A Flexible-Bandwidth Model with Channel Reservation and Channel Aggregation for Three-layered Cognitive Radio Networks

Marcos R. M. Falcão[1], Andson M. Balieiro[2] and Kelvin L. Dias[1]
[1]Universidade Federal de Pernambuco, [2]Universidade de Pernambuco

***Abstract*— The Fifth Generation (5G) of wireless communication systems is expected to meet a large demand for mobile traffic and provide higher data rates, supporting bandwidth-hungry applications. In this respect, Cognitive Radio and Channel Aggregation (CA) are envisioned as key 5G enablers providing additional spectrum resources through the Dynamic Spectrum Access, and higher data rates through multiple contiguous or non-contiguous spectrum aggregation. Moreover, since 5G networks should comprise heterogeneous applications that may have different Quality of Service (QoS), Quality of Experience (QoE) and security requirements, multiple service class support becomes a must, and thus multiple priorities have been assigned for different flow types in current wireless standards. Previous works have studied Cognitive Radio Networks (CRN) as homogeneous two-priority queueing systems, composed of primary (PUs) and secondary (SUs) users, however, those are usually not capable of analyzing SUs with different QoS requirements. In addition, most authors are concerned about proving the efficiency of QoS provisioning approaches such as channel reservation or channel aggregation, frequently using separate models in unloaded scenarios. This paper proposes and analyzes an M/M/N/N three-layered system in which the unlicensed traffic is detached in two priority classes (i.e., high and low), encompassing all possible bandwidth arrangements, a multi-level reservation feature and multiple aggregation strategies. Previous works on CA have shown that, regardless the network state, this technique should always boost the overall performance, differently from the reservation process that presents high inefficiency in overloaded networks. For this reason, CA was enabled to mitigate the reservation's drawbacks while scaling the benefits of both techniques, in a single model.**

***Index Terms*— Cognitive Radio Networks, Dynamic Spectrum Access, Continuous-Time Markov Chains.**

I. Introduction

The mobile traffic has experienced exponential growth due to the bandwidth requirement shift that supports demanding applications such as video streaming and massive machine-to-machine (M2M) communication [1]. The fifth generation (5G) of mobile communications is expected to meet a broader range of applications, which are beyond the capability of previous technologies, requiring spectrum resources to be efficiently managed [2]. Cognitive Radio (CR) has been envisioned as a 5G enabler [3] that helps addressing the strict spectrum requirement, supporting the opportunistic use of the underutilized licensed spectrum through Dynamic Spectrum Access (DSA) and allowing low-cost expansion for wireless systems [4]. Besides, in order to achieve 5G high date rates (10 to 100 times the current 4G speeds), channel aggregation (CA) may scale CR benefits by enabling multiple contiguous or non-contiguous spectrum aggregation that may be of two kinds: macro and micro scale. The first utilizes fixed sized spectrum fragments (e.g., resources blocks - RBs in LTE systems), while the latter uses varying spectrum fragments to be considered in the process [5]. Therefore, higher throughput and flexibility may be achieved through the opportunistic use of underutilized licensed bands (e.g., TV bands) in conjunction to the regular licensed band (e.g., LTE band).

In Cognitive Radio Networks (CRNs), Secondary Users (SUs) opportunistically access the spectrum that is temporarily unused by the licensed users also known as Primary Users (PUs). Since 5G networks should comprise heterogeneous applications (e.g., ultra-high definition video streaming, Web browsing, online banking, and tactile Internet) that have different Quality of Service (QoS), Quality of Experience (QoE) and security requirements, multiple service class support becomes a must. Current standards such as the IEEE 802.11p for Vehicular Ad Hoc Networks (VANETs) were prepared for the traffic collision possibility, where vital flow types such as those from safety applications may be hindered by the infotainment traffic. So as to protect relevant data, this standard proposes multiple priorities for different flow types and by following this idea, previous authors have built queue-based analytical models to analyze service classes in CRNs, considering either two [6-13] or three priorities [14,15,16,17].

Most authors have proposed inflexible models that are either restricted to cases where the PU bandwidth is larger than the SU´s [11, 14, 15] or the opposite [9, 13, 17]. Hence, scenarios where: (1) a smart grid network (e.g. an advanced metering infrastructure) is employed as a secondary network and ATSC digital TV signals with a bandwidth of 6 MHz characterizes the primary network; and (2) a wireless multimedia streaming network for connected home (e.g. by using 802.11af) [18] is the secondary network and NTSC analog TV signals with 100 kHz bandwidth form the primary system [19] can be rarely represented by a single model. The work developed in [10] is an exception, but is limited for a two-user priority system (homogeneous secondary system).

This work was supported by the Science and Technology Foundation of Pernambuco (FACEPE)/Brazil under Grant IBPG -0571-1.03/10.

This paper aims to accommodate three user categories in a single CRN (heterogeneous secondary system) with the following access priority order: Primary (PUs), first class SU (SU_1) and second class SU (SU_2), allowing all users' bandwidth requirements to be analyzed simultaneously with channel reservation and channel aggregation.

Channel reservation is an admission control mechanism used to guarantee the QoS [6,11,12,15], by avoiding network overuse. However, differently from previous works that have adopted single-level channel reservation and thus defined the same number of reserved channels for all SU types [15], the current model admits multi-level channel reservation. This allows limiting a specific channel number for each SU, i.e., the channel reservation mechanism may be fitted to each SU layer. Finally, although the literature has thoroughly explored such technique, its drawbacks are often disguised by poor scenario exploration (e.g., low network occupation). Hence, knowing that channel aggregation offers performance boosting in any circumstance [17], this feature is also contemplated in the system for the most vulnerable user layer (SU_2), according to the adopted priority order.

This paper evaluates the secondary system in terms of blocking and dropping probabilities that are derived from a multi-dimensional Markov chain model with three-state variables, one for each user layer. It was noted that previous works poorly explored their scenarios, first due to an intrinsic limitation that focused either on larger bandwidth PUs or SUs and secondly because a low PU density was usually applied, favoring the secondary communication. On the contrary, this work employs multiple PU density scales for every experiment, showing useful insights both in congested and uncongested networks. The results show that channel reservation not always provides reasonable performance tradeoffs for the secondary network, as most authors have suggested. In fact, this technique highly depends on the CRN load to be useful, i.e., it slightly increases the blocking probability but substantially decreases the forced termination rate in some cases. For this reason, dynamic channel aggregation was applied together with channel reservation to mitigate the blocking event increase, while keeping the low forced termination rate.

In brief, the main contributions of this paper are outlined as follows:

- A more complete model for CRN is proposed and validated. Previous works just explore a subset of features (channel aggregation, three-layer priority and channel reservation) in an isolate manner or have constraints in terms of users' bandwidth requirements or do not consider heterogeneous secondary system; in counterpart, our CRN model is unconstrained regarding the bandwidth arrangement between user types and encompasses all features, allowing the analysis of their joint effects on the CRN and more complex scenarios may be analyzed.
- A multi-level channel reservation is proposed and tested in a non-typical scenario, enabling the reservation mechanism to be fitted to each SU layer.
- The joint use of dynamic channel aggregation and channel reservation is proposed and tested to mitigate the negative reservation effects.
- A comprehensive model analysis on the CRN performance, by applying multiple PU density scales for every experiment, showing useful insights both in crowded and uncongested networks.

In a realistic scenario, the model could help the service provider to dimension the CRN by answering the following issues: Justify the possibility to employ secondary services, how to structure these at a reasonable QoS and to establish the limits of channel aggregation, fragmentation or reservation for performance boosting. The remainder of this paper is organized as follows: Section II presents the related works, section III explores the analytical model while the performance metrics and the numerical results are detailed in sections IV and V. Finally, section VI provides a conclusion.

II. RELATED WORKS

The literature has thoroughly explored queuing-based models to evaluate the QoS of Cognitive Radio Networks (CRNs) [6-17]. However, most works are not flexible to represent all bandwidth arrangements between user classes, being generally limited to one of the following cases: a single channel is applied for both PUs and SUs [12], the PU bandwidth is necessarily larger than the SU´s [6,11,14,15] or the opposite [7,8,9,16,17]. For instance, the authors of [6,11] considered bandwidth requirements where each PU uses a single channel whereas the SUs are constrained to sub-bands (smaller than one channel unit), and besides, several mistakes in [6] were pointed by [20,21]. On the counterpart, [7,8,9,13,17] support multiple bandwidth requirements with the following condition: the SU's bandwidth must be larger than the PU's. The work proposed in [10] is a rare exception; it presents a more flexible model allowing any arrangement between the user bandwidths besides describing a more complete analysis by evaluating two scenarios: real-time and elastic traffic. On the other hand, since the model only supports a single SU type, these are analyzed separately, that is, it does not cover a heterogeneous secondary system in a single model. The works [14,15,16,17] differ in this sense as three user layers are adopted, with two SUs types that may have different bandwidth requirements, characterizing a heterogeneous secondary system.

In the CRN context, secondary transmissions can collide with primary transmissions since the PUs have the priority over the SUs, i.e., in any moment, PUs can access a channel occupied by a SU, causing a forced interruption. Hence, this model explores a prioritization mechanism known as channel reservation that enables QoS provisioning in CRNs [6,11,12,15]. By doing so, new secondary arrivals are blocked even if there are enough available channels, lessening secondary terminations due to a future PU arrival. For instance, the authors of [15] provide channel reservation, but in a limited manner as a single level reservation (not distinguishing SU_1 and SU_2) is adopted in both layers. In other words, if the SU_1's access is limited, it necessarily impacts the SU_2 as well. In particular, we investigate the effects of multi-

Tab. 1. *Related work comparison*

Paper	Priority Levels	Secondary Traffic	Bandwidth Constraints	Aggregation Technique(s)	Channel Reservation
[6]	Two	Homogeneous	PU bandwidth ≥ SU bandwidth	Static CA	Single-Level CA
[7]	Two	Homogeneous	SU bandwidth ≥ PU bandwidth	Static CA	N/A
[8]	Two	Heterogeneous	Minimum SU bandwidth ≥ PU bandwidth	Dynamic CA+CF	N/A
[9]	Two	Homogeneous	Minimum SU bandwidth ≥ PU bandwidth	Dynamic CA+CF	N/A
[10]	Two	Homogeneous	No Constraint	Dynamic CA+CF	N/A
[11]	Two	Homogeneous	PU bandwidth ≥ Minimum SU bandwidth	Static CA	Single -Level CA
[12]	Two	Homogeneous	PU bandwidth = SU bandwidth	Static CA	Single -Level CA
[13]	Two	Homogeneous	Minimum SU bandwidth ≥ PU bandwidth	Dynamic CA+CF	N/A
[14]	Three	Heterogeneous	PU bandwidth ≥ SU bandwidth	Static CA	N/A
[15]	Three	Heterogeneous	PU bandwidth ≥ SU bandwidth	Static CA	Single -Level CA
[16]	Three	Heterogeneous	Minimum SU bandwidth ≥ PU bandwidth	Dynamic CA	N/A
[17]	Three	Heterogeneous	Minimum SU bandwidth ≥ PU bandwidth	Dynamic CA+CF	N/A
This Paper	Three	Heterogeneous	No Constraint	Dynamic CA+CF	Multi-Level CA

level channel reservation by allowing each secondary layer to have an individual number of reserved channels, differently from [11,12,15]. Furthermore, [12] proposes a scheme that exclusively reserves a number of channels to the SUs, not allowing the PUs to access these, although they have the highest access priority. On the other hand, the current work ensures the PU's highest access priority by allowing it to occupy any channel, even those reserved to the SUs, which reflects the opportunistic spectrum access (OSA) concept.

To diminish both blocking and forced termination events by enabling an SU to access multiple channels simultaneously, fixed channel aggregation (FCA) has been proposed [6,7,14,15]. It can be accompanied by spectrum handover that enables an ongoing SU service to hop onto another non-occupied channel and in some cases to change its bandwidth according to the network usage, such as in [17]. In the current paper, such feature is named dynamic channel aggregation, which can be further divided into integer (macro scale) [16] or fractionary (micro scale) [8,9,10,13,17]. Both can be applied to the lowest priority layer (SU_2) whereas the other two layers (PU and SU_1) will be allowed to aggregate a fixed number of channels. Therefore, this model encompasses a heterogeneous secondary layer, differently from [6,7,9,10,12,13] that supports only a homogeneous secondary network.

The key difference from this paper to our previous work [17] is that the latter compared the performance of three aggregation techniques in a limited model, where the bandwidths of both PUs and SU_1s are restricted to one channel unit. Similarly, [14,15,16] also could not provide an unique set of transitions in order to evaluate multiple bandwidth arrangements between the three network classes. On the other hand, the current paper eases this constraint by allowing both PUs and SU_1s to assemble a fixed number of channels that can be greater than, equal or smaller than the bandwidth set for the SU_2. Furthermore, [17] does not explores channel reservation (neither the single approach nor the multi-level) while the present work focuses on the use of the aggregation techniques as a solution for suppressing the drawbacks that may occur when utilizing the first approach. In brief, this paper explores a range of scenarios where some important features were not put to proof in a single model, besides experimenting under a more diverse scenario where multiple PU loads were applied.

The related works were assessed in terms of 'Priority Levels', 'Secondary Traffic', 'Bandwidth Constraints', 'Aggregation Technique(s)' and 'Channel Reservation'. These features are summarized in Tab. 1

III. System Model

A continuous time Markov chain (CTMC) was used to model the CRN, assuming three user types (PU, SU_1 and SU_2) sharing N channels. The user arrivals are Poisson processes with rates λ_P, λ_{SU1} and λ_{SU2} and the service times are exponentially distributed with rates μ_P, μ_{SU1} and μ_{SU2} for the PU, SU_1 and SU_2, respectively. This model allows the service rate to vary depending on the bandwidth requirement, i.e., if M channels are assembled, then the service rate is tuned to $M * \mu$, which diminishes the service duration.

The CRN addressed by the current model uses a centralized overlay approach and a common control channel to map the resource´s status along operation. Like other studies, we do not take into account the time overhead imposed by the spectrum sensing and system's collision delay as they are small compared to the transmission duration. For the latter, this means that no time is spent if a higher priority user drops a lower priority one, taking control of its resources.

With regard to the priority mechanism, when a new SU_1 arrives, it randomly selects a set of free channels (not used by PU or SU_1). If the chosen set is occupied by an SU_2, this user will move to another idle channel as long as the CRN has enough. On the contrary, the SU_2 will be discontinued. Similarly, the PU follows the same procedure towards both SU_1 and SU_2. Once a PU enters the network, it only leaves when its service time is completed as opposed to the SUs that may be forced out before service completion.

In this paper, B_{PU}, B_{SU1} and B_{SU2}^{min} refer to the bandwidths for each user layer PU, SU_1 and SU_2 (minimum required bandwidth), while R_1 and R_2 indicate the restricted amount of spectrum resources (in channel units) from the SU_1 and SU_2, respectively. In other words, during operation a SU_1 can statically access N - R_1 channels, with $N \geq R_1$ and similarly, a SU_2 can access N - R_2, with $N \geq R_2$ and ($R_1 + R_2 \leq$ N), whereas the PU can access all channels. Then, each state can be represented as a tuple (i, j, k), where i, j and k are the number of active PUs, SU_1s and SU_2s, respectively. Finally, the

feasible state space is $\Omega = \{(i,j,k) | 0 \le i \le, \left\lfloor \frac{N}{B_{PU}} \right\rfloor, 0 \le j \le \left\lfloor \frac{N-R_1}{B_{SU1}} \right\rfloor, 0 \le k \le \left\lfloor \frac{N-R_2}{B_{SU2}^{min}} \right\rfloor$, provided that $(i * B_{PU}) + (j * B_{SU1}) + (k * B_{SU2}^{min}) \le N\}$.

Since prioritization may greatly hinder the SU_2´s communication, the use of channel aggregation for leveraging the SU_2´s performance is permitted in this model. Other than FCA, two approaches named Dynamic Channel Aggregation (DCA) and Dynamic Channel Aggregation and Fragmentation (DCAF) can be applied in the lowest layer (SU_2s) together with the *equal sharing algorithm* (ESA). ESA distributes evenly the available spectrum among ongoing SU_2s, given that the bandwidth requirement of each user is not violated. The ESA will be executed only for adjusting the SU_2´s bandwidth, but for every arrival or service completion event as follows: (1) When a SU_2 arrives, it will try to occupy the maximum allowed bandwidth B_{SU2}^{max}, otherwise, the available channels will be equally shared by the ongoing SU_2s and by the newcomer if every user gets at least B_{SU2}^{min} resources. If not, then the new SU_2 will be blocked. (2) When a new PU or SU_1 arrives, they pick B_{PU} or B_{SU1} channels as long as they are idle or occupied by SU_2s. In the latter, these SU_2s will either need to move to another idle channel or share the available spectrum by ESA, ensuring that the SU_2's bandwidth is kept between $[B_{SU2}^{min}, B_{SU2}^{max}]$, otherwise, the target SU_2(s) will be forcibly terminated. (3) Once a user completes transmission and leaves the network, the residual SU_2s will equally share the vacant spectrum by ESA, given that B_{SU2}^{max} is respected. Differently from the SU_2's elastic property, both PU and SU_1 keep their bandwidth amounts either until service completion (PUs or SU_1s) or forced termination (solely for SU_1s), regardless the availability.

For the following examples in Figs. 1, 2 and 3, consider the PU and SU_1's bandwidths one channel unit, while the SU_2's bandwidth varies according to the aggregation strategy. In DCAF (Fig. 1), the SU_2's bandwidth can dynamically be adjusted depending on the load, as channel aggregation (CA) and channel fragmentation (CF) are performed adaptively and the number of assembled resources may be a real positive number between B_{SU2}^{min} and B_{SU2}^{max} (micro scale aggregation). Furthermore, by using the DCA (Fig. 2), the SU_2's bandwidth can also be dynamically adjusted according to availability, but only by adaptive CA; consequently, the number of assembled resources may be an integer positive number between B_{SU2}^{min} and B_{SU2}^{max} (macro scale aggregation). In Fig. 2, the SU_1 performs channel handoff (from C_2 to C_1) due to the PU arrival at T4, i.e., this user vacates the channel C_2 (requested by the arriving PU) and uses the channel C_1 to resume its communication, which forces the SU_2's communication to be terminated. The last approach known as fixed channel aggregation (FCA) allows a user to use multiple channels simultaneously (Fig. 3), however, bandwidth adaptation is not permitted during operation, hence $B_{SU2}^{min} = B_{SU2}^{max}$.

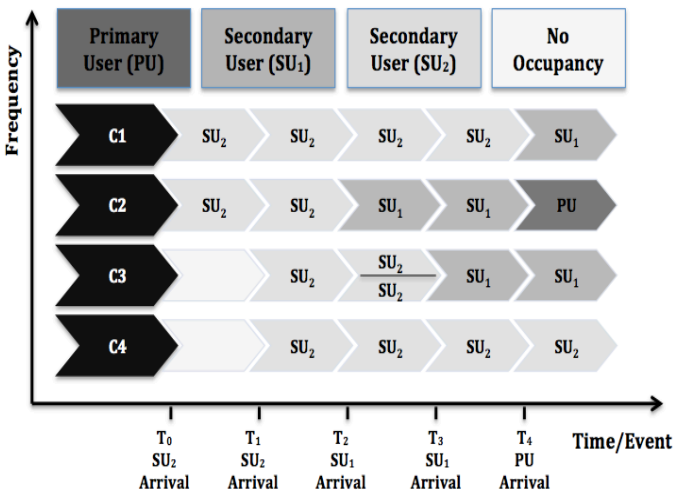


Fig. 1. *DCAF example in a four channel CRN where* $B_{SU2}^{min} = 1$ *and* $B_{SU2}^{max} = 2$

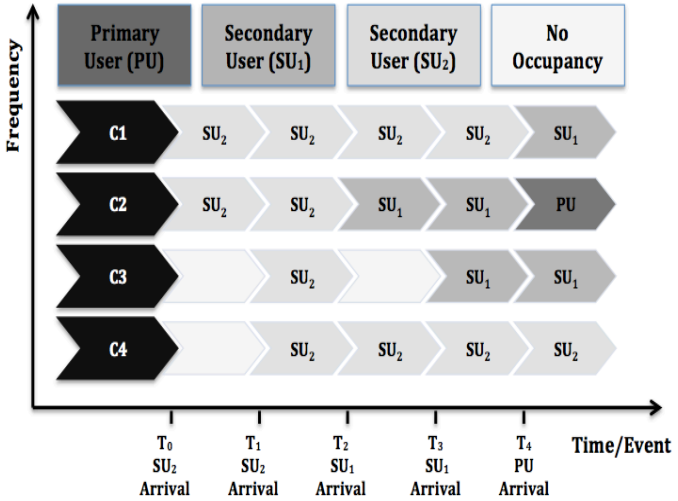


Fig. 2. *DCA example in a four channel CRN where* $B_{SU2}^{min} = 1$ *and* $B_{SU2}^{max} = 2$

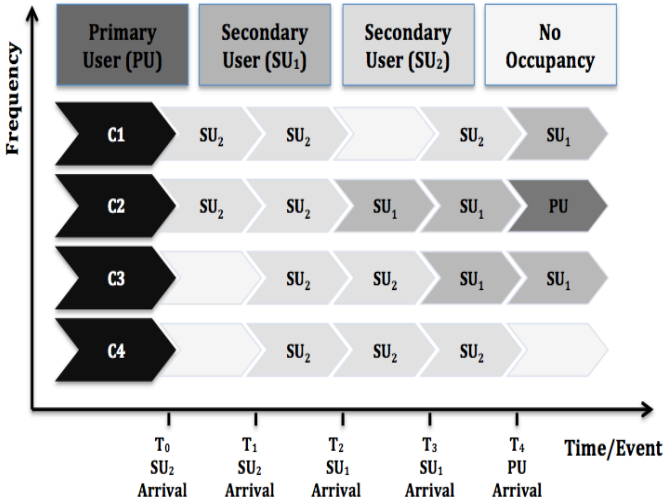


Fig. 3. *FCA example in a four channel CRN where* $B_{SU2}^{min} = 2$ *and* $B_{SU2}^{max} = 2$

The SU_2's bandwidth may vary according to the network occupancy and adopted aggregation strategy. In DCAF, for instance, the bandwidth and service rate of each SU_2 in Ω is expressed by (1) and (2), respectively. For DCA, the bandwidth and service rate will differ from DCAF's expression due to the integer channel aggregation property, represented by (3) and (4). Lastly, in order to enable a fixed aggregation rule, it is possible to obtain FCA from either DCAF or DCA by setting equal values to B_{SU2}^{min} and B_{SU2}^{max}, which will result in $B_{SU2,FCA}(i,j,k) = B_{SU2}^{min} = B_{SU2}^{max}$ and $\mu_{SU2,FCA} = B_{SU2,FCA}(i,j,k) * \mu_{SU2}$. Note that for Eqs. (1-4), the ESA is ensured as the resulting bandwidth should be applied to all active SU_2s.

(1)

$$\mathrm{B_{SU2,DCAF}(i,j,k)} = \begin{cases} \min\left\{\mathrm{B_{SU2}^{max}}, \max\left\{\mathrm{B_{SU2}^{min}}, \dfrac{\mathrm{N-(i*B_{PU})-(j*B_{SU1})-R_2}}{\mathrm{k}}\right\}\right\}, \\ \text{if } 0 \le (\mathrm{i*B_{PU}}) + (\mathrm{j*B_{SU1}}) \le \left\lfloor \mathrm{N-B_{SU2}^{min}} \right\rfloor, 1 \le \mathrm{k} \le \left\lfloor \dfrac{\mathrm{N}}{\mathrm{B_{SU2}^{min}}} \right\rfloor; \\ 0, \text{otherwise.} \end{cases}$$

$$\mu_{\mathrm{SU2,DCAF}}(\mathrm{i,j,k}) = \mathrm{B_{SU2,DCAF}(i,j,k)} * \mu_{\mathrm{SU2}} \quad (2)$$

(3)

$$\mathrm{B_{SU2,DCA}(i,j,k)} = \begin{cases} \min\left\{\mathrm{B_{SU2}^{max}}, \max\left\{\mathrm{B_{SU2}^{min}}, \left\lfloor \dfrac{\mathrm{N-(i*B_{PU})-(j*B_{SU1})-R_2}}{\mathrm{k}} \right\rfloor\right\}\right\}, \\ \text{if } 0 \le (\mathrm{i*B_{PU}}) + (\mathrm{j*B_{SU1}}) \le \left\lfloor \mathrm{N-B_{SU2}^{min}} \right\rfloor, 1 \le \mathrm{k} \le \left\lfloor \dfrac{\mathrm{N}}{\mathrm{B_{SU2}^{min}}} \right\rfloor; \\ 0, \text{otherwise.} \end{cases}$$

$$\mu_{\mathrm{SU2,DCA}}(\mathrm{i,j,k}) = \mathrm{B_{SU2,DCA}(i,j,k)} * \mu_{\mathrm{SU2}} \quad (4)$$

The expressions in subsections *A* and *B* describe all possible state transitions for the system, which are classified as user requests or service completions. The state transitions $\gamma_{(i,j,k)}^{(i',j',k')}$ occur from one feasible state (i,j,k) to another (i',j',k') and classified in normal and forced termination cases, where the first stands for those arrivals that do not imply in user interruptions while the latter will necessarily cause a user interruption. For the following set of equations, the SU_2's variable bandwidth and service rate should be represented by B_{SU2} and μ_{SU2} respectively, and thus be replaced by the desired approach's equivalents (DCA or DCAF) described in Eqs (1-4) or by placing equal values for B_{SU2}^{min} and B_{SU2}^{max} if FCA is selected.

A. Transitions from State (i,j,k) to Other States

Consider the number of idle resources in state *(i,j,k)* to be $idle = N - (i * B_{PU}) - (j * B_{SU1}) - (k * B_{SU2}^{min})$ and each feasible state transition is associated to a given probability/rate. In this way, the transitions from state *(i,j,k)* to all other states are described as follows and then summarized in Tab. 2.

1) Primary user requests service:

a. If the number of idle resources is greater than or equal to the PU bandwidth, i.e., $B_{PU} \le idle$, the arriving PU will be assigned $\mathrm{B_{PU}}$ channels without any user being forced to terminate.

$$\gamma_{(\mathrm{i,j,k})}^{(\mathrm{i+1,j,k})} = \lambda_{\mathrm{PU}}$$

b. If the number of idle resources is smaller than the PU bandwidth, but the sum of the idle channels and resources occupied by SU_2s is greater than or equal to the PU bandwidth, i.e., $idle < B_{PU} \le idle + (k * B_{SU2})$, then the arriving PU will be assigned $\mathrm{B_{PU}}$ channels and $z = \lceil (B_{PU} - idle)/B_{SU2} \rceil$ SU_2s will be terminated.

$$\gamma_{(\mathrm{i,j,k})}^{(\mathrm{i+1,j,k-z})} = \lambda_{\mathrm{PU}}$$

c. If the sum of the idle channels and the ones occupied by SU_2s is not enough for accommodating an arriving PU, but this number summed up with the amount of resources occupied by SU_1s is, i.e., $idle + (k * B_{SU2}) < B_{PU} \le idle + (k * B_{SU2}) + (j * B_{SU1})$, then $y = \lceil ((i * B_{PU}) + (j * B_{SU1}) - N + B_{PU})/B_{SU1} \rceil$ SU_1s and k SU_2s will be terminated.

$$\gamma_{(\mathrm{i,j,k})}^{(\mathrm{i+1,j-y,0})} = \lambda_{\mathrm{PU}}$$

2) First class secondary user requests service:

a. If the number of idle resources is greater than or equal the sum of the SU_1's bandwidth and their number of reserved channels, i.e. $B_{SU1} + R_1 \le idle$, the arrival SU_1 will be assigned $\mathrm{B_{SU1}}$ channels without any SU_2 being forced to terminate.

$$\gamma_{(\mathrm{i,j,k})}^{(\mathrm{i,j+1,k})} = \lambda_{\mathrm{SU1}}$$

b. If the number of idle resources is smaller than the sum of the SU_1's bandwidth and their number of reserved channels, but the sum of unoccupied channels and channels occupied by SU_2s is greater than or equal to that value, i.e., $idle < B_{SU1} + R_1 \le idle + (k * B_{SU2}^{min})$, then the SU_1 arrival is assigned $\mathrm{B_{SU1}}$ channels. Thus, $z = \lceil (B_{SU1} - idle)/B_{SU2}^{min} \rceil$ SU_2s should be terminated.

$$\gamma_{(\mathrm{i,j,k})}^{(\mathrm{i,j+1,k-z})} = \lambda_{\mathrm{SU1}}$$

3) Second class secondary user requests service:

a. If the number of idle resources is greater than or equal to the sum of the minimum SU_2's bandwidth and their number of reserved channels, i.e., $B_{SU2}^{min} + R_2 \le idle$, the arrival SU_2 will be assigned $B_{SU2}(i,j,k+1)$ channels (please refer to Eqs. (1-4)). The output for $B_{SU2}(i,j,k+1)$ should be a value in between the pre-defined lower and upper bounds.

$$\gamma_{(\mathrm{i,j,k})}^{(\mathrm{i,j,k+1})} = \lambda_{\mathrm{SU2}}$$

4) Primary user completes service:

$$\gamma_{(\mathrm{i,j,k})}^{(\mathrm{i-1,j,k})} = \mathrm{i} * \mu_{\mathrm{PU}}$$

5) First class secondary user completes service:

$$\gamma_{(\mathrm{i,j,k})}^{(\mathrm{i,j-1,k})} = \mathrm{j} * \mu_{\mathrm{SU1}}$$

6) Second class secondary user completes service:

$$\gamma_{(\mathrm{i,j,k})}^{(\mathrm{i,j,k-1})} = \mathrm{k} * \mu_{\mathrm{SU2}}$$

B. Transitions from Other States to State (i,j,k)

The transitions from all other states to *(i,j,k)* are as follows and then summarized in Tab. 2.

1) Primary user requests service:

a. If the number of free resources is greater than or equal to the PU bandwidth, i.e., $B_{PU} \leq idle\ (i-1, j, k)$, the arriving PU will be assigned B_{PU} channels without any other user being forced to terminate. In such case, there will be only one possible state from which the system changes to state *(i, j, k)*.

$$\gamma^{(i,j,k)}_{(i-1,j,k)} = \lambda_{PU}$$

b. If the number of free resources is smaller than the PU's bandwidth, i.e., $B_{PU} > idle\ (i-1, j, k+z')$, $k > 0$, then $z' = \lceil B_{PU}/B^{min}_{SU2} \rceil$ SU_2s will be removed, if this amount is enough for admitting the newly arrived PU.

$$\gamma^{(i,j,k)}_{(i-1,j,k+z')} = \lambda_{PU}$$

c. If the number of free resources is lower than the PU's bandwidth, i.e., $B_{PU} > idle\ (i-1, j, k+z')$ and $k = 0$, then four situations might occur:

i. $B_{PU} < B^{min}_{SU2}$ *and* $B_{PU} \geq B_{SU1}$. If $k + z' > 0$, then only one SU_2 will be dropped, i.e., $z'=1$. Otherwise, $k + z'= 0$, then there are no SU_2s to be dropped, so $y' = \lceil B_{PU}/B_{SU1} \rceil$ SU_1s will be interrupted.

ii. $B_{PU} < B^{min}_{SU2}$ *and* $B_{PU} < B_{SU1}$. If $k + z' > 0$, then only one SU_2 will be dropped, i.e., $z'=1$ *and* $y'=0$. Otherwise, if $k + z'= 0$, then only one SU_1 will be dropped, i.e., $z'=0$ *and* $y'=1$.

iii. $B_{PU} \geq B^{min}_{SU2}$ *and* $B_{PU} \geq B_{SU1}$. In this case, y' may vary from zero up to B_{PU}/B_{SU1}, whereas z' varies from zero to B_{PU}/B^{min}_{SU2}, as long as, $y' * B_{SU1} + z' * B^{min}_{SU2}$ is equal to $B_{PU} - idle\ (i-1, j+y', k+z')$.

iv. $B_{PU} \geq B^{min}_{SU2}$ *and* $B_{PU} < B_{SU1}$. If $B_{PU} \leq idle\ (i-1, j+y', k+z') + (k+z') * B^{min}_{SU2}$, then z' varies from zero to B_{PU}/B^{min}_{SU2}, as long as, $z' * B^{min}_{SU2}$ is equal to $B_{PU} - idle\ (i-1, j+y', k+z')$ and $y'= 0$.

$$\gamma^{(i,j,k)}_{(i-1,j+y',k+z')} = \lambda_{PU}$$

2) First class secondary user requests a service:

a. If the number of free resources is greater than or equal to the sum of the SU_1's bandwidth, i.e., $B_{SU1} \leq idle\ (i, j-1, k)$, the arriving SU_1 will be assigned B_{SU1} channels without any SU_2 being forced to terminate.

$$\gamma^{(i,j,k)}_{(i,j-1,k)} = \lambda_{SU1}$$

b. If there are no free resources for SU_1 arrivals but, by dropping SU_2s, enough space can be made available, i.e., $idle\ (i, j+y', k+z') + (k+z') * B_{SU2} \geq B_{SU1} > idle\ (i, j-1, k+z')$, then there might be more than one situation from which the system changes to (i, j, k).

i. If $B_{SU1} \leq B^{min}_{SU2}$ and $(k + z') > 0$ then $z' = 1$.

ii. If $B_{SU1} > B^{min}_{SU2}$ then z' assumes up to $\lceil B_{SU1}/B^{min}_{SU2} \rceil$.

$$\gamma^{(i,j,k)}_{(i,j-1,k+z')} = \lambda_{SU1}$$

3) Second class secondary user requests a service:

a. If the number of idle resources is greater than or equal to the sum of the minimum SU_2's bandwidth and their number of reserved channels, i.e., $B^{min}_{SU2} + R_2 \leq idle\ (i, j, k-1)$, the arriving SU_2 will be assigned $B_{SU2}(i, j, k+1)$ channels.

$$\gamma^{(i,j,k)}_{(i,j,k-1)} = \lambda_{SU2}$$

4) Primary user completes service:

$$\gamma^{(i,j,k)}_{(i+1,j,k)} = (i+1) * \mu_{PU}$$

5) First class secondary user completes service:

$$\gamma^{(i,j,k)}_{(i,j+1,k)} = (j+1) * \mu_{SU1}$$

6) Second class secondary user completes service:

$$\gamma^{(i,j,k)}_{(i,j,k+1)} = (k+1) * \mu_{SU2}(i,j,k)$$

A state transition diagram when $N = 2$, $B_{PU} = 1$, $B_{SU1} = 1$, $B^{min}_{SU2} = 1$, $R_1 = 0$ *and* $R_2 = 0$ is depicted in Fig. 4, i.e., no reservation is assumed and so, the variables R_1 and R_2 are set to zero. The diagram depends only on B^{min}_{SU2} because B^{max}_{SU2} is not considered for the feasible state space, so Fig. 4 can represent FCA, DCA or DCAF as long as B^{min}_{SU2} is the same.

Tab. 2. *Transition summary*

Event	Type	Departing from state (i,j,k) to other states		Departing from other states to state (i,j,k)	
		Transition	Value	Transition	Value
PU Arrival	Normal	$\gamma^{(i+1,j,k)}_{(i,j,k)}$	λ_{PU}	$\gamma^{(i,j,k)}_{(i-1,j,k)}$	λ_{PU}
PU Arrival	Dropping	$\gamma^{(i+1,j,k-z)}_{(i,j,k)}$	λ_{PU}	$\gamma^{(i,j,k)}_{(i-1,j,k+z')}$	λ_{PU}
PU Arrival	Dropping	$\gamma^{(i+1,j-y,0)}_{(i,j,k)}$	λ_{PU}	$\gamma^{(i,j,k)}_{(i-1,j+y',k+z')}$	λ_{PU}
SU_1 Arrival	Normal	$\gamma^{(i,j+1,k)}_{(i,j,k)}$	λ_{SU1}	$\gamma^{(i,j,k)}_{(i,j-1,k)}$	λ_{SU1}
SU_1 Arrival	Dropping	$\gamma^{(i,j+1,k-z)}_{(i,j,k)}$	λ_{SU1}	$\gamma^{(i,j,k)}_{(i,j-1,k+z')}$	λ_{SU1}
SU_2 Arrival	Normal	$\gamma^{(i,j,k+1)}_{(i,j,k)}$	λ_{SU2}	$\gamma^{(i,j,k)}_{(i,j,k-1)}$	λ_{SU2}
PU Departure	Normal	$\gamma^{(i-1,j,k)}_{(i,j,k)}$	$i * \mu_{PU}$	$\gamma^{(i,j,k)}_{(i+1,j,k)}$	$(i+1) * \mu_{PU}$
SU_1 Departure	Normal	$\gamma^{(i,j-1,k)}_{(i,j,k)}$	$j * \mu_{SU1}$	$\gamma^{(i,j,k)}_{(i,j+1,k)}$	$(j+1) * \mu_{SU1}$
SU_2 Departure	Normal	$\gamma^{(i,j,k-1)}_{(i,j,k)}$	$k * \mu_{SU2}(i,j,k)$	$\gamma^{(i,j,k)}_{(i,j,k+1)}$	$(k+1) * \mu_{SU2}(i,j,k)$

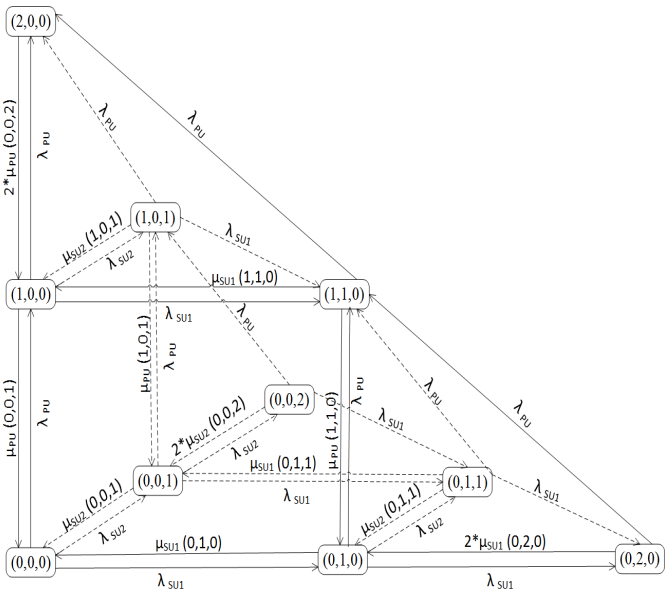


Fig. 4. *State transition diagram for a CRN with N = 2, B_{PU}= B_{SU1} = B_{SU2}^{min}=1, R_1= R_2 = 0*

Channel reservation can be activated in one or more secondary network layers, reducing the number of accessible channels for arriving SUs and providing better QoS to the ongoing SUs. The reserved channels can be used by the SU to resume its communication when a higher priority user arrives and requests its current channel, for instance. Using the same example as in Fig. 4, this feature can be enabled by setting positive integers for the variables R_1 and R_2. For example, when $R_1 = 1$ and $R_2 = 0$, the SU_1s will be able to access $N - R_1$ channels (see Fig. 5). On the other hand, Fig. 6 takes the configuration where $R_1 = 0$ and $R_2 = 1$, thereupon the SU_2 should be able to access only $N - R_2$ channels. Moreover, the example in Fig. 7 shows channel reservation being used in both secondary levels simultaneously, i.e., $R_1 = 1$ and $R_2 = 1$. It allows adjusting the accepted users in the network, being an important mechanism for performance regulation.

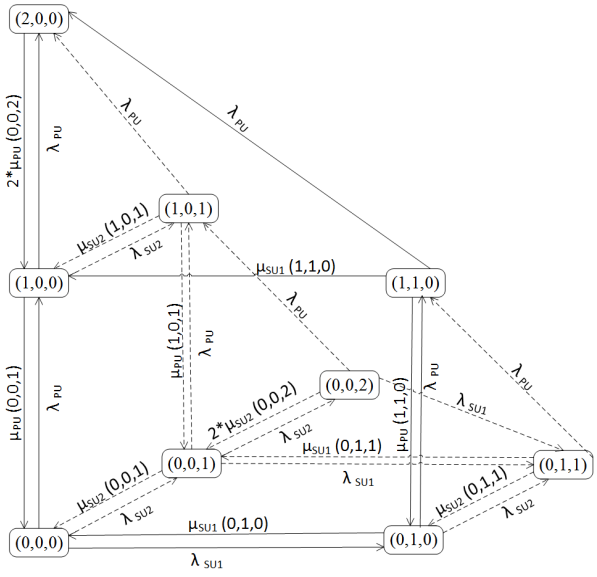


Fig. 5. *Example CRN with channel reservation applied to SU_1*

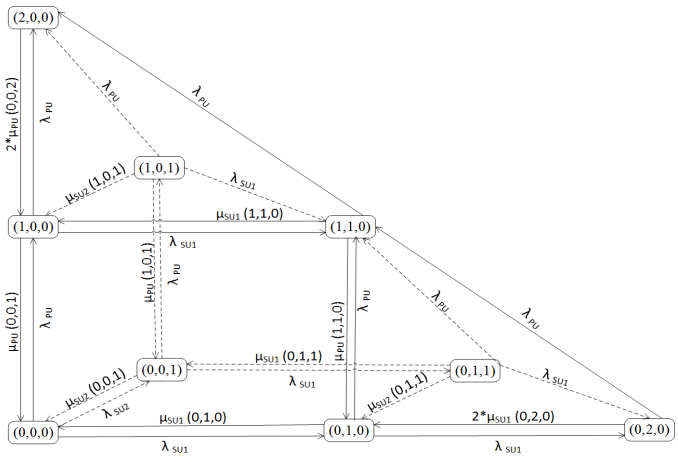


Fig. 6. *Example CRN with channel reservation applied to SU_2*

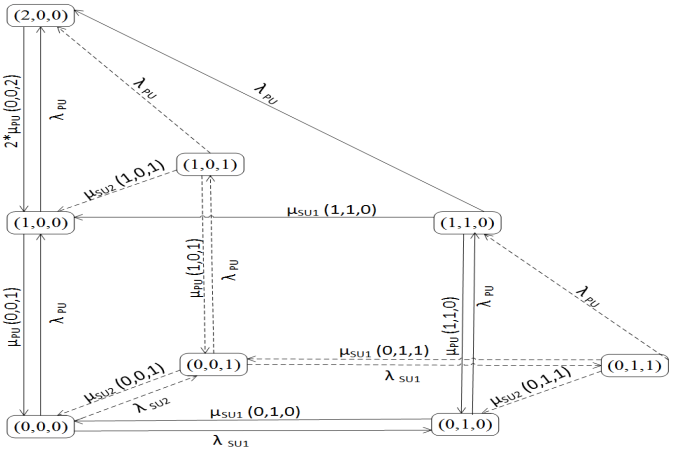


Fig. 7. *Example CRN with channel reservation applied to SU_1 and SU_2*

IV. Performance Metrics

Considering that $\pi(i,j,k)$ is the steady state probability for state (i,j,k), the probability values can be found by solving a linear system resulting from two equation blocks, where (5) is known as balance equations and (6) the normalization condition.

$$\sum_{i\prime=0}^{\left\lfloor \frac{N}{B_{PU}} \right\rfloor} \sum_{j\prime=0}^{\left\lfloor \frac{N-R_1}{B_{SU1}} \right\rfloor} \sum_{k\prime=0}^{\left\lfloor \frac{N-R_2}{B_{SU2}^{min}} \right\rfloor} \pi(i,j,k) * \gamma_{(i,j,k)}^{(i',j',k')} = \sum_{i\prime=0}^{\left\lfloor \frac{N}{B_{PU}} \right\rfloor} \sum_{j\prime=0}^{\left\lfloor \frac{N-R_1}{B_{SU1}} \right\rfloor} \sum_{k\prime=0}^{\left\lfloor \frac{N-R_2}{B_{SU2}^{min}} \right\rfloor} \pi(i',j',k') * \gamma_{(i',j',k')}^{(i,j,k)}$$
$$\text{with } 0 \le i \le \left\lfloor \frac{N}{B_{PU}} \right\rfloor, 0 \le j \le \left\lfloor \frac{N-R_1}{B_{SU1}} \right\rfloor, 0 \le k \le \left\lfloor \frac{N-R_2}{B_{SU2}^{min}} \right\rfloor \text{ and } (i',j',k') \ne (i,j,k). \quad (5)$$

$$\sum_{i=0}^{\left\lfloor \frac{N}{B_{PU}} \right\rfloor} \sum_{j=0}^{\left\lfloor \frac{N-R_1}{B_{SU1}} \right\rfloor} \sum_{k=0}^{\left\lfloor \frac{N-R_2}{B_{SU2}^{min}} \right\rfloor} \pi(i,j,k) = 1. \quad (6)$$

The solution for the linear system formed by (5) and (6) is the steady state probability vector, which is used to formulate the blocking and forced termination probabilities that are relevant to analyze the secondary communication QoS, and can be calculated in the following.

A. Blocking Probability

An SU is blocked when it tries to access the CRN but there are no available channels, so, the blocking probability is the secondary requests´ percentage that are not accepted. Let BP_{SU1} and BP_{SU2} denote the SU_1's (7) and SU_2's (8) blocking probability, which will be equal to the sum of the steady probabilities of all states that characterize a full network.

$$\mathrm{BP_{SU1}} = \sum_{i=0}^{\left\lfloor \frac{N}{B_{PU}} \right\rfloor} \sum_{j=\left\lceil \frac{N-R_1-(i*B_{PU})}{B_{SU1}} \right\rceil}^{\left\lfloor \frac{N-R_1}{B_{SU1}} \right\rfloor} \sum_{k=0}^{\left\lfloor \frac{N-R_2}{B_{SU2}^{min}} \right\rfloor} \pi(i,j,k). \quad (7)$$

$$\mathrm{BP_{SU2}} = \sum_{i=0}^{\left\lfloor \frac{N}{B_{PU}} \right\rfloor} \sum_{j=0}^{\left\lfloor \frac{N-R_1}{B_{SU1}} \right\rfloor} \sum_{k=\left\lceil \frac{N-R_2-(i*B_{PU})-(j*B_{SU1})}{B_{SU2}^{min}} \right\rceil}^{\left\lfloor \frac{N-R_2}{B_{SU2}^{min}} \right\rfloor} \pi(i,j,k). \quad (8)$$

B. Forced Termination Probability

Once the SUs are admitted, their communication may be abruptly interrupted by higher priority user arrivals, causing secondary QoS degradation. The SU_1 is terminated if there are no available resources for an arriving PU, i.e., $idle + (k * B_{SU2}) < B_{PU}$. Considering that the system state is *(i,j,k)* and using idle to denote the number of available resources in state *(i,j,k)*, the number of forcibly terminated SU_1s is $y = \lceil (((i+1) * B_{PU}) + (j * B_{SU1}) - N)/B_{SU1} \rceil$. Then, the SU_1´s forced termination rate will be $\lambda_{PU} * \pi(i,j,k) * y(i,j,k)$ and by dividing it by the admitted SU_1's rate, $(1 - BP_{SU1}) * \lambda_{SU1}$, the SU_1 forced termination probability can be obtained by (9). For this metric, we provide a feasible state indication function $I(i,j,k) = \begin{cases} 1, if\ (i,j,k) \in \Omega; \\ 0, otherwise. \end{cases}$ that guarantees that only the relevant states are considered.

$$\mathrm{FTP_{SU1}} = \frac{\lambda_{PU}}{(1 - \mathrm{BP_{SU1}}) * \lambda_{SU1}} \sum_{i=0}^{\left\lfloor \frac{N}{B_{PU}} \right\rfloor} \sum_{j=0}^{\left\lfloor \frac{N-R_1}{B_{SU1}} \right\rfloor} \sum_{k=0}^{\left\lfloor \frac{N-R_2}{B_{SU2}^{min}} \right\rfloor} \pi(i,j,k) * y * I(i,j,k) * I(i+1, j-y, k). \quad (9)$$

The lowest user priority class is subject to forced termination events in the following situations: First, an arriving PU may drop a SU_2 if there are no available resources upon a PU arrival, i.e., $idle < B_{PU}$ and, in the same way, by an arriving SU_1, where the condition is $idle < B_{SU1} + R_1$. The number of SU_2s to be dropped is $z_1 = \lceil (-idle + B_{PU})/B_{SU2} \rceil$ in the first case and $z_2 = \lceil (-idle + B_{SU1})/B_{SU2}^{min} \rceil$ in the latter. Thus, the SU_2's dropping rates due to first (C_1) and second (C_2) cases are given by (10) and (11), respectively. Thereof, the SU_2 forced termination probability is denoted as the ratio between the rate of dropped SU_2s and the rate of admitted SU_2s (12).

$$\mathrm{C_1} = \sum_{i=0}^{\left\lfloor \frac{N}{B_{PU}} \right\rfloor} \sum_{j=0}^{\left\lfloor \frac{N-R_1}{B_{SU1}} \right\rfloor} \sum_{k=0}^{\left\lfloor \frac{N-R_2}{B_{SU2}^{min}} \right\rfloor} \lambda_{PU} * \pi(i,j,k) * z_1 * I(i,j,k) * I(i+1, j, k-z). \quad (10)$$

$$\mathrm{C_2} = \sum_{i=0}^{\left\lfloor \frac{N}{B_{PU}} \right\rfloor} \sum_{j=0}^{\left\lfloor \frac{N-R_1}{B_{SU1}} \right\rfloor} \sum_{k=0}^{\left\lfloor \frac{N-R_2}{B_{SU2}^{min}} \right\rfloor} \lambda_{SU1} * \pi(i,j,k) * z_2 * I(i,j,k) * I(i, j+1, k-z). \quad (11)$$

$$\mathrm{FTP_{SU2}} = \frac{1}{(1 - \mathrm{BP_{SU2}}) * \lambda_{SU2}} * (\mathrm{C_1} + \mathrm{C_2}). \quad (12)$$

V. Numerical Results

In this section, simulation (markers) and analytical (continuous lines) results are analyzed under variable network conditions. The evaluation follows this sequence: the flexibility analysis towards the bandwidth and multi-level channel reservation are described in *sections A* and *B*, while *section C* presents a final experiment demonstrating how the aggregation techniques can be useful for suppressing the negative effects caused by channel reservation. The simulation was provided from a mean of one hundred executions with simulation time set to 10^4 time units and the results presented a 95% confidence level according to the statistical significance assessment, although no bars were drawn due to a small difference between upper and lower bounds.

A discrete-event simulator was used to map the system's operation so as to reproduce the behavior of a centralized CRN, based on a discrete event sequence (in time), each of which occurs at an unique instant and may cause a state change, as opposed to real-time simulations. No change is assumed to occur between consecutive events and the performance metrics are not analytically derived from probability distributions, but rather as averages from different runs. In addition, the FIFO service discipline was adopted and the arrival order is obtained after merging the three independent event lists, one for each user type.

A. Multiple Bandwidth Arrangements

The switchover to digital television frees up valuable spectrum chunks, but can still present active PUs that will usually use wideband transmissions. Because CR is expected to allow the coexistence between PUs and SUs, the literature frequently considers the PUs to have larger or equal bandwidths compared to the opportunistic users. However, narrowband primary applications may also exist in those frequency bands, e.g., IEEE 802.22 (wireless regional area network) that uses cognitive radio to avoid interference with incumbent TV broadcasting (wideband primary application) and low power licensed devices such as wireless microphone operation (narrowband primary application).

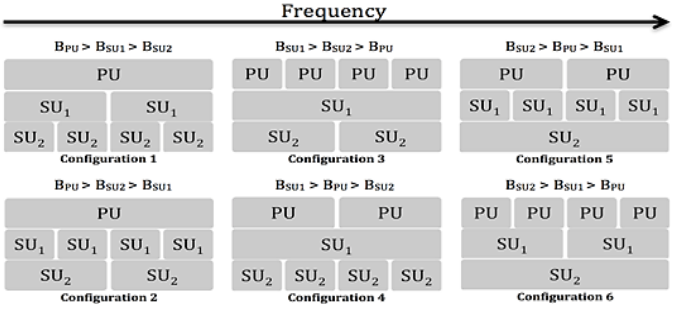
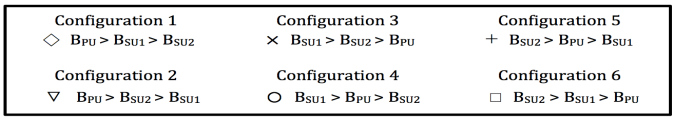


Fig. 8. *Six possible bandwidth configurations addressed by this model*

In this experiment, the requested bandwidth was varied for each user level, not considering dynamic bandwidth but rather using a fixed strategy (FCA). Then, the blocking and forced termination rates were computed under different network conditions by tuning the PU arrival rate. Previous works have limitations towards the bandwidth requirement, whereas this

model allows any combination of such parameter. Six different input configurations were depicted in Fig. 8, where the number of channels per user is one, two and four. For this experiment, the total number of channels (N) was set to four, but any amount could have been used, provided that N is always greater or equal than the highest required bandwidth. The analytical (solid lines) and simulation (markers) results were obtained using the following parameters: $N = 4$, $\lambda_{SU1} = \lambda_{SU2} = 1, \mu_P = 1$, $\mu_{SU1} = 1$ and $\mu_{SU2} = 1$.

According to Fig. 9, the requested bandwidth significantly impacts on the SU_1 blocking probability. For instance, larger idle bandwidth amounts may be hardly available depending on the requested size and primary network occupation, which can hinder network acceptance. For this reason, configurations 3 & 4 ('x' and circle markers, respectively) that assign the largest bandwidth B_{SU1} present the highest SU_1 blocking probabilities. The same applies to Fig. 10, where larger B_{SU2} configurations 5 & 6 (cross and square markers, respectively) experience higher SU_2 blocking probability. Contrarily, larger bandwidth diminishes service duration, i.e., when the PU requires more bandwidth it frees up its channel resources faster, enabling less SU_1 and SU_2 blocking events. This is the key fact behind the difference between configurations 3 (highest blocking) & 4 in Fig. 9 and configurations 5 & 6 (highest blocking) in Fig. 10, as the PU bandwidth varies in these configurations.



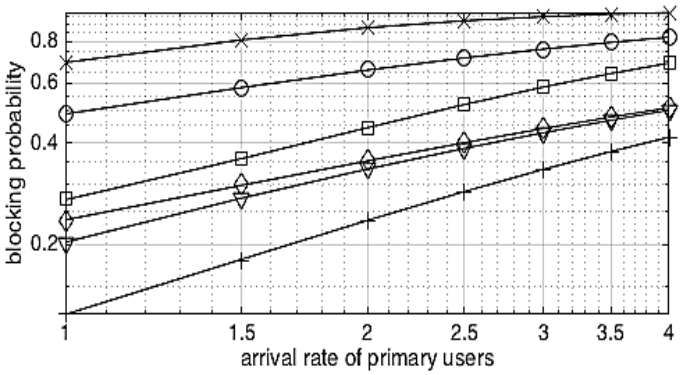


Fig. 9. *SU_1 Blocking Probability as a function of* λ_{PU}

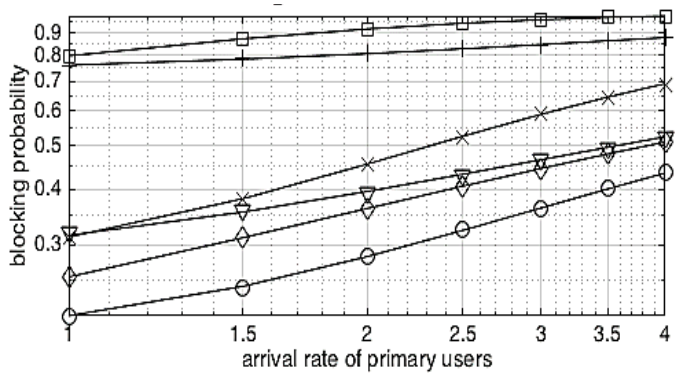


Fig. 10. *SU_2 Blocking Probability as a function of* λ_{PU}

The best performances for blocking probability in both secondary layers were found in configurations 2 & 5 for Fig. 9 and in configurations 1 & 4 for Fig. 10, since smaller bandwidth requests enable the users to be easily accepted by the CRN, even under crowded network conditions. With regard to blocking probability, although prioritization plays a significant role, the bandwidth amounts also impacts on user resource allocation. In Fig. 10, for example, configuration 4 under a dense scenario ($\lambda_{PU} = 4$) has a smaller SU_2's blocking probability (42.5%) than the corresponding SU_1's blocking probability in Fig. 9 (50%) on a sparse scenario ($\lambda_{PU} = 1$), even though the SU_1s have access priority over SU_2s.

According to Figs. 11 and 12, smaller bandwidth setups such as in configuration 2 (inverted triangle marker - Fig. 11) and configuration 1 (diamond marker - Fig. 12) may not allow fast service completion, increasing their chances to be terminated due to user arrivals. Similarly, large secondary bandwidth configurations were also not effective in terms of forced termination probability because, although shorter service duration is enabled, they also pose some threat as higher-class user arrivals may require the channels in use by the secondary system, forcing these to vacate and search for available channels to resume its communication. However, larger idle bandwidth amounts may be hardly available. On the contrary, intermediate bandwidth setups, e.g., configuration 6 (square marker) in Fig. 11 for the SU_1 and configuration 3 ('x' marker) in Fig. 12 for the SU_2, had the best tradeoff between occupied amount of resources and total service time, which mitigated the forced termination probabilities.

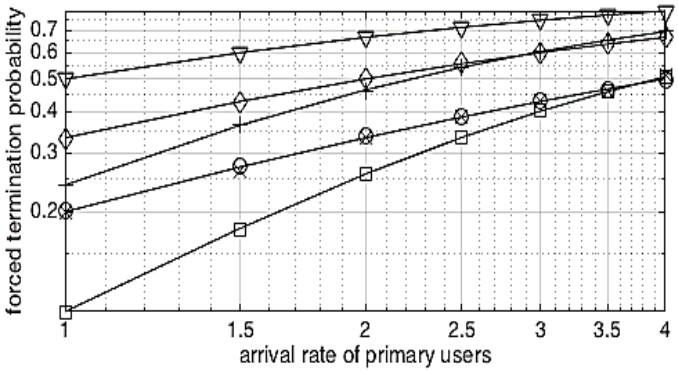


Fig. 11. *SU_1 Forced Termination Probability as a function of* λ_{PU}

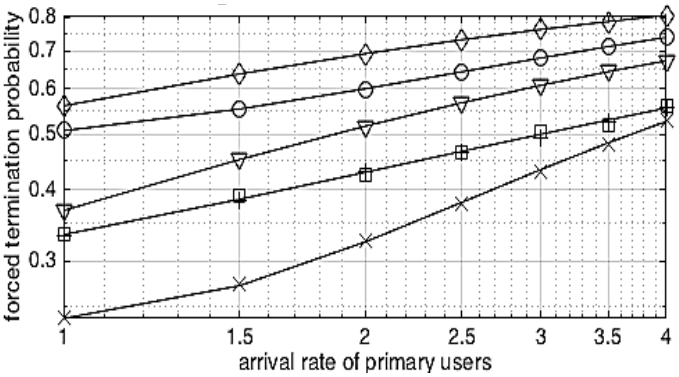


Fig. 12. *SU_2 Forced Termination Probability as a function of* λ_{PU}

This section provided insights on how different applications (with specific bandwidth requirements) could be structured in a CRN, aiming at the secondary service quality that was calculated in terms of blocking and forced termination probabilities. The results illustrated how the bandwidth configuration and priority level impact on the secondary communication for a three-layered CRN under different PU loads. As far as we are concerned, this is the most flexible model considering that previous authors have not addressed all bandwidth possibilities simultaneously in a unique system.

B. Multi-Level Channel Reservation

In this section, another model feature known as channel reservation that has been previously presented in the literature for QoS provisioning in CRNs is modified and analyzed. In this work, allowing channel reservation means diminishing the network availability for a determined set of users (SUs), which necessarily increases the blocking events but lessens the number of secondary sessions to be forcibly terminated. Consequently, it enables a tradeoff between forced termination and blocking probabilities that can be tuned according to the QoS requirements of the secondary traffic.

The problem with most works is that the experiments are adjusted to maximize the channel reservation´s benefits, i.e., most scenarios are built under low PU loads. For example, in [15] the PU arrival rate´s ranges from 0.1 to 1 and a PU service rate is equal to 30. In other words, this results in an extremely low PU load. For this reason, we have opted for higher PU loads to compute the afore-mentioned tradeoff under a harsh scenario. Also, we have provided two channel reservation variables R_1 and R_2, one for each secondary layer, which causes different effects on the secondary´s performance, differing from previous approaches such as [15], that had a single variable for both secondary layers.

Being R_1 and R_2 the number of reserved channels from each secondary layer (SU_1 and SU_2), each user layer may access only $N - R_1$ and $N - R_2$ channels. Four different input configurations were computed: $R_1 = R_2 = 0$ (configuration 1), $R_2 = 0,\ R_1 = 1$ (configuration 2), $R_1 = 1,\ R_2 = 0$ (configuration 3) and $R_1 = R_2 = 1$ (configuration 4) and the total number of channels in this CRN was set to four. Moreover, we have configured all three layers to use fixed channel aggregation (FCA) and all user bandwidths were set to one unit. Two curves for each configuration were plotted for our performance metrics. The results from the second experiment are divided into two groups, but it presents SU_1 metrics separately from the SU_2 metrics.

Because there are four configurations, the images should provide four different curves (eight if counting theoretical and simulation). Figs. 13 and 14 seem to show only half of the outputs, however, they correctly present the expected number of curves, but some overlap. Such behavior occurs because only one user layer, in this case the PUs, pressures the SU_1s for their resources, ergo, configurations one and two ($R_1 = 0$) will naturally provide the same values while configurations three and four ($R_1 = 1$) will result in another equal set.

Isolating the point where the PU arrival rate is 1.4 (highest PU load), we have depicted the tradeoff (Fig. 15) between the SU_1's blocking (Fig. 13) and forced termination (Fig. 14) probabilities. In configurations 1 & 2 the reservation for SU_1 is not considered ($R_1 = 0$), whereas in configurations 3 & 4 $R_1 = 1$, disabling one channel from the SU_1´s perspective. It was noted that when channel reservation is triggered, the blocking probability approximately doubles its value while the forced termination probability reduces by a factor of only 1.5, hence, in this scenario channel reservation does not pay.

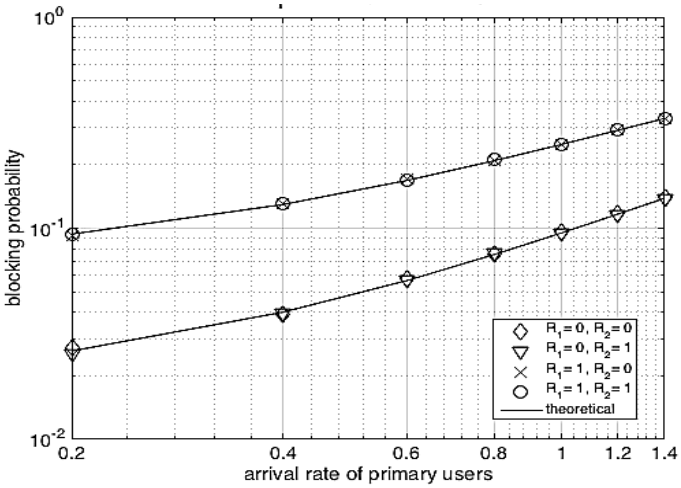


Fig. 13. *SU_1 Blocking Probability*

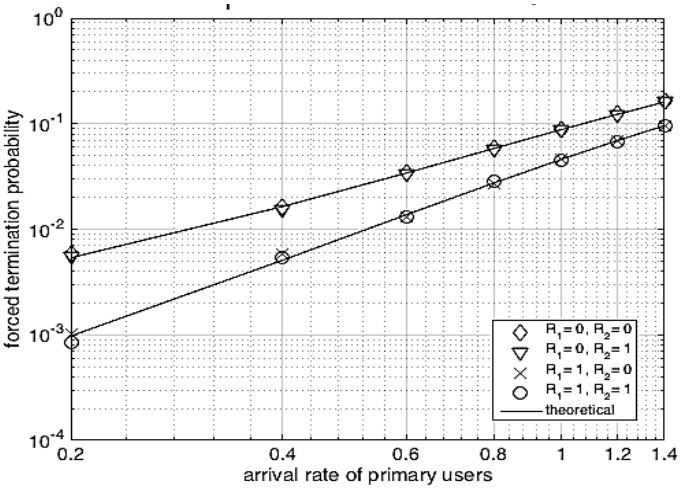


Fig. 14. *SU_1 Forced Termination Probability*

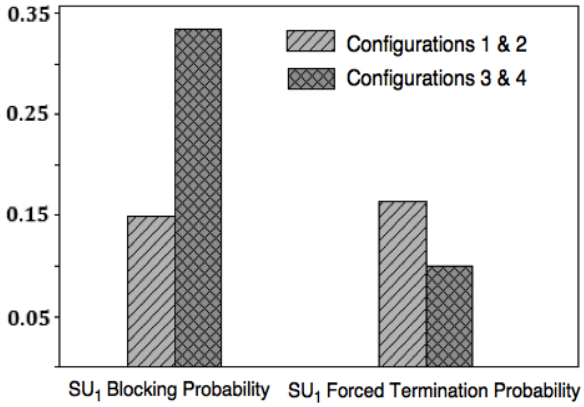


Fig. 15. *Tradeoff between blocking and forced termination probabilities for a PU arrival rate of 1.4*

Differently from the SU_1, Figs. 16 and 17 show four configurations producing distinct curves since the SU_1 is prioritized and also influences the SU_2 behavior. Thus, analyzing the blocking probability in Fig. 16, for instance, configuration three ($R_1 = 1, R_2 = 0$) has the lowest blocking values followed by configuration one ($R_1 = R_2 = 0$). In brief, configuration three provides fewer channels ($N - R_1$) for SU_1 admission, but gives full resources for the SU_2s ($N - R_2 = N$). Consequently, fewer SU_1s undermine the SU_2s´ performance. Curiously, regarding the forced termination probability (Fig. 17), it was observed that configuration three provides the third worse performance, which might seem contradictive at a first sight. However, because configurations two and four have $R_2 = 1$, it implies fewer SU_2s are being admitted in the CRN, which lessens the probability of a SU_2 being forcibly terminated. Briefly, for this experiment, configuration three seems to have the best compromise between the adopted metrics since it had the best blocking rate and a reasonable forced termination probability, with a difference of only 5.5% compared to the best alternative for this metric (configuration four). On the other hand, again for a crowded network, channel reservation seems not to be worthwhile.

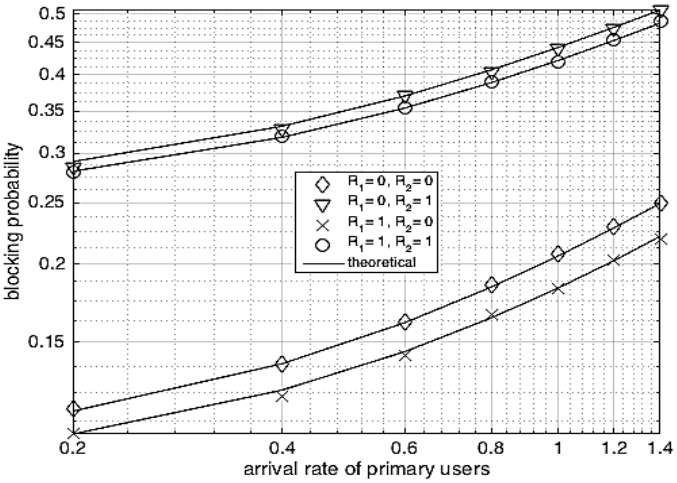


Fig. 16. *SU_2 Blocking Probability*

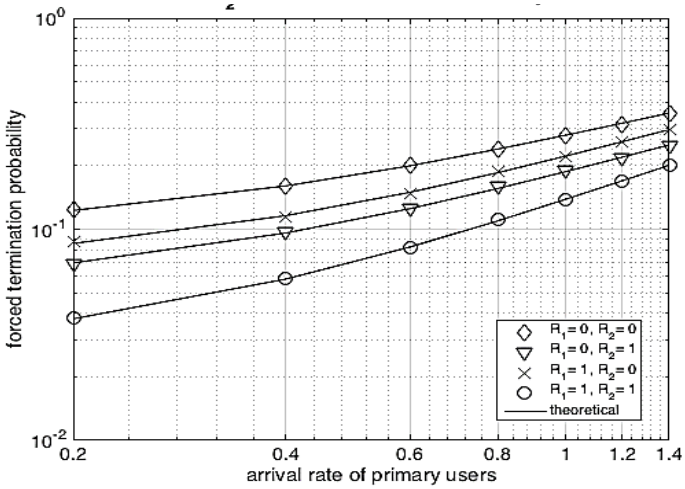


Fig. 17. *SU_2 Forced Termination Probability*

C. Combined use of Channel Aggregation and Reservation

The current section shows the achievable performance improvement triggered by both channel aggregation and channel reservation, simultaneously. As previously discussed, channel reservation might not be an effective approach depending on the network load; however, the aggregation techniques should always enhance the system's performance, regardless the network's state. The following experiment joins both techniques to mitigate the negative effects that may be caused by channel reservation on the lowest secondary layer (SU_2s). For this experiment, DCA was preferred over DCAF due to its feasibility in real scenarios and the performance similarity, as previously discussed in [17]. It will be compared to the fixed aggregation approach (FCA) together will variable reservation values for R_1 and R_2. Tab. 3 outlines the four configurations to be tested and, thus, four curves were generated for each metric.

Tab. 3. *Bandwidth and reserved channels for the fourth experiment*

n°	N° of channels	B_{PU}	B_{SU1}	$[B_{SU2}^{min}, B_{SU2}^{max}]$	R_1	R_2
1	12	1	2	FCA [4,4]	0	2
2	12	1	2	FCA [4,4]	0	4
3	12	1	2	FCA [4,4]	4	6
4	12	1	2	DCA [1,4]	4	6

The PU arrival rate varied from 1 to 4 with a step of 0.5 and the remaining arrival and service rates were set to: $N = 4$, $\lambda_{SU1} = \lambda_{SU2} = 1, \mu_P = 1$, $\mu_{SU1} = 1$ and $\mu_{SU2} = 1$. Again, these values were chosen according to Little's law, providing low and high PU loads.

Configuration 1 ($R_1 = 0$ and $R_2 = 2$) enables channel reservation in the lowest priority user layer, which means that for the SU_2 there are not 12 channels to be used, but 12 – 2 = 10 channels. For such scenario, even though it is not capable of using the full network, it achieves the lowest blocking probability (see Fig. 18) since the other configurations have higher reservation values $R_2 = 4$, $R_2 = 6$ and $R_2 = 6$, respectively, i.e., they may use fewer channels than those available for configuration 1. Surprisingly, configuration 4 that applies channel aggregation through the DCA technique, performs very similarly to configuration 1, although it uses only 12 – 6 = 6 channels. Moreover, configuration 4 achieves much lower blocking values compared to configuration 3. These inputs differ by the aggregation technique, which in configuration 3 is FCA [4,4] and in configuration 4 is DCA [1,4]. It was noticed that the lower bandwidth bound of a single channel unit for configuration 4 directly contributes for its performance compared to the input that applies FCA, which in this case will experience higher blocking rates because its minimum bandwidth is set to four channel units, i.e., 1/3 of the network total. Such condition hampers the chances of an SU_2 to be accepted by the CRN.

Regarding the forced termination probability results in Fig. 19, as more channels are reserved from the SU_2, less termination is experienced. Naturally, the sequence from the greatest to the smallest values for this metric is: FCA with $R_2 = 2$ and $R_2 = 4$, FCA with $R_2 = 6$ and DCA with $R_2 = 6$ where much lower values were registered, mainly because it uses multi-level channel reservation for denying some SU_1s but

also because they reserve more channels *($R_2 = 6$)* for the SU_2. For theses inputs, a performance switch happens when the PU arrival rate is around 3.5, making FCA with $R_2 = 6$ the configuration with the smallest forced termination probability. This might seem unusual since DCA [1,4] allows dynamic CA, but one should consider that DCA [1,4] has a much lower blocking probability compared to FCA ($R_2 = 6$), therefore, more SU_2s are accepted by the network. A performance switch was noted when the PU arrival rate is set to 3.5 (Fig. 18), where the users began to experience more service interruptions than those in FCA with $R_2 = 6$.

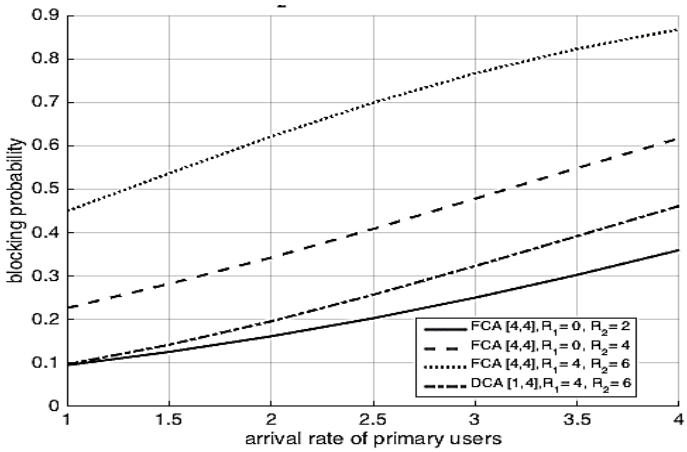


Fig. 18. *SU_2 Blocking Probability*

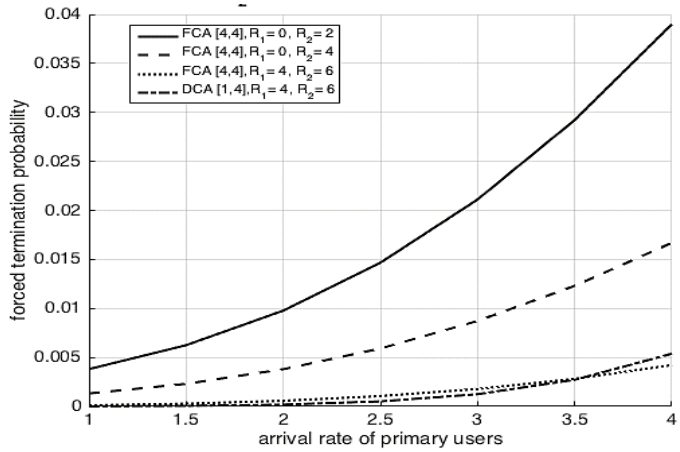


Fig. 19. *SU_2 Forced Termination Probability*

VI. CONCLUSION

This paper proposes and analyses a heterogeneous CRN queue-based analytical model that can help to overcome various challenges such as network dimensioning and secondary QoS guarantees. Although previously explored, many authors seem to limit their analysis by considering specific network conditions. Hence, from the resource allocation perspective, a more complete model was outlined, encompassing multiple bandwidth combinations among primary and secondary users, multi-level channel reservation and different channel aggregation approaches. The developed analytical model is validated by extensive simulations and evaluated in terms of blocking and forced termination probabilities. The results have shown that channel reservation not always provides reasonable performance tradeoffs as most works suggested, on the contrary, its success highly depends on the network's state, i.e., it can slightly increase the blocking probability but substantially decrease the forced termination rate, if the PU/SU load is reasonably low. However, either for high PU or SU loads it causes a skyrocketing blocking rate increase and minimum termination reduction. On the other hand, channel aggregation was proven to be efficient in any given scenario, thus, it was combined with channel reservation to mitigate the blocking event increase, while maintaining a low number of forced terminations, sustaining the best properties of both.

REFERENCES

[1] D. Jiang, G. Liu ."An Overview of 5G Requirements", In: 5G Mobile Communications, pp. 3-26, Springer, 2017.

[2] C. Yang, J. Li, M. Guizani, A. Anpalagan and M. Elkashlan, "Advanced spectrum sharing in 5G cognitive heterogeneous networks," IEEE Wireless Communications, vol. 23, no. 2, pp. 94-101, April 2016.

[3] O. Adigun, M. Pirmoradian, and Christos Politis, "Cognitive Radio for 5G Wireless Networks" In Fundamentals of 5G Mobile Networks, pp. 149-163, Wiley, 2015.

[4] X. Hong, J. Wang, C-X. Wang, J. Shi, "Cognitive radio in 5G: a perspective on energy-spectral efficiency trade-off", IEEE Communications Magazine, vol. 52, no. 7, pp. 46-53, 2014.

[5] H. Bogucka, P. Kryszkiewicz, and A. Kliks. "Dynamic spectrum aggregation for future 5G communications." IEEE Communications Magazine, vol. 53, n. 5, pp. 35-43, 2015.

[6] X. Zhu, L. Shen and T. S. P. Yum, "Analysis of Cognitive Radio Spectrum Access with Optimal Channel Reservation," IEEE Communications Letters, vol. 11, no. 4, pp. 304-306, 2007.

[7] L. Jiao, V. Pla and F. Y. Li, "Analysis on channel bonding/aggregation for multi-channel cognitive radio networks", European Wireless Conference (EW), pp. 468-474, 2010.

[8] L. Jiao, F. Y. Li and V. Pla, "Modeling and Performance Analysis of Channel Assembling in Multichannel Cognitive Radio Networks with Spectrum Adaptation," IEEE Transactions on Vehicular Technology, vol. 61, no. 6, pp. 2686-2697, 2012.

[9] L. Li, S. Zhang, K. Wang and W. Zhou, "Queuing method in combined channel aggregation and fragmentation strategy for dynamic spectrum access," IEEE 23rd International Symposium on Personal, Indoor and Mobile Radio Communications - (PIMRC), pp. 1214-1219, 2012.

[10] L. Jiao, I. A. M. Balapuwaduge, F. Y. Li and V. Pla, "On the Performance of Channel Assembling and Fragmentation in Cognitive Radio Networks," IEEE Transactions on Wireless Communications, vol. 13, no. 10, pp. 5661-5675, 2014.

[11] A. N. Dudin, M. H. Lee, O. Dudina and S. K. Lee, "Analysis of Priority Retrial Queue with Many Types of Customers and Servers Reservation as a Model of Cognitive Radio System", IEEE Transactions on Communications, vol. 65, no. 1, pp. 186-199, 2017.

[12] G. Ding and Q. Zhao, "Analysis on the performance of special channel reservation mechanism in cognitive radio," First IEEE International Conference on Computer Communication and the Internet (ICCCI), pp. 37-40, 2016.

[13] S. Ai, L. Jiao, F. Y. Li and M. Radin, "Channel aggregation with guard-band in D-OFDM based CRNs: Modeling and performance evaluation," IEEE Wireless Communications and Networking Conference, pp. 1-6, 2016.

[14] T. M. C. Chu, H. Phan and H. J. Zepernick, "Dynamic Spectrum Access for Cognitive Radio Networks with Prioritized Traffics" IEEE Communications Letters, vol. 18, no. 7, pp. 1218-1221, 2014.

[15] T. M. C. Chu, H. J. Zepernick and H. Phan, "Channel reservation for dynamic spectrum access of cognitive radio networks with prioritized traffic", 2015 IEEE International Conference on Communication Workshop (ICCW), pp. 883-888, 2015.

[16] I. A. M. Balapuwaduge, L. Jiao, V. Pla and F. Y. Li, "Channel Assembling with Priority-Based Queues in Cognitive Radio Networks: Strategies and Performance Evaluation" in IEEE Transactions on Wireless Communications, vol. 13, no. 2, pp. 630-645, 2014.

[17] M. Falcao, G. A. Silva, A. Balieiro and K. Dias, "Three-layered prioritized cognitive radio networks with channel aggregation and fragmentation techniques," 8th IEEE Latin-American Conference on Communications (LATINCOM), pp. 1-5, 2016.

[18] M. Fitch, M, Nekovee, S, Kawade, K. Briggs, and R. Mackenzie, “Wireless Service Provision in TV White Space with Cognitive Radio Technology: A Telecom Operator’s Perspective and Experience”, IEEE Communication Magazine, vol. 49, pp. 64-73, 2011.

[19] J. Wang, M. Gosh and K. Challapali, “Emerging Cognitive Radio Applications: A Survey”, IEEE Communications Magazine, vol 49, pp. 74-81, 2011.

[20] J. Martinez-Bauset, V. Pla and D. Pacheco-Paramo, "Comments on "analysis of cognitive radio spectrum access with optimal channel reservation"," IEEE Communications Letters, vol. 13, no. 10, pp. 739-739, 2009.

[21] W. Ahmed, J. Gao, H. A.Suraweera, M. Faulkner, "Comments on" Analysis of cognitive radio spectrum access with optimal channel reservation." IEEE Transactions on Wireless Communications, vol. 8, n. 9, pp. 4488-4491, 2009.