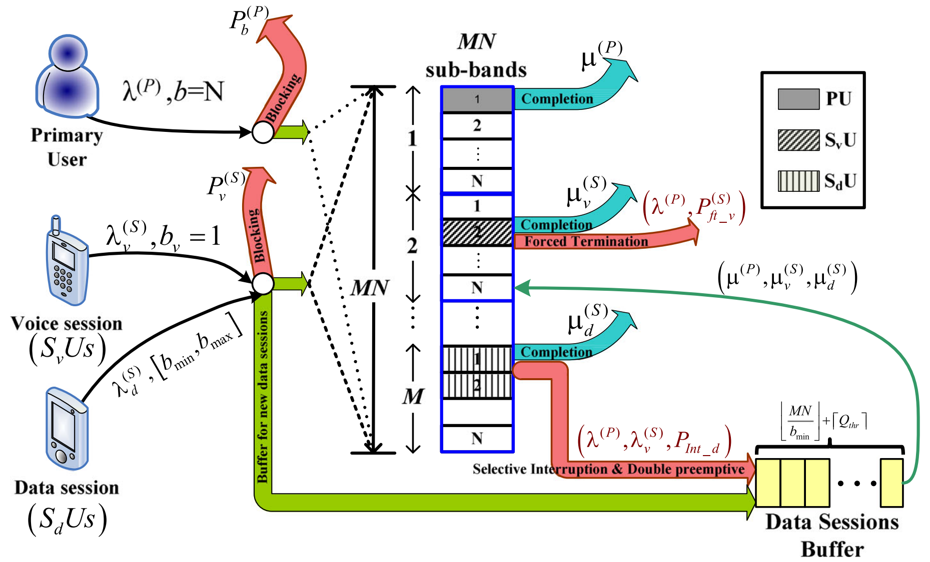

Fig. 2 Block diagram of strategy E3

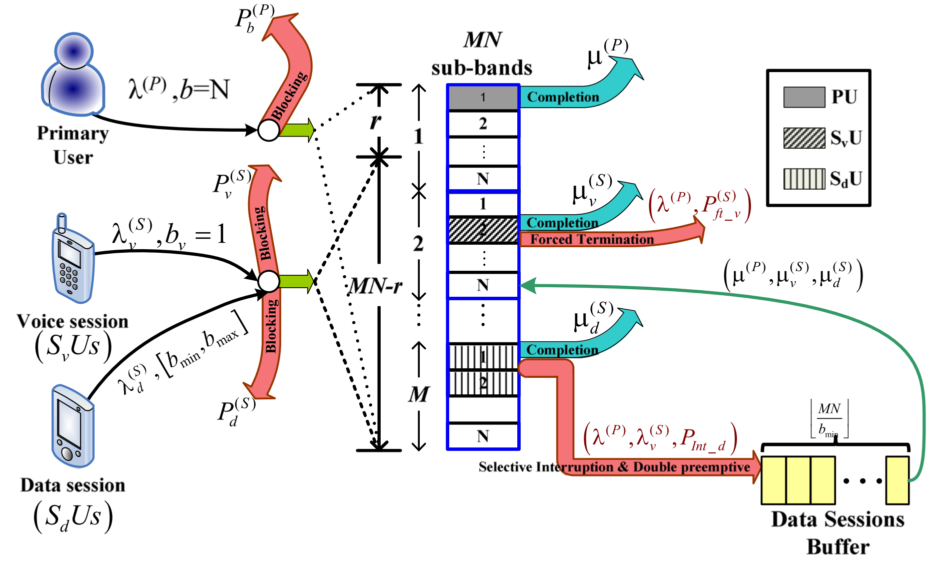

Fig. 3 Block diagram of strategy E4

the newly arrived S_vU call to join the network. As such, a newly arrived S_vU call is blocked only if the number of available channels is smaller than one and no ongoing S_dUs exist in the system.

Under this spectrum adaptation strategy, upon the departure of either PUs or SUs, the available channels are assigned to users at the queue. If after this process there remain vacant channels, a spectrum compensation procedure for secondary data calls is triggered. The compensation procedure works as follows. When it is necessary to compensate ongoing data calls, a cyclic process exists in which ongoing data calls with the smaller amount of assembled channels are compensated first and calls with the larger number of assembled channels are compensated last. Calls are compensated one channel at a time. This helps to achieve the closer to a uniform distribution of channels among data users. The compensation process ends until either all the vacant channels are assigned or when all the existing S_dU calls have assembled b_{max} channels. By using the compensation mechanism, the transmission rate of S_dUs is speeded up when channels become vacant. As such, the channel occupancy times by delay-tolerant traffic calls is reduced.

On the other hand, the spectrum degradation procedure described above is also triggered by the arrival of a PU to

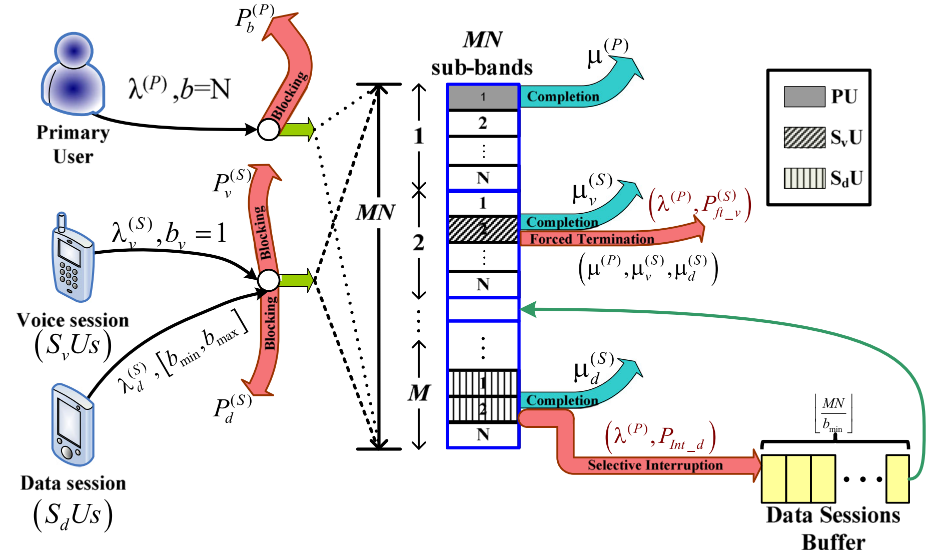

Fig. 4 Block diagram of strategy E5

mitigate forced termination of interrupted SU calls. Again, it is taken advantage of the flexibility and delay tolerance of elastic traffic in benefit of QoS satisfaction of S_vUs at the price of transmission delay for S_dUs. In this case, if the degradation procedure (as well as the preemptive priority mechanism used to protect S_vU calls) fails, the interrupted secondary call is forcedly terminated (that is, no victim buffer is considered for voice sessions).

Note that the previous description applies to strategies E3, E4, and E5 under the following considerations: (i) When no new data sessions are allowed to be queued in the buffer, as in strategy E4, the value of the threshold Q_{thr} is set to 0 (see block diagram of Fig. 3 and compare with block diagram of Fig. 2); (ii) When neither channel reservation nor the new data sessions are buffered, as in strategy E5, both r_{v} and Q_{thr} are set to 0 (see block diagram of Fig. 4 and compare with block diagram of Fig. 2). Additionally, it is important to note that strategy E5 does not consider the double preemptive mechanism where real-time sessions interrupt ongoing elastic traffic sessions, as it is done in strategies E3 and E4. As such, in strategy E5, whenever a secondary voice user arrives to the system and finds no idle resources to be served, the preemption priority mechanism is not triggered and the new S_vU session request is simply blocked.

4 Teletraffic analysis

In this section, the teletraffic analysis for the performance evaluation of the studied FRA strategies is developed. As strategies E1, E2, E4, E5, E6, and E7 are special cases of strategy E3, a mathematical analysis for strategy E3 is developed and, then, it is explained how it must be modified to evaluate the performance of the other strategies. According to the considered spectrum adaptation strategy, resources are assigned in such a way as to achieve as close as possible equal resource sharing among secondary data users.

The state variable of the system is represented by the vector $\mathbf{k} = [k_0, k_1, k_2]$, where k_0 is the number of PUs in the system, k_1 is the number of S_vUs being served in the secondary system, and k_2 is the number of S_dUs (both buffered and being served) in the secondary system. Building on this, a continuous time Markov chain (CTMC) is built in order to perform the teletraffic analysis for the strategy E3. In Fig. 5, the state transition diagram of this system is shown. Let us represent by vector $\mathbf{e}_i$ of size 3 whose all entries are zero except the entry at the ith position which is 1 ($i = 0, 1, 2$). Let Ω be the set of feasible states as

$$\Omega = \left\{ \begin{array}{l} \mathbf{k} = (k_0, k_1, k_2):\ 0 \le k_0 \le M \cap 0 \le k_1 \le MN \\ \cap\ 0 \le k_2 \le \lfloor MN/b_{min} \rfloor + \lceil Q_{thr} \rceil\ \cap\ 0 \le k_1 + b(k_2) \le MN \end{array} \right\} \tag{1}$$

where $b(k_2)$ represents the number of resources occupied by S_dUs. Note that in order to have at least one S_dU being served, the system must satisfy the following conditions: $MN - Nk_0 - k_1 \ge b_{\min} \cap k_2 > 0$. From this observation, it is straightforward to see that the number of resources occupied by S_dUs is given by:

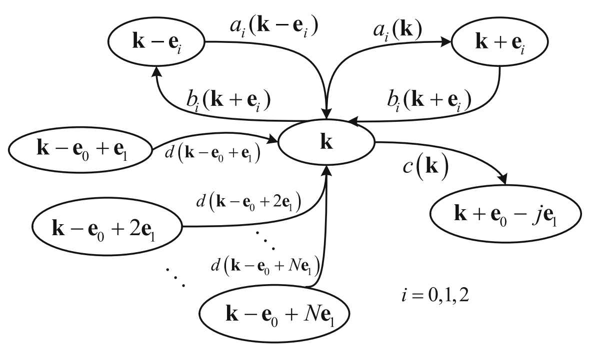

Fig. 5 State transition diagram

$$b(k_2) = \begin{cases} \min(MN - Nk_0 - k_1, b_{\max}k_2) & ; Nk_0 + k_1 + b_{\min}k_2 < MN \\ MN - Nk_0 - k_1 & ; Nk_0 + k_1 + b_{\min}k_2 \ge MN \cap MN - Nk_0 - k_1 \ge b_{\min} \\ 0 & ; \text{otherwise} \end{cases} \tag{2}$$

Note that even if variable k_2 represents only the total number of S_dUs both active and queued in the buffer, it can be used to calculate the number of resources used by an active S_dU. For a particular state $\mathbf{k}$, the maximum number of active S_dUs is $b(k_2)/b_{\min}$. Hence, if k_2 is higher than $b(k_2)/b_{\min}$, it means that there are S_dUs waiting to be served in the buffer. Building on this, let κ_{a} represent the maximum number of active data sessions in the system. Then,

$$\kappa_{\text{a}} = \min\left(k_{2,} \left\lfloor \frac{b(k_2)}{b_{\min}} \right\rfloor\right) \tag{3}$$

and the average number of resources used by active S_dUs is:

$$b_{\text{avg}} = \min\left(\left\lfloor \frac{MN - Nk_0 - k_1}{\kappa_{\text{a}}} \right\rfloor, b_{\max}\right) \tag{4}$$

As such, the number of data sessions using $b_{\text{avg}} + 1$ resources is calculated as $\kappa = \text{mod}(b(k_2), b_{\text{avg}}\kappa_{\text{a}})$ and the number of data sessions using b_{avg} resources is given by $\kappa_a - \kappa$. The remaining $k_2 - \kappa_a$ S_dUs are buffered in the queue. Transition rates due to arrivals and service terminations of PUs or SUs, as well as departure of SUs from the queue are summarized in Table 3, where:

$$a_0(\mathbf{k}) = \begin{cases} \lambda^{(\text{P})} & ; k_0 \ge 0 \cap Nk_0 + k_1 + \gamma b_{\min}k_2 \le (M-1)N \\ 0 & ; \text{otherwise} \end{cases}$$

where $\gamma = 1$ for strategies E1 and E2 because these do not employ the selective interruption mechanism, and $\gamma = 0$ for the other strategies.

Table 3 Transitions from a reference state k = [k_0,…, k_2]

Activity	Rate	Destination state
A PU arrives to the system and there are enough resources in order to be served without interrupting a S_vU.	$a_0(\mathbf{k})$	$(\mathbf{k}+\mathbf{e}_0)$
A S_vU arrives to the system without interrupting a S_dU, or S_dUs are interrupted and buffered.	$a_1(\mathbf{k})$	$(\mathbf{k}+\mathbf{e}_1)$
A S_dU arrives to the system and there are at least $b_{\min}$ resources available.	$a_2(\mathbf{k})$	$(\mathbf{k}+\mathbf{e}_2)$
A PU terminates its service and leaves the system.	$b_0(\mathbf{k})$	$(\mathbf{k}-\mathbf{e}_0)$
A S_vU terminates its service and leaves the system.	$b_1(\mathbf{k})$	$(\mathbf{k}-\mathbf{e}_1)$
A S_dU terminates its service and leaves the system.	$b_2(\mathbf{k})$	$(\mathbf{k}-\mathbf{e}_2)$
A PU arrives to the system and interrupts *j* S_vUs.	$c(\mathbf{k})$	$(\mathbf{k}+\mathbf{e}_0-j\mathbf{e}_1)$

In this transition, the number of resources occupied by PUs and S_vUs is lower or equal than $(M-1)N$ and no SU is interrupted. It is worth mentioning the importance of representing by k_2 the total number of data sessions (both active and queued): the value of k_2 is not modified when secondary data sessions are interrupted due to the fact that these interrupted sessions go from active state to buffered state. As such, the complexity of the analysis is reduced. In order to also employ this expression of $a_0(\mathbf{k})$ as incoming transition rate, the number of PUs has to be at least zero. In fact, for the arrival rate functions $a_i(\mathbf{k})$ $(i = 0, 1, 2)$, it is necessary to specify that the corresponding number of sessions has to be equal or larger than zero to use it as incoming transition rate to the reference state **k**.

To reserve, on average, a real number r_v of channels, $\lfloor r_v \rfloor + 1$ channels has to be reserved with probability $p_v = r_v - \lfloor r_v \rfloor$ and $\lfloor r_v \rfloor$ channels with the complementary probability [31].

$$a_1(\mathbf{k}) = \begin{cases} \lambda_v^{(S)} & ; Nk_0 + k_1 < MN - \lfloor r_v \rfloor - 1 \cap k_1 \geq 0; \\ \lambda_v^{(S)}(1 - p_v) & ; Nk_0 + k_1 = MN - \lfloor r_v \rfloor - 1 \cap k_1 \geq 0; \\ 0 & ; \text{otherwise} \end{cases}$$

This transition rate indicates that new sessions of S_vUs are accepted into the system whenever the number of occupied resources by PUs and S_vUs is either lower than $MN - \lfloor r_v \rfloor - 1$ or with probability $(1 - p_v)$ when the number of occupied resources by PUs and S_vUs equals $MN - \lfloor r_v \rfloor - 1$. Note that the number of resources occupied by S_dUs is not considered in this point due to the double preemptive scheme and the use of k_2 for active and queued S_dUs. In this case, the number of S_vUs has to be higher or equal to zero.

$$a_2(\mathbf{k}) = \begin{cases} \lambda_d^{(S)}; & Nk_0 + k_1 + b_{\min}k_2 \leq MN - b_{\min} \cap k_2 \geq 0 \\ \lambda_d^{(S)}; & k_2 - \kappa_a < \lfloor Q_{\text{thr}} \rfloor \cap Nk_0 + k_1 + b_{\min}k_2 > MN - b_{\min} \cap k_2 \geq 0, \\ \lambda_d^{(S)} p_q; & k_2 - \kappa_a = \lfloor Q_{\text{thr}} \rfloor \cap Nk_0 + k_1 + b_{\min}k_2 > MN - b_{\min} \cap k_2 \geq 0 \\ 0; & \text{otherwise,} \end{cases}$$

where $p_q = Q_{\text{thr}} - \lfloor Q_{\text{thr}} \rfloor$. For this state transition, new sessions of S_dUs can be accepted if at least $b_{\min}$ resources can be obtained by the system as a consequence of the service degradation scheme if they are not being used by PUs or S_vUs. If this degradation cannot be performed, the new sessions can be queued into the buffer if the occupation of the buffer is lower than $\lfloor Q_{\text{thr}} \rfloor$ or equal to $\lfloor Q_{\text{thr}} \rfloor$ with probability p_q or blocked with probability $(1 - p_q)$. When the buffer occupation is higher than the threshold $\lfloor Q_{\text{thr}} \rfloor$, new data secondary sessions are blocked. In this case, the number of S_dUs (active or waiting in the buffer) has to be higher or equal to zero.

$$b_0(\mathbf{k}) = \begin{cases} k_0 \mu^{(P)} & ; k_0 \leq M \\ 0 & ; \text{otherwise} \end{cases}.$$

This transition occurs when a PU completes its service. Hence, the departure rate is proportional to the number of active PUs. When this departure happens, there can be a resource reassignment process but, as already mentioned, the value of k_2 is not altered. In this case, to employ this departure rate as incoming rate to the reference state **k**, the number of sub-channels used has to be lower or equal to the total number of sub-channels in the network.

$$b_1(\mathbf{k}) = \begin{cases} k_1 \mu_v^{(S)} & ; Nk_0 + k_1 + b(k_2) \leq MN \\ 0 & ; \text{otherwise} \end{cases}.$$

This transition rate indicates that the departure rate of S_vUs is also proportional to the number of active S_vUs in the system. Now, the number of occupied sub-channels either by SUs or PUs has to be lower or equal to the total number of sub-channels in the system.

$$b_2(\mathbf{k}) = \begin{cases} b(k_2)\mu_d^{(S)}; Nk_0 + k_1 + b(k_2) \leq MN \cap b_{\min}k_2 \leq MN + b_{\min} \lceil Q_{\text{thr}} \rceil. \\ \\ 0 & ; \text{otherwise} \end{cases}$$

$b\ (k_2)$ In this transition, the departure rate of S_dUs is proportional to the number of resources occupied by the S_dUs and the inverse of the mean service time of these users when they employ only one channel. Note that in order to calculate the death rate of secondary data calls, it is not necessary to know the specific distribution of resources between the different data users. The rationale behind this is that the death rate is given by the sum of the individual death rates of each secondary user which are in turn proportional to the number of assigned resources to each one. Hence, the total death rate of $\mathrm{S_dUs}$ is the product of the death rate of $\mathrm{S_dUs}$ when only one channel is being used multiplied by the total number of resources that are assigned to the $\mathrm{S_dUs}$. This is the reason why our analysis requires less state variables than that presented in [10]. In this case, to employ this departure rate as incoming rate to the reference state $\mathbf{k}$, the total number of sub-channels occupied by SUs and PUs has to be lower or equal to the total number of sub-channels in the system. Also, the total number of active $\mathrm{S_dUs}$ and waiting in the buffer has to be lower or equal to the maximum number of users that can be accepted in the network, $\left\lfloor {MN}\big/{b_{\min}} \right\rfloor + \lceil Q_{\mathrm{thr}} \rceil$.

$$c(\mathbf{k}) = \begin{cases} \lambda^{(\mathrm{P})} & ; k_0 < M \cap Nk_0 + k_1 > (M-1)N \\ 0 & ; \text{otherwise} \end{cases}.$$

This transition rate indicates that new PU sessions are accepted with rate $\lambda^{(\mathrm{P})}$ when the number of occupied resources by PUs is lower than M and the total number of sub-channels used by primary and secondary users is higher than $(M - 1)N$. The corresponding incoming transition rate to the reference state $\mathbf{k}$ in this case is

$$d(\mathbf{k}) = \begin{cases} \lambda^{(\mathrm{P})} & ; 0 < k_0 \le M \cap Nk_0 + k_1 > (M-1)N. \\ 0 & ; \text{otherwise} \end{cases}$$

Given feasible states and their transitions in the CTMC, the global set of balance equations can be constructed as follows

$$\left[\sum_{i=0}^{2} a_i(\mathbf{k}) + \sum_{i=0}^{2} b_i(\mathbf{k}) + c(\mathbf{k})\right] P(\mathbf{k}) = \sum_{i=0}^{2} a_i(\mathbf{k}-\mathbf{e}_i) P(\mathbf{k}-\mathbf{e}_i) + \sum_{i=0}^{2} b_i(\mathbf{k}+\mathbf{e}_i) P(\mathbf{k}+\mathbf{e}_i) + \sum_{j=1}^{N} d(\mathbf{k}-\mathbf{e}_0 + j\mathbf{e}_1) P(\mathbf{k}-\mathbf{e}_0 + j\mathbf{e}_1) \tag{5}$$

Based on these equations and the normalization equation, the state probabilities $P(\mathbf{k})$ are obtained. The performance metrics of interest in this CRN when the spectrum adaptation strategy is enabled are as follows. Let $P_{\mathrm{b}}^{(\mathrm{P})}$, $P_{\mathrm{b\_v}}^{(\mathrm{S})}$, and $P_{\mathrm{b\_d}}^{(\mathrm{S})}$ represent, respectively, the new call blocking probabilities of PUs, $\mathrm{S_vUs}$, and $\mathrm{S_dUs}$, while $P_{\mathrm{ft\_v}}^{(\mathrm{S})}$ represents the forced call termination probability of $\mathrm{S_vUs}$.

- Blocking probability for PUs, which occurs when the total number of channels of the system are being used by PUs:

$$P_{\mathrm{b}}^{(\mathrm{P})} = \sum_{\Omega | k_0 = M} P(\mathbf{k}) \tag{6}$$

- Blocking probability for secondary voice users depends on whether it is used simple or double preemption priority. Then, two different expressions for the new call blocking probability of secondary voice users are derived.
- *Simple preemption priority.* For the strategies that employ the simple preemption priority mechanism, blocking of secondary voice user requests occurs when the total number of idle sub-channels is lower than $\lfloor r_{\mathrm{v}} \rfloor + 1$ or with probability p_{v}, when the total number of idle sub-channels equals $\lfloor r_{\mathrm{v}} \rfloor + 1$. Then, the blocking probability for secondary voice users is given by

$$P_{b\_v}^{(\mathrm{S})} = \sum_{\Omega | Nk_0 + k_1 + b_{\min} k_2 > MN - \lfloor r_v \rfloor - 1} P(\mathbf{k}) + \sum_{\Omega | Nk_0 + k_1 + b_{\min} k_2 = MN - \lfloor r_v \rfloor - 1} p_{\mathrm{v}} P(\mathbf{k}) \tag{7a}$$

- *Double preemption priority.* For the strategies that employ double preemption priority mechanism, blocking of secondary voice user requests occurs when the total number of sub-channels used by both PUs and SvUs is greater than $MN - \lfloor r_{\mathrm{v}} \rfloor - 1$ or with probability p_{v}, when the total number of sub-channels used by both PUs and SvUs equals $MN - \lfloor r_{\mathrm{v}} \rfloor - 1$. Then, the blocking probability for secondary voice users is given by

$$P_{b\_v}^{(\mathrm{S})} = \sum_{\{\mathbf{k} \in \Omega | Nk_0 + k_1 > MN - \lfloor r_{\mathrm{v}} \rfloor - 1\}} P(\mathbf{k}) + \sum_{\{\mathbf{k} \in \Omega | Nk_0 + k_1 = MN - \lfloor r_{\mathrm{v}} - 1 \rfloor\}} p_{\mathrm{v}} P(\mathbf{k}) \tag{7b}$$

- Forced termination probability for secondary voice users depends on whether prioritized interruption is used or not. Then, two different expressions for the forced termination probability of secondary voice users are derived.
- *Without prioritized interruption*. For the strategies that do not employ the prioritized interruption mechanism, secondary voice sessions are interrupted by PU users. This occurs with the probability that the PU is assigned the same channels being used by secondary voice users when the number of occupied channels is higher than $(M - 1)N$. Voice sessions are interrupted with the same rate that PU find more than $(M - 1)N$ occupied channels multiplied by the probability that a SvU is using the same channel assigned to a new PU given as $k_1/(k_1 + b_{\min}k_2)$ and divided by the rate that voice calls are accepted into the system given by $\lambda_{\mathrm{v}}^{(\mathrm{S})}$

$\left(1-P_{\mathrm{b\_v}}^{(\mathrm{S})}\right)$. Then, the forced termination probability of secondary voice users is given by

$$P\mathrm{ft\_v}(\mathrm{S}) = \frac{\lambda^{(\mathrm{P})}}{\lambda_{\mathrm{v}}^{(\mathrm{S})}\left(1-P_{\mathrm{b\_v}}^{(\mathrm{S})}\right)} \sum_{\left\{\mathbf{k}\in\Omega \middle| \begin{array}{l} k_0 < M \cap k_1 > 0 \\ \cap N k_0 + k_1 + b_{\min} k_2 > (M-1)N \end{array}\right\}} \frac{k_1}{k_1 + b_{\min} k_2} P(\mathbf{k}) \tag{8a}$$

- *With prioritized interruption.* For the strategies that employ the prioritized interruption mechanism, forced termination of secondary voice users occurs when, upon the arrival of PUs, the number of occupied resources is higher than $(M - 1)N$ and there are no active data sessions that can be interrupted. Then, the forced termination probability of secondary voice users is given by

$$P_{\mathrm{ft\_v}}^{(\mathrm{S})} = \frac{\lambda^{(\mathrm{P})}}{\lambda_{\mathrm{v}}^{(\mathrm{S})}\left(1-P_{\mathrm{b\_v}}^{(\mathrm{S})}\right)} \sum_{\Omega \left| \begin{array}{l} k_0 < M \cap k_1 > 0; b(k_2) = 0 \\ \cap N k_0 + k_1 + b_{\min} k_2 > (M-1)N \end{array}\right.} P(\mathbf{k}) \tag{8b}$$

- Blocking probability for secondary data users which occurs when the total number of idle sub-channels is lower than $b_{\min}$ and the buffer is no longer accepting new sessions of $\mathrm{S_dU}$s:

$$P_{\mathrm{b\_d}}^{(\mathrm{S})} = \sum_{\Omega \left| \begin{array}{c} N k_0 + k_1 + b_{\min} k_2 > MN - b_{\min} \\ \cap k_2 - \kappa_{\mathrm{a}} > \lfloor Q_{\mathrm{thr}} \rfloor \end{array}\right.} P(\mathbf{k}) + \sum_{\Omega \left| \begin{array}{c} N k_0 + k_1 + b_{\min} k_2 > MN - b_{\min} \\ \cap k_2 - \kappa_{\mathrm{a}} = \lfloor Q_{\mathrm{thr}} \rfloor \end{array}\right.} \left(1-p_{\mathrm{q}}\right) P(\mathbf{k}) \tag{9}$$

- Average queue length which is the sum of $\mathrm{S_dU}$s waiting in the queue in each state multiplied by the probability of being in that state:

$$E\{L\} = \sum_{\Omega} (k_2 - \kappa_{\mathrm{a}}) P(\mathbf{k}) \tag{10}$$

where $(k_2 - \kappa_{\mathrm{a}})$ is the total number of queued secondary sessions (of both data and voice service sessions).

- Average number of resources used by each data session

$$B_{\mathrm{avg}} = \frac{\sum_{\Omega|\kappa_{\mathrm{a}}>0} \frac{b(k_2)}{\kappa_{\mathrm{a}}} P(\mathbf{k})}{\sum_{\Omega|\kappa_{\mathrm{a}}>0} P(\mathbf{k})} \tag{11}$$

All numerical results obtained for the different performance metrics are obtained analytically but for the average normalized transmission delay. The average normalized transmission delay is evaluated through simulations and is calculated as:

$$\frac{E\{W\}}{E\left\{X_d^{(S)}\right\}} = \frac{\sum_{i=1}^{n} \quad x_{death}^{i} \; - x_{born}^{i} \; - x_{d\_ideal}^{(S,i)}}{n E\left\{X_d^{(S)}\right\}} \tag{12}$$

where x_{death}^{i} and x_{born}^{i} are the instant when the ith $\mathrm{S_dU}$ successfully completes its service and arrives to the system respectively. On the other hand, $x_{\mathrm{d\_ideal}}^{(\mathrm{S,i})}$ is the service time when the maximum amount of resources $b_{\max}$ is used by a given $\mathrm{S_dU}$, and n is the number of data calls successfully completed. As shown in Appendix 1 of [40], the call interruption process of secondary calls due to the arrival of primary users in cognitive radio networks is not a Poissonian one. As such, the interruption probability of secondary users cannot be obtained in a straightforward manner. This is the reason why the average transmission delay of SU elastic flows was not been computed by Little's Law.

For resource management strategies that allow neither new data sessions to be queued in the buffer nor the use of channel reservation for new secondary voice session requests, parameters Q_{thr} and r_{v} are set to zero; this is the case of strategies E1, E2, E6, and E7. Additionally, due to the fact that strategies E1 and E6 do not consider the use of spectrum adaptation $b_{\min} = b_{\max} = 1$. The set of feasible states Ω for the strategies E1, E2, E6, and E7 should include the condition $Mk_0 + k_1 + b(k_2) \le MN$ because they do not buffer interrupted data sessions.

Similarly, due to the fact that strategies E1, E2, E5, E6, and E7 do not employ the double preemptive mechanism, the transition from state $\mathbf{k}$ to state $(\mathbf{k} + \mathbf{e}_1)$ is only possible when the total number of busy resource is less or equal to $MN - 1$. Thus, for strategies E1, E2, E5, E6, and E7, the arrival rate $a_1(\mathbf{k})$ must be rewritten as follows:

$$a_1(\mathbf{k}) = \begin{cases} \lambda_{\mathrm{v}}^{(\mathrm{S})} & ; N k_0 + k_1 + b_{\min} k_2 < MN - \lfloor r_{\mathrm{v}} \rfloor - 1 \cap \; k_1 \ge 0 \\ \lambda_{\mathrm{v}}^{(\mathrm{S})} (1 - p_{\mathrm{v}}) & ; N k_0 + k_1 + b_{\min} k_2 = MN - \lfloor r_{\mathrm{v}} \rfloor - 1 \cap \; k_1 \ge 0. \\ 0 & ; \text{otherwise} \end{cases}$$

For strategies that do not consider either selective interruption (E1 and E2) or buffer for interrupted data sessions (E1, E2, E6, and E7), when a new PU arrival occurs and the total number of white spaces is less than N, it is imperative to interrupt enough ongoing secondary sessions (voice and/or data) in order to have at least N available channels to attend this primary arrival call. The number of channels occupied by secondary sessions that needs to be released to complete the N available resources can be computed as $n = Nk_0 + k_1 + b_{\min}k_2 - (M - 1)N$. In these situations, there exists different ways of interrupting ongoing secondary (voice and/or data) sessions to obtain the n required available resources. In this case, the primary arrival event triggers the transition from state $\mathbf{k}$ to the state $\mathbf{k} + \mathbf{e}_0 - x\mathbf{e}_1 - y\mathbf{e}_2$, where $x = 0, 1, \ldots, \min(n, k_1)$ represents the number of interrupted voice sessions while $y = n - x$ represents the number of interrupted secondary data calls. Notice that for the incoming rate to the state $\mathbf{k}$ in strategies E1 and E2, x and y must be modeled as random variables due to the fact that there exist more than one possible state from which the state $\mathbf{k}$ can be reached. Thus, for every state under which occurs a primary session request and the number of available resources is greater than $(M - 1)N$, the probability that x ongoing voice secondary sessions and y ongoing data secondary sessions be interrupted is given by the following hyper-geometric distribution

$$P(\mathbf{k}, \mathbf{X} = x, \mathbf{Y} = y) = \frac{\binom{k_1}{x}\binom{b(k_2)}{y}}{\binom{k_1 + b(k_2)}{n}}$$

for $x = 0, 1, \ldots, \min(n, k_1)$ and $y = n - x$.

Thus, in order to analyze strategies E1 and E2, the transition rate $c(\mathbf{k})$ involved with the transition from state $\mathbf{k}$ to state $\mathbf{k} + \mathbf{e}_0 - x\mathbf{e}_1 - y\mathbf{e}_2$ must be computed as follows.

$$c(\mathbf{k}, x, y) = \begin{cases} P(\mathbf{k}, \mathbf{X} = x, \mathbf{Y} = y)\lambda^{(\mathrm{P})} & ; k_0 < M \cap k_1 \geq 0 \cap k_2 \geq 0 \\ & \cap (M-1)N < Nk_0 + k_1 + b_{min}k_2 \leq MN \\ 0 & ; \text{otherwise} \end{cases}.$$

The corresponding incoming transition rate to the reference state $\mathbf{k}$ in this case is

$$d(\mathbf{k}, x, y) = \begin{cases} P(\mathbf{k}, \mathbf{X} = x, \mathbf{Y} = y)\lambda^{(\mathrm{P})} & ; 0 < k_0 \leq M \cap k_1 \geq 0 \cap k_2 \geq 0 \\ & \cap (M-1)N < Nk_0 + k_1 + b_{min}k_2 \leq MN \\ 0 & ; \text{otherwise} \end{cases}.$$

On the other hand, due to the fact that strategies E6 and E7 employ selective interruption, the number of interrupted voice and data sessions (x and y, respectively) is deterministic quantities. Thus, in order to analyze strategies E6 and E7, $P(\mathbf{k}, \mathbf{X} = x, \mathbf{Y} = y) = 1$ and the transition rates $c(\mathbf{k}, x, y)$ and $d(\mathbf{k}, x, y)$ equals $\lambda^{(\mathrm{P})}$ only when $(x > 0 \cap k_2 = y) \cup (x = 0 \cap k_2 = n)$.

Finally, the global set of balance equations for strategies E1, E2, E6, and E7 can be constructed as follows:

$$\begin{aligned} &\left[\sum_{i=0}^{2} a_i(\mathbf{k}) + \sum_{i=0}^{2} b_i(\mathbf{k}) + \sum_{x=0}^{\min(n,k_1)} c(\mathbf{k}, x, y)\right] P(\mathbf{k}) \\ &= \sum_{i=0}^{2} a_i(\mathbf{k} - \mathbf{e}_i) P(\mathbf{k} - \mathbf{e}_i) + \sum_{i=0}^{2} b_i(\mathbf{k} + \mathbf{e}_i) P(\mathbf{k} + \mathbf{e}_i) \\ &+ \sum_{x=1}^{N} d(\mathbf{k} - \mathbf{e}_0 + x\mathbf{e}_1 + (n-x)\mathbf{e}_2, x, y) P(\mathbf{k} - \mathbf{e}_0 + x\mathbf{e}_1 + (n-x)\mathbf{e}_2). \end{aligned}$$

5 Numerical results

The goal of the numerical evaluations presented in this section is to verify the applicability as well as the accuracy and robustness of the mathematical model developed in Section 4 to investigate the performance of spectrum adaptation strategies for CRN with heterogeneous traffic described in previous sections.

The performance of the studied flexible resource allocation (FRA) strategies is evaluated in terms of the maximum Erlang capacity of the CRN (defined as the maximum offered traffic load of secondary users for which QoS requirements are guaranteed in terms of blocking probability of new voice and data calls and forced termination probability of secondary voice calls) [35]. In order to calculate the Erlang capacity of the CR system, the maximum acceptable value of the new call blocking probability for voice and data secondary traffic and forced termination probability of voice secondary calls equals 2%. In this work, a bound on the transmission delay of SU elastic flows to determine the Erlang capacity is not established since it is assumed that secondary data users have a type of service where data integrity is much more important than delay. These users can experience long queuing delay without degrading their QoS. Additionally, in this work we compare the performance of strategies where secondary data sessions experience forced termination (i.e., E1, E2, E6, and E7) against that of strategies where secondary data sessions do not experience forced termination (i.e., E3, E4, and E5). As such, strategies E3, E4, and E5 entail higher transmission delays due to the fact that no data sessions of secondary users are forced to terminate. Then, for these reasons, establishing a bound on the transmission delay of SU elastic flows to determine the Erlang capacity would yield both not suitable performance evaluation of the system (because secondary data sessions do not have delay requirement) and unfair performance comparison among the different resource allocation strategies (because some strategies guarantee the reliability of the delivery of information at the expense of transmission delay and others not).

It is important to note that the forced termination probability for S_dUs under the strategies E3, E4, and E5 is equal to zero due to the fact that the interrupted data calls are not forced to terminate their transmissions, instead preempted data calls are queued into the buffer. For strategies E3 and E4, maximum Erlang capacity is achieved by optimizing the control parameters Q_{thr} and r_v, respectively. Maximum Erlang capacity for E4 is obtained by optimizing the number of reserved channels to prioritize new voice call attempts over data call requests. The optimal value of the number of reserved channels r_v is systematically searched by using the fact that new call blocking (forced termination) probability for S_vUs is a monotonically increasing (decreasing) function of the number of reserved channels. Once a value of the number of reserved channels for which the new call blocking probability for S_vUs and/or S_dUs and/or forced termination probability for S_vUs achieve its maximum acceptable value is found, another offered traffic load is tested. The capacity maximization procedure ends when the new call blocking probability for S_vUs and S_dUs, and forced call termination probability for S_vUs achieve their maximum acceptable values. A similar procedure is followed for calculation of the threshold Q_{thr} under the strategy E3. The threshold Q_{thr} is systematically searched by using the fact that new call blocking probability for S_dUs is a monotonically decreasing function of threshold Q_{thr}. For strategies E1, E2, E6, and E7, the maximum value of the traffic load is systematically searched by using the fact that, for these strategies, new call blocking and forced termination probabilities for S_vUs and S_dUs are monotonically increasing functions of the offered traffic. Thus, different values for the traffic load are tested (using, for example, the well-known bisection algorithm). The maximum traffic load (Erlang capacity) is found when at least one of the performance metrics reaches its maximum acceptable value and the others are below of their respective maximum acceptable values.

Unless otherwise specified, the following system and teletraffic parameters are employed in this section: total number of primary bands $M = 6$; number of channels per primary band $N = 3$; minimum and maximum number of channels required for S_dUs calls (elastic traffic) $b_{min} = 1$ and $b_{max} = 3$; the number of channels required for S_vUs calls (real-time traffic) is fixed to 1; of course, a fixed number of N channels is required by any primary call. We use as reference the evaluation scenario considered in [10] where mean service time for primary users is 1/0.5, voice services have a mean service time of 1/0.6, and secondary data users have a mean time of 1/0.82, and the proportion of voice traffic $f_v = 0.4767$. The mean service times for PUs, S_vUs, and S_dUs are $1/\mu^{(P)} = 1/0.5$, $1/\mu_v^{(S)} = 1/0.6$, and $1/\mu_d^{(S)} = 1/0.82$, respectively.

Figures 6 and 7, respectively, show the maximum Erlang capacity and the normalized mean transmission delay as function of the *utilization factor of primary channels* (defined as the ratio between the primary carried load and the total number of primary channels). From Fig. 6, it is observed that, for all the analyzed spectrum adaptation strategies, as the *utilization factor of primary channels* (hereafter denoted by ρ) increases, the Erlang capacity decreases. This is an expected result due to the fact that as the primary traffic load increases, more secondary calls are interrupted in detrimental of system performance. Figure 6 shows that, for values of ρ less than about 0.2, our reference strategy (E3) diminishes this effect compared to the other strategies. On the other hand, the improvement due to the use of spectrum adaptation can be quantitative and qualitatively obtained from Fig. 6 by comparing strategy E2 versus strategy E1. For instance, for a value of $\rho = 0.2$, the Erlang capacity of strategy E2 increases 30.5% compared to the one achieved by strategy E1. Also, capacity improvement due to the jointly use of new data call buffering and double preemptive mechanisms can be obtained by comparing strategy E3 versus strategy E5. For instance, for a value of $\rho = 0.2$, the Erlang capacity of strategy E3 increases 70.2% compared to the one achieved by strategy E5. In general, Fig. 6 shows that, for values of ρ smaller than about 0.2, our reference E3 strategy considerably outperforms all the other strategies. For instance, for a value of $\rho = 0.2$, the Erlang capacity of strategy E3 increases 60% (72%) compared to the one achieved by strategy E2 (E4). This is mainly due to the use of a buffer for new secondary data calls; this buffer allows the strategy E3 (compared against strategies E2 and E4) to exploit more efficiently the elasticity of data traffic in benefit of system capacity. Figure 7 shows that this capacity gain is

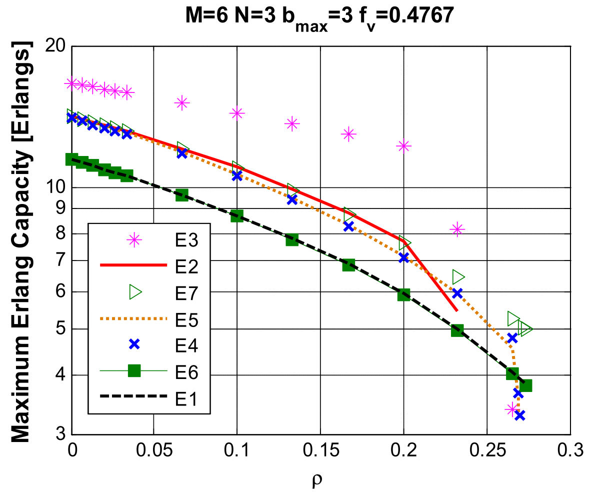

Fig. 6 Maximum Erlang capacity as a function of ρ, for $f_v = 0.4767$

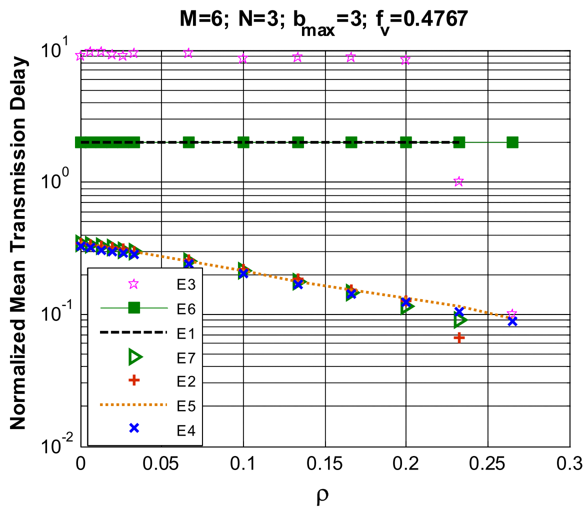

Fig. 7 Normalized mean transmission delay as a function of ρ, for $f_v = 0.4767$

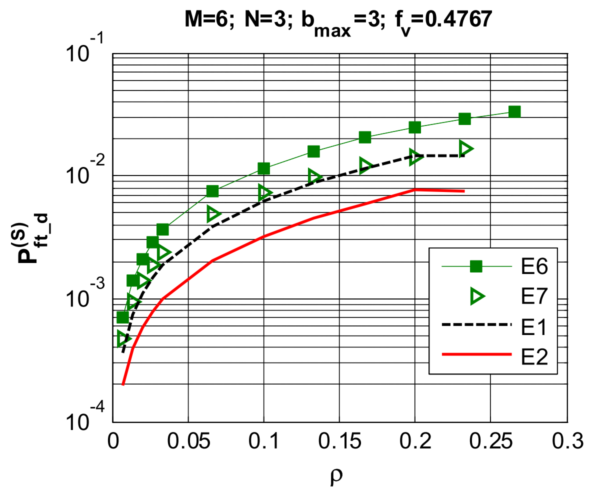

Fig. 8 Forced termination probability of data traffic as a function of ρ, for $f_v = 0.4767$

achieved at expenses of increasing the mean transmission delay of data (elastic) traffic. Additionally, under the strategy E3 (and its variants: strategies E4 and E5), the forced termination probability of data calls is equal to zero (i.e., the integrity of accepted data traffic is guaranteed) due to the use of a buffer for interrupted secondary data calls. This feature is especially useful in heterogeneous CRNs where the data traffic is represented by background services or some interactive applications with relaxed delay constraints. Notice from Fig. 7 that under strategy E3, the normalized mean transmission delays is always lower than 10. Notice from Fig. 7 that the mean transmission delay of SUs decreases with the primary traffic load; this behavior may be perceived as counterintuitive due to the fact that the opposite effect would be expected; however, this behavior can be explained as follows. First of all, we have to mention that the behavior observed in Fig. 7 has to be interpreted in conjunction with the results presented in Fig. 6. Indeed, Fig. 6 shows that the Erlang capacity of the cognitive radio system decreases as the PU traffic load increases. Recall that this Erlang capacity corresponds to the maximum offered traffic load of secondary users where the QoS requirements are guaranteed in terms of blocking probability of new voice and data calls and forced termination probability of secondary voice calls. Hence, as the PU traffic load increases; less new SUs are admitted to the network which in turn decreases the resource occupation. As such, transmission delay is reduced as the PU traffic load increases. From this, Fig. 7 shows numerical results considering that the system is operating with the PU traffic load presented in Fig. 6. Figure 8 shows forced termination probability of data traffic as a function of ρ for strategies E1, E2, and its variants E6 and E7. Notice that, due to the use of selective interruption, strategy E6 (E7) achieves a higher forced termination probability than strategy E1 (E2).

On the other hand, for values of ρ greater than 0.2, Fig. 6 shows that strategy E3 achieves less Erlang capacity than most of the analyzed strategies; moreover, strategies E4 and E6 are the ones that have the slowest decreasing rates of Erlang capacity as the utilization factor of primary channels increases. This positive behavior of strategy E4 is achieved by the use of channel reservation and by precluding the use of a buffer for new data calls (that is, strategy E4 interchange capacity by blocking probability of new data calls). Similarly, as Fig. 8 shows, the beneficial behavior of strategy E6 regarding Erlang capacity is paid with forced termination probability of data calls (due to the use of selective interruption). Thus, channel reservation mechanism and avoiding the use of queue for buffering new data calls are relevant aspects for improving system performance in scenarios with high primary traffic load.

By comparing in Fig. 6 the plots that correspond to the strategies E4 and E5, it is observed that the performance of both strategies is practically the same for low to moderate primary traffic load ($\rho < 0.25$). This behavior indicates that, in these scenarios, the double preemptive and channel reservation mechanisms do not contribute to improve system Erlang capacity. On the other hand, and for scenarios where $\rho < 0.2$, by comparing strategy E4 against strategy E3, it is observed from Fig. 6 that strategy E3 considerably improves system Erlang capacity. This behavior indicates that buffering new (blocked) secondary calls is an important mechanism to improve system capacity under low primary traffic load. Similarly, for scenarios where $\rho < 0.2$, by comparing strategies E1 against E6 (and E2 against E7), it is observed from Fig. 6 that their performance are practically the same. This behavior indicates that the capacity improvement due to the use of selective interruption is negligible. Finally, Table 4 shows the maximum percentage on Erlang capacity gain given for every single considered resource management mechanism when $M = 6$, $N = 3$, $f_v = 0.4767$, $b_{\max} = 3$, and a maximum value of $\rho = 0.23$ is considered. From Table 4, and the previous discussion of this section, it is concluded that the most relevant mechanisms that improve system performance in CRNs with heterogeneous traffic are spectrum adaptation and buffering new data calls.

Figures 9, 10 and 11, respectively, show numerical results for new call blocking probability of voice SUs, new call blocking probability of data SUs, and forced termination probability of voice SUs as function of the utilization factor of primary channels. For every value of the utilization factor of primary channels considered in Figs. 9, 10, 11, the performance of each resource management strategy is evaluated under its optimal operating configuration (that is, the configuration that maximizes Erlang capacity). In this sense, for every value of the utilization factor of primary channels shown in Figs. 9, 10, 11, the corresponding value of the secondary traffic load shown in Fig. 7 is considered. In Figs. 9, 10 and 11, label "A" ("S") stands for analytical (simulation) results. From Figs. 9, 10 and 11, perfect agreement between analytical and simulation results is observed; this verifies the accuracy of the mathematical model developed in Section 4. From Fig. 9 (Fig. 10), it is observed that, for values of the utilization factor of primary channels smaller than 0.2, the secondary blocking probability for voice (data) users attained for strategies E4 and E5 (strategy E3) reaches its maximum allowable value. This means that for values of the utilization factor of primary channels smaller than 0.2, the secondary blocking probability is the metric that limits Erlang capacity. Similarly, from Fig. 11, it is observed that the secondary forced termination probability for voice users attained by the different resource management strategies reaches its threshold for values of utilization factor of primary channels greater than 0.2. This means that under these scenarios, the secondary forced termination probability is the metric that limits Erlang capacity.

6 Conclusions

The Erlang capacity achieved by the separate or joint use of different resource management mechanisms commonly considered in the literature to mitigate the effects of secondary call interruptions in cognitive radio networks (CRNs) with heterogeneous traffic was evaluated and compared. Novel adaptive spectrum allocation strategies that jointly employ different resource management techniques to take advantage of the flexibility and delay tolerance features of elastic traffic for the considered networks were analyzed and evaluated. To this end, a general teletraffic analysis was developed. For the mathematical modeling, spectrum handoff, channel reservation, elastic-traffic buffering, spectrum adaptation, and non-homogenous channels were considered. Because the proposed analysis reduces the number of state variables employed in [10], this is a major contribution of this work. The performance of the considered system was evaluated in terms of maximum Erlang capacity. Numerical results show that call buffering and spectrum adaptation are the mechanisms that best exploit the elasticity of delay-tolerant traffic in heterogeneous traffic CRNs and, therefore, most significantly improve their performance. Also, it concluded that channel reservation for ongoing secondary calls allows a slow decreasing rates of Erlang capacity as the utilization factor of primary channels increases. Finally, our numerical results show that the Erlang capacity of delay-sensitive secondary users is very low even for reduced primary traffic loads (such that the blocking probability is lower than 1%). As previously commented, the most effective mechanism for guarantee QoS for real-time and interactive applications in CRNs is to reserve spectrum band for exclusive use of

Table 4 Maximum percentage on Erlang capacity gain given for every single considered resource management mechanism when $M = 6$, $N = 3$, $f_v = 0.4767$, $b_{\max} = 3$ and, a maximum value of $\rho = 0.23$

Resource management mechanism	Compared strategies	Maximum Erlang capacity gain [Erlangs]	Maximum percentage on Erlang capacity gain [%]
Buffer for new data calls	E3 vs E4	5.15 (y $P_{\text{ft\_d}}^{(S)} = 0$, at the expense of a longer transmission delay)	72.36
Spectrum adaptation	E2 vs E1	1.65 (and lower $P_{\text{ft\_d}}^{(S)}$ and transmission delay)	43.1
Double preemptive and channel reservation	E5 vs E4	0.16 (the same $P_{\text{ft\_d}}^{(S)}$ and transmission delay)	1.13
Selective interruption	E6 vs E1 {E7 vs E2}	– 0.03[a] {– 0.11[a]} (higher $P_{\text{ft\_d}}^{(S)}$ and similar transmission delay)	-0.43[a] {–1.37[a]}
Buffer for interrupted data calls	E5 vs E7	– 0.47[a] ($P_{\text{ft\_d}}^{(S)} = 0$ for E5 strategy. The transmission delay is the same for both, E5 and E7, strategies)	–7.23[a]

[a] As explained in the text, there is a loss in capacity by queueing the interrupted data calls. In this table entry, the maximum loss obtained for the range of values considered for ρ is shown

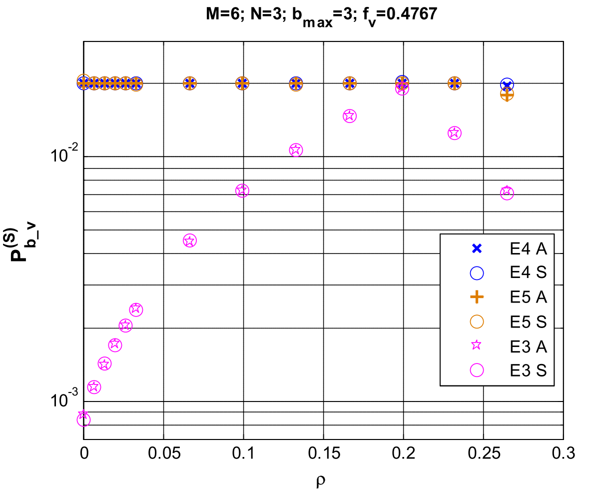

Fig. 9 New call blocking probability of secondary voice users for the corresponding optimal configuration of each strategy as function of the primary channel utilization factor

the secondary network (CCRNs) [3, 5–8]. The use of spectrum partitioning is not considered in this work; however, the extension of the resource management mechanisms considered in the paper for CCRNs is straightforward and subject of future research work.

7 Endnotes

[1] In the context of cellular systems, spectrum adaptation is equivalent to the Flexible Resource Allocation (FRA) concept, where resource compensation and degradation mechanisms are employed [39, 41, 42]. In the context of the LTE-Advanced standard, spectrum adaptation is analogous to the concept of (intra-band) spectrum aggregation [24].

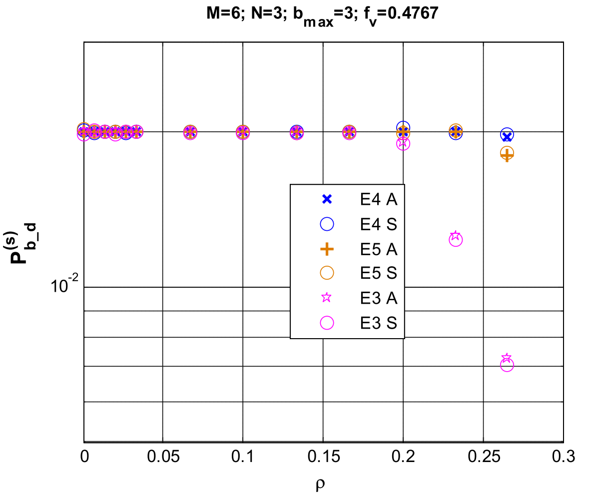

Fig. 10 New call blocking probability of secondary data users for the corresponding optimal configuration of each strategy as function of the primary channel utilization factor

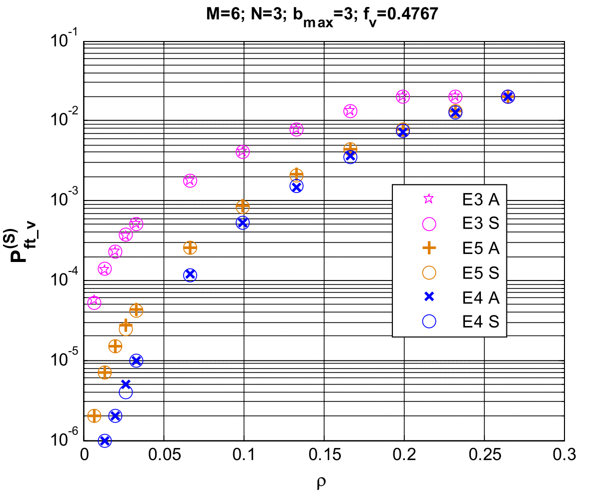

Fig. 11 Forced termination blocking probability of secondary voice users for the corresponding optimal configuration of each strategy as function of the primary channel utilization factor

[2] It is important to comment that initially, we did include channel reservation in this reference strategy. However, we noticed that the simultaneous use of channel reservation and limiting the number of secondary data users to the system has a redundant effect.

[3] Note that the value of Q_{thr} ranges from 0 up to $\lfloor MN/b_{\min}\rfloor$. Then, the assumption that the buffer size is long enough to guarantee that any interrupted secondary data call can be accommodated into the buffer is practical in reality.

Acknowledgements
None.

Authors' contributions
SLCL developed and proposed the adaptive spectrum allocation strategies to mitigate the effects of secondary call interruptions in cognitive radio networks (CRNs). She also participated in the mathematical analysis derived in this work and the numerical results obtained to evaluate the performance of the system. FACP developed the mathematical model and Markov chains used to evaluate the proposed mechanism. Also, he identified the most relevant resource management mechanisms to improve the performance of the considered networks and proposed, analyzed, and modeled the optimization of procedures to maximize the achievable Erlang capacity. GHV was involved in the development of the mathematical model as well as identifying, analyzing, and studying the main resource management mechanisms commonly considered in the literature (spectrum aggregation, spectrum adaptation, call buffering, channel reservation, selective interruption, and preemptive prioritization) to mitigate the effects of secondary call interruptions in cognitive radio networks (CRNs). MERA was involved in the mathematical analysis as well as the evaluation of the studied strategies in heterogeneous traffic CRNs. All authors read and approved the final manuscript.

Funding
PRODEP Contract No.916040 provides us resources to acquire the computing equipment for the numerical evaluation of the studied systems.

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

Publisher's Note

Author details

[1]Electronics Department, UAM-A, Mexico City, Mexico. [2]Electrical Engineering Department, CINVESTAV-IPN, Mexico City, Mexico. [3]Communication Networks Laboratory, CIC-Instituto Politécnico Nacional, Mexico City, Mexico.

Received: 9 April 2017 Accepted: 2 December 2017

References

1. YC Liang, KC Chen, GY Li, P Mähönen, Cognitive radio networking and communications: an overview. IEEE Trans. Veh. Technol. **60**(7), 3386–3407 (2011)
2. M. M. Buddhikot, Cognitive radio, DSA and Self-X: toward next transformation in cellular networks, in Proc. IEEE DySPAN 2010, Singapore, 6–9 Apr. 2010.
3. X. Mao, H. Ji, V. C. M. Leung, M. Li, Performance enhancement for unlicensed users in coordinated cognitive radio networks via channel reservation, in Proc. IEEE GLOBECOM'2010, Miami, FL., 6–10 Dec. 2010.
4. J. Sachs, I. Maric, and A. Goldsmith, Cognitive cellular systems within the TV spectrum, Proc. IEEE Symp. New Frontiers in Dynamic Spectrum (DySPAN'2010), Singapore, 6–9 Apr. 2010.
5. G. Liu. X. Zhu, and, L. Hanzo, Dynamic spectrum sharing models for cognitive radio aided ad hoc networks and their performance analysis, Proc. IEEE GLOBECOM'2011, Houston, Texas, USA, 5–9 Dec. 2011.
6. SK Jayaweera, G Vazquez-Vilar, C Mosquera, Dynamic spectrum leasing: a new paradigm for spectrum sharing in cognitive radio networks. IEEE Trans. on Veh. Technol. **59**(5), 2328–2339 (2010)
7. M. A. Ramirez-Reyna, F. A. Cruz-Perez, M. E. Rivero-Angeles, and G. Hernandez-Valdez, Dynamic Spectrum Leasing Strategies for Coordinated Cognitive Radio Networks with Delay-Tolerant Traffic, IEEE 25th Annual IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC 2014), Washington, DC USA, 2–5 Sept. 2014.
8. M. A. Ramirez-Reyna, F. A. Cruz-Perez, M. E. Rivero-Angeles, and G. Hernandez-Valdez, Performance Analysis of Dynamic Spectrum Leasing Strategies in Overlay Cognitive Radio Networks, Proc. 80th IEEE Vehicular Technology Conference (VTC2014-Fall), Vancouver, Canada, 14–17 Sept. 2014.
9. SL Castellanos-López, FA Cruz-Peréz, ME Rivero-Ángeles, G Hernández-Valdez, Performance analysis of coordinated cognitive radio networks under fixed-rate traffic with hard delay constraints. J. Commun. Netw. Spec Issue Cogn Networking **16**(2), 130–139 (2014)
10. L Jiao, FY Li, V Pla, Modeling and performance analysis of channel assembling in multichannel cognitive radio networks with spectrum adaptation. IEEE Trans. on Veh. Technol. **61**(6), 2686–2697 (2012)
11. T Jiang, H Wang, S Leng, Channel allocation and reallocation for cognitive radio networks. Wirel. Commun., Mob. Comput. **13**(12), 1073–1081 (2013)
12. J. Martínez-Bauset, A. Popescu, V. Pla, and A. Popescu, Cognitive radio networks with elastic traffic, in Proc. 8th Euro-NF Conference on Next Generation Internet (NGI), Karlskrona, Sweden, 25–27 June 2012.
13. L. Jiao, F. Y. Li, V. Pla, Greedy versus Dynamic Channel Aggregation Strategy in CRNs: Markov Models and Performance Evaluation. In: Casares-Giner V., Manzoni P., Pont A. (eds) NETWORKING 2011 Workshops. NETWORKING 2011. Lecture Notes in Computer Science, vol 6827. Springer, Berlin, Heidelberg.
14. L. Yu, T. Jiang, P. Guo, Y. Cao, D. Qu, and P. Gao, Improving achievable traffic load of secondary users under GoS constraints in cognitive wireless networks, in Proc. IEEE Globecom'2011, Houston, Texas, USA, 5–9 Dec. 2011.
15. J Mar, HC Nien, J-C Cheng, Intelligent data control in cognitive mobile heterogeneous networks. IEICE Trans. Commun **E95-B**(4), 1161–1169 (2012)
16. Y Konishi, H Masuyama, S Kasahara, Y Takahashi, Performance analysis of dynamic spectrum handoff scheme with variable bandwidth demand of secondary users for cognitive radio networks. J Wirel. Netw **19**(5), 607–617 (2013)
17. J Lee, J So, Channel aggregation schemes for cognitive radio networks. IEICE Trans. Commun. **E95-B**(5), 1802–1809 (2012)
18. L. Li, S. Zhang, K. Wang, and W. Zhou, Combined channel aggregation and fragmentation strategy in cognitive radio networks, CoRR abs/1203. 4913: (2012).
19. T. M. N. Ngatched, S. Dong, A. S. Alfa, and J. Cai, Performance analysis of cognitive radio networks with channel assembling and imperfect sensing, in Proc. IEEE ICC'2012, Ottawa, Canada, 10–15 June 2012
20. S. M. Kannappa and M. Saquid, Performance analysis of a cognitive network with dynamic spectrum assignment to secondary users, in Proc. IEEE ICC'2010, Cape Town, South Africa, 23–27 May 2010.
21. L. Jiao, F. Y. Li, and V. Pla, Dynamic channel aggregation strategies in cognitive radio networks with spectrum adaptation, in Proc. IEEE GLOBECOM'2011, Houston, Texas, USA, 5–9 Dec. 2011.
22. L. Jiao, V. Pla, and F. Y. Li, Analysis on channel bonding/aggregation for multi-channel cognitive radio networks, in Proc. European Wireless Conference (EW'2010), Tuscany, Italy, 12–15 Apr. 2010.
23. J. Lee and J. So, Analysis of cognitive radio networks with channel aggregation, in Proc. IEEE WCNC'2010, Sydney, Australia, 18-210April 2010.
24. H Lee, S Vahid, K Moessner, A survey of radio resource management for spectrum aggregation in LTE-advanced. IEEE Commun. Surv. Tutorials **16**(2), 745–760 (2014)
25. S. L. Castellanos-Lopez, F. A. Cruz-Perez, and G. Hernandez-Valdez, Channel reservation in cognitive radio networks with the RESTART retransmission strategy, in Proc. CROWNCOM 2012, Stockholm, Sweden, 18–20 June 2012.
26. X Zhu, L Shen, TSP Yum, Analysis of cognitive radio spectrum access with optimal channel reservation. IEEE Commun. Lett **11**(4), 304–306 (2007)
27. J. Lai, R. P. Liu, E. Dutkiewicz, and R. Vesilo, Optimal channel reservation in cooperative cognitive radio networks, in Proc. IEEE 73rd VTC'2011-Spring, Budapest, Hungary, 15–18 May 2011.
28. H Ahmed, SA AlQahtani, Performance evaluation of joint admission and eviction controls of secondary users in cognitive radio networks. Arab. J. Sci. Eng. **40**(12), 3469–3481 (2015)
29. Y. Zhang, Dynamic spectrum access in cognitive radio wireless networks, in Proc. IEEE ICC'2008, Beijing, China, 19–23 May 2008.
30. J Zhou, CC Beard, A controlled preemption scheme for emergency applications in cellular networks. IEEE Trans. on Veh. Technol. **58**(7), 3753–3764 (2009)
31. JL Vázquez-Avila, FA Cruz-Pérez, L Ortigoza-Guerrero, Performance analysis of fractional guard channel policies in mobile cellular networks. IEEE Trans. on Wireless Commun. **5**(2), 301–305 (2006)
32. H Wang, NB Mandayam, Opportunistic file transfer over a fading channel under energy and delay constraints. IEEE Trans. on Commun. **53**(4), 632–644 (2005)
33. IAM Balapuwaduge, L Jiao, V Pla, FY Li, Channel assembling with priority-based queues in cognitive radio networks: strategies and performance evaluation. IEEE Tran. Wireless Commun **13**(2), 630–645 (2014)
34. S.M. Kannappa, T. Ali, and M. Saquib., Analysis of a buffered cognitive wireless network with dynamic spectrum assignment, in Proc. IEEE International Conference on Communications (ICC'2014), Sydney, Australia, 10–14 June 2014.
35. SL Castellanos-López, FA Cruz-Peréz, ME Rivero-Ángeles, G Hernández-Valdez, Joint connection level and packet level analysis of cognitive radio networks with VoIP traffic. IEEE J. Sel. Areas Commun. **32**(3), 601–614 (2014)
36. M.A. Ramírez-Reyna, F.A. Cruz-Pérez, S.L. Castellanos-López, G. Hernández-Valdez, and M.E. Rivero-Angeles, Analysis of spectrum adaptation and spectrum leasing in heterogeneous traffic cognitive radio networks, 12th International IEEE Conference on Wireless and Mobile Computing, Networking and Communications (WiMob 2016), New York, NY, 17–19 October, 2016.
37. RP Murillo-Perez, CB Rodriguez-Estrello, FA Cruz-Perez, Call admission control with fractional buffer size. IEICE Trans. on Commun. **95**(9), 2972–2975 (2012)
38. S. L. Castellanos-Lopez, F. A. Cruz-Perez, and G. Hernandez-Valdez, Performance of Cognitive Radio Networks under Resume and Restart Retransmission Strategies, 7th IEEE International Conference on Wireless and Mobile Computing Networking, and Communications (WiMob 2011), Shanghai, China, 10–12 October 2011.
39. FA Cruz-Pérez, L Ortigoza-Guerrero, Flexible resource allocation strategies for class-based QoS provisioning in mobile networks. IEEE Trans. Veh. Technol. **53**(3), 805–819 (2004)
40. A.L.E. Corral-Ruiz, F.A. Cruz-Perez, S.L. Castellanos-Lopez, G. Hernandez-Valdez, and M.E. Rivero-Angeles, Modeling and performance analysis for mobile cognitive radio cellular networks, J Wireless Com Network, 2017: 159. https://doi.org/10.1186/s13638-017-0940-1.
41. L Ortigoza-Guerrero, FA Cruz-Pérez, H Heredia-Ureta, Call level performance analysis for multi-services wireless cellular networks with adaptive resource allocation strategies. IEEE Trans. Veh. Technol. **54**(4), 1455–1472 (2005)
42. S Tang, W Li, An adaptive bandwidth allocation scheme with preemptive priority for integrated voice/data mobile networks. IEEE Trans. Wireless Commun. **5**(10), 2874–2886 (2006)