Pós-Graduação em Ciência da Computação

MARCOS ROCHA DE MORAES FALCÃO

"CHANNEL RESERVATION AND SPECTRUM ADAPTATION STRATEGIES IN A MULTI-LEVEL PRIORITIZED COGNITIVE RADIO NETWORK"

MSc. Dissertation

Universidade Federal de Pernambuco
posgraduacao@cin.ufpe.br
www.cin.ufpe.br/~posgraduacao

RECIFE
2016

MARCOS ROCHA DE MORAES FALCÃO

“CHANNEL RESERVATION AND SPECTRUM ADAPTATION STRATEGIES IN A MULTI-LEVEL PRIORITIZED COGNITIVE RADIO NETWORK "

THIS DISSERTATION HAS BEEN SUBMITTED TO THE GRADUATE PROGRAM IN COMPUTER SCIENCE OF THE INFORMATICS CENTER OF THE FEDERAL UNIVERSITY OF PERNAMBUCO AS A PARTIAL REQUIREMENT TO OBTAIN THE MASTER DEGREE IN COMPUTER SCIENCE.

ADVISOR: KELVIN LOPES DIAS
CO-ADVISOR: ANDSON MARREIROS BALIEIRO

RECIFE
2016

Catalogação na fonte
Bibliotecária Monick Raquel Silvestre da S. Portes, CRB4-1217

F178c Falcão, Marcos Rocha de Moraes
Channel reservation and spectrum adaptation strategies in a multi-level prioritized cognitive radio network / Marcos Rocha de Moraes Falcão. – 2016.
95 f.: il., fig., tab.

Orientador: Kelvin Lopes Dias.
Dissertação (Mestrado) – Universidade Federal de Pernambuco. CIn, Ciência da Computação, Recife, 2016.
Inclui referências e apêndices.

1. Redes de computadores. 2. Teoria de filas. I. Dias, Kelvin Lopes (orientador). II. Título.

004.6 CDD (23. ed.) UFPE- MEI 2016-150

Marcos Rocha de Moraes Falcão

Channel Reservation and Spectrum Adaptation Strategies in a Multi-level Prioritized Cognitive Radio Network

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de Pernambuco, como requisito parcial para a obtenção do título de Mestre em Ciência da Computação.

Aprovado em: 29/07/2016

BANCA EXAMINADORA

Prof. Dr. Stênio Flávio de Lacerda Fernandes
Centro de Informática / UFPE

Prof. Dr. Eduardo de Melo Vasconcelos
Instituto Federal de Pernambuco/Campus Garanhuns

Prof. Dr. Kelvin Lopes Dias
Centro de Informática / UFPE
(**Orientador**)

Acknowledgments

I would like to thank God for providing me incredible people that supported me during this work's design. Besides my family, Marcos, Teresa and Juliana, I would like to thank my team, composed by Kelvin, Andson and Gabriel, who were key to my success.

Abstract

Wireless technologies have dominated the communication's market by offering reasonable speeds and convenience at low deployment costs. However, due to the significant growth of mobile computing devices and their bandwidth demands, together with the paradigm shift brought by the Internet of things, future wireless networks should become highly dense and heterogeneous, which will hardly cope with the traditional fixed spectrum allocation policy. Some standards such as the Long Term Evolution – Advanced (LTE-A), have already set the precedent for carrier aggregation (CA), aiming at scaling up bitrates, which partially helps solving the problem. However, cognitive radio (CR) has been put forward as the most promising solution to handle this complex ecosystem since it may provide better spectrum utilization and user coordination through non-traditional mechanisms. Among other features, it allows non-licensed users, known as secondary users (SUs) to opportunistically use temporarily idle licensed bands that are used by licensed clients called primary users (PUs). Once PUs and SUs are expected to share the same spectrum bands, a critical issue is to concomitantly avoid primary interference while supporting QoS for the secondary services. This dissertation studies the synergistic integration of cognitive radio networks (CRNs), Dynamic Spectrum Access DSA techniques and resource allocation strategies (e.g., CA) that combined, should improve the overall system's performance. We have proposed a layered M/M/N/N queue-based model that addresses three user priorities, flexible bandwidth choices, multi-level channel reservation and two channel aggregation strategies. Different network load conditions for each feature were evaluated in terms of four performance metrics: blocking probability, forced termination probability, spectrum utilization and throughput. Such study is particularly useful for understanding the effects of each of these approaches in the secondary network. To the best of our knowledge, our model fulfills almost completely the user bandwidth's possibilities, improves the existing channel reservation formulation and demonstrates that our proposed dynamic channel aggregation strategy performs similarly to a more complex simultaneous channel aggregation and fragmentation approach, but can be technically more feasible.

Keywords: Cognitive Radio Networks. Channel Reservation. Channel Aggregation. Queueing Theory.

Resumo

Tecnologias sem fio têm dominado o mercado das comunicações, oferecendo velocidades razoáveis e conveniência a um baixo custo de implantação. No entanto, devido ao crescimento significativo do número de plataformas computacionais móveis e de suas demandas por largura de banda, acrescido do advento da Internet das Coisas, as redes sem fio do futuro devem passar a ser muito mais densas e heterogêneas, sendo difíceis de se adequar a política tradicional de alocação espectral fixa. Recentemente, o método de agregação de portadora (AP) fora proposto no padrão Long Term Evolution – Advanced (LTE-A), com o propósito de aumentar as taxas de bit, mitigando assim parte do problema. Todavia, rádio cognitivo (RC) foi apresentada como a solução mais promissora para lidar com este ecossistema complexo, uma vez que pode proporcionar uma melhor utilização do espectro e coordenação de usuários através de mecanismos não-tradicionais. Entre outras características, isso permite que usuários não-licenciados também conhecidos como usuários secundários (USs) utilizem de forma oportunista bandas licenciadas temporariamente ociosas, cujos clientes licenciados são também chamados de usuários primários (UPs). Como os UPs e os USs devem compartilhar as mesmas bandas, uma questão crítica é evitar interferência primária e concomitantemente apoiar a qualidade de serviço prestada aos USs. Esta dissertação estuda a integração sinérgica das redes de rádio cognitivas, técnicas de acesso dinâmico ao espectro e estratégias de alocação de recursos (AP), que combinados, devem melhorar o desempenho do sistema. Neste trabalho, propomos um modelo baseado em filas do tipo M/M/N/N, que inclui três prioridades de usuário, opções de largura de banda, reserva de canal multi-nível e duas estratégias de agregação de canal. Para cada recurso estudado, empregamos diferentes condições de carga de rede e avaliamos os resultados em termos de quatro métricas: probabilidade de bloqueio, a probabilidade de terminação forçada, utilização espectral e vazão. Este estudo é particularmente útil para compreender os efeitos de cada uma destas abordagens em relação à rede secundária. O modelo fornecido cumpre quase completamente as possibilidades do largura de banda de cada nível de usuário, melhora a formulação de reserva de canal existente e demonstra que estratégia de agregação de canais proposta possui performance similar a uma abordagem mais complexa de agregação e fragmentação simultânea, mas que seria tecnicamente mais viável.

Palavras-chave: Redes de Rádio Cognitivo. Reserva de Canais. Agregação de Canais. Teoria de Filas.

Figures Index

Tables Index

Abbreviations and Acronyms Index

CA	*Channel Aggregation*
CF	*Channel Fragmentation*
CDF	*Cumulative Distribution Function*
CR	*Cognitive Radio*
CTMC	*Continuous-time Markov Chains*
DSA	*Dynamic Spectrum Access*
DCA	*Dynamic Channel Aggregation*
DCAF	*Dynamic Channel Aggregation and Fragmentation*
ESA	*Equal Sharing Algorithm*
FCFS	*First Come, First Served*
FCA	*Fixed Channel Aggregation*
HetNet	*Heterogeneous Networks*
IEEE	*Institute of Electrical and Electronics Engineers*
IID	*Independent and Identically Distributed*
ISP	*Internet Service Provider*
LCFS	*Last Come, First Served*
LTE	*Long Term Evolution*
MAC	*Media Access Control*
OSA	*Opportunistic Spectrum Access*
PU	*Primary User*
QoE	*Quality of Experience*
QoS	*Quality of Service*
TVWS	*TV White Space*
SIRO	*Service In Random Order*
SP	*Service Provider*
SU	*Secondary User*
Wi-Fi	*Wireless Fidelity*
WiMAx	*Worldwide Interoperability for Microwave Access*
WLAN	*Wireless Local Area Network*

Contents

Chapter 1 - Introduction

A number of wireless access technologies (e.g., LTE-A, IEEE 802.11, IEEE 802.15.4, WiMax,), with different geographic coverage capacities (e.g. wide, local, and personal area) and for diverse purposes (e.g., wireless vehicular networks, wireless sensor networks) have emerged so as to meet the market's demands for speed, reliability and accessibility (WEN; TIWARY; LE-NGOC, 2013a). However, wireless networks are now experiencing heavier amounts of mobile traffic (HASEGAWA et al., 2014) caused by the significant growth of computing platforms (e.g. smart phones and tablets) and the paradigm shift brought by the Internet of things. In the early days of telecommunication, services were made available mainly for people communicating between themselves (e.g., cellular networks), but now, machines are able to the same (e.g., production monitoring in a factory floor), which highly increases the wireless traffic load. Due to the number of different services and their characteristics, this shift implies a future dense and heterogeneous wireless network that will hardly be compliant with the traditional spectrum allocation policy that establishes fixed spectrum bands for operation (STAPLE; WERBACH, 2004). This static division causes certain frequency bands, which are desirable for communication, to be crowded whereas others remain completely without use. Cognitive radio (CR) has been put forward as a promising solution to handle this complex ecosystem since it may provide better spectrum utilization and user coordination through non-traditional network mechanisms.

Most of the useful spectrum is allocated to licensed users, also known as primary users (PUs) (eg. mobile and TV companies). If the spectrum is opened for unlicensed use e.g., short-range networks, it is likely that new services will appear, for example, Wi-Fi and Bluetooth are now highly popular while operating in unlicensed bands. However, before opening the licensed spectrum to new users in a dynamic spectrum access (DSA) basis, it is necessary to guarantee that the PU will not be interfered. CR has been chosen to enable such DSA behavior due to its built-in cognition capability. A cognitive radio system can observe, learn and adjust radio parameters according to the environment conditions or application demands (AKYILDIZ et al., 2006). Among other features, this allows non-licensed users, that is, secondary users (SUs) to opportunistically use licensed bands. For example, A CR device may handoff to a different frequency band because its original network has suddenly become crowded, reducing the user's throughput. Hence, differently from the traditional fixed spectrum access idea that does not allow such move, CR enables the DSA approach that should support the continuous development of wireless services. Though, service providers

(SP) can increase their revenues by temporarily sharing unused spectrum bands to opportunistic users, thus introducing a new business model (MITOLA, 2006).

Once PUs and SUs are expected to share the same spectrum bands, a critical issue regarding spectrum sharing is to concomitantly avoid primary communication interference while supporting Quality of Service (QoS) for the secondary services. In other words, a cognitive radio network (CRN) should support the network's reliability and yet keep reasonable performance for both types of users. Since CRNs naturally introduces user prioritization and because the PU's performance ideally should not be diminished, the SUs may eventually suffer consequences that can range from service temporary/permanent interruption, throughput reduction or even service blocking (not being served at all).

Since Mitola introduced the cognitive radio concept in 1999 (MITOLA, 2006), the literature has investigated a number of techniques in order to simultaneously control interference and enhance the SU's performance. One of these mechanisms is known as admission control that among other features offers channel reservation while other are recently being deployed such as channel aggregation, specified in 2013 for LTE-A release 10 (3GPP SPECIFICATION, 2013).

Channel reservation is one way to avoid primary communication interference as it disables part of the network for unlicensed usage, thus the network becomes partitioned. In other words, part of total spectrum space is shared by PUs and SUs while solely the PUs use the other part. The effects on the SU is the increase in the blocking events, i.e., less network resources imply that it necessarily becomes crowded faster while it simultaneously experiences less dropping events, that is, less secondary services are interrupted during operation, mainly because there are less active SUs.

As in the Shannon's theorem, channel capacity is proportional to channel-width (bandwidth). Hence, higher throughput may be achieved if an SU is able to access multiple channels at the same time. Technically, this behavior known as carrier or channel aggregation (CA) can be implemented based on orthogonal frequency division multiplexing (OFDM). To increase flexibility, CA can be accompanied by spectrum adaptation. The meaning of spectrum adaptation is a twofold. Firstly, it is inherited from spectrum handover, allowing an ongoing SU service to jump onto another channel that is not occupied. Secondly, it is meant that ongoing SU services can adaptively adjust the number of aggregated channels according to the availability.

The motivation to propose the analysis of spectrum adaptation for channel aggregation is triggered by the following observation: the performance of secondary networks with channel aggregation when spectrum adaptation is enabled has been analyzed through mathematical models (ZHU et al., 2007), (LI et al., 2012) and (CHU et al., 2015), most of which considered only two user layers (PUs and SUs). The current work proposes a three-layered study with the secondary layer being divided into two segments (higher and lower priority). But either because the essential hardware tools for cognitive radio (e.g., Universal Software Radio Peripheral) are still tender or relatively expensive for implementing large networks, many researchers apply queueing theory to model such strategies.

This work provides an alternative three-layered M/M/N/N queue-based model that addresses three user priorities, flexible bandwidth choices, multi-level channel reservation and two channel aggregation strategies. To the best of our knowledge, our model fulfills almost completely the user bandwidth's possibilities, improves the existing channel reservation formulation and demonstrates that our proposed pure channel aggregation strategy performs similarly to a more complex simultaneous channel aggregation and fragmentation approach. For each feature, we have proposed a high number of scenarios characterized by different network load conditions and evaluated the system in terms of four performance metrics: blocking probability, forced termination probability, spectrum utilization and throughput.

1.1 Objectives

The main objective of this work is to model channel reservation and channel aggregation onto a prioritized three-layered CRN. Our main advantage compared to other models is the possibility of instantiating many bandwidth sizes to each of the three layers while other authors limit their analysis to the case where the PU bandwidth is greater than the SU's. The following specific objectives were listed in order to achieve our main goal:

- Study the most relevant existing solutions in order to understand how to model key features.
- To model the interactions between the three user layers with different sized bandwidths.
- To model a channel reservation approach.
- To model channel aggregation.
- To validate and verify the designed model.

1.2 Organization of the Dissertation

This dissertation is structured as follows. Chapter 2 examines the technical background material for this work, which includes the motivation for adopting cognitive radio, its concepts and categories. Furthermore, we present a brief overview of random variables, discrete event simulation, queueing theory and continuous-time Markov processes. In Chapter 3, there is critical review of the current literature, which includes the main features that previous CRN models have already addressed, as well as a discussion on their negative and positive characteristics. In addition, a classification of the related works is proposed. The proposed CRN model is described in Chapter 4. The analytical and simulated results are discussed in Chapter 5 and finally, Chapter 6 provides our concluding remarks and highlights future work directions.

Chapter 2 - Technical Background

In this Chapter we will review the main concepts that the reader should know in order to better understand the following chapters. First of all, we present the cognitive radio motivation, concept and categories, along with some QoS provisioning mechanisms examples and a brief outline on random variables and queueing theory.

2.1 Cognitive Radio

Modern wireless communications are prone to suffer due to the current static spectrum allocation policy. Cognitive radio (CR) is a promising paradigm for allowing dynamic spectrum access (DSA) and thereby addressing higher spectrum usage efficiency (AKYILDIZ et al., 2006). This section presents the CR concept and categories and the main quality of service (QoS) provisioning and admission control mechanisms derived from the DSA policy.

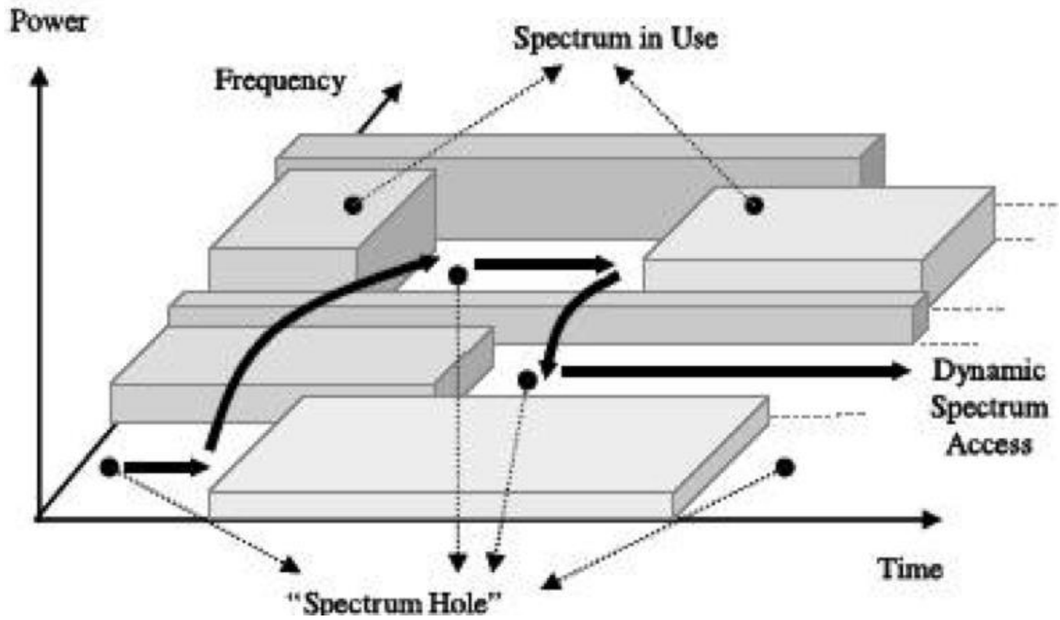
2.1.1 Motivation, Concept and Categories

The exponential rise in the amount of mobile traffic was caused by a significant growth in mobile platforms such as smart phones and by bandwidth-hungry applications (e.g. high definition video streaming). However, the natural resource that sustains wireless communications has not increased its capacity, making it dependent from new technologies for satisfying this great demand for connection. In order to ensure minimal interference among users, the current static allocation process delivers a given spectrum band to each licensed service. This scheme grants exclusive access to frequency bands by the license holders excluding any other application, even if those firstly allocated are not transmitting. The main problem regarding this policy is that PU operation (transmission or reception) might be considerably low, but still, in many countries it would not be allowed to share its spectrum band, causing spectrum underutilization (XIN; SONG, 2012).

Most researches agree on the fact that spectrum underutilization is the common condition among many countries. They also indicate that the traditional fixed spectrum allocation policy developed in the early 1900s for radio broadband (STAPLE; WERBACH, 2004) led to this problem. The specialists are currently questioning this strict paradigm and the concepts of CR and DSA emerged as promising approaches to solve the problem.

Cognitive radio is a radio that can change its transmitter/reception parameters according to environment interactions and user requirements (AKYILDIZ et al., 2006). CR is able to locate temporarily idle licensed spectrum bands, enabling non-licensed users, often referenced as secondary users (SUs), to access the licensed band opportunistically. When the PU returns to its spectrum band, the SU should vacate and search for another spectrum hole so as to resume its communication, in a process called spectrum handoff. Briefly, the combination of CR and DSA should enable opportunistic spectrum access pictured in Fig. 2.1.

Fig. 2.1. Dynamic spectrum access example



Source: (AKYILDIZ et al., 2006)

According to its definition, CR has two main characteristics - cognitive capability and re-configurability (AKYILDIZ et al., 2006). The former refers to the ability to sense and pick information from the radio environment. Thus, spectrum opportunities, location of neighboring users, type of available systems/services and so on, can be identified to estimate the best spectrum band and the most appropriate operational parameters (e.g. frequency carrier, modulation scheme, transmission power, access technology) to be used by SU in its communication. The latter enables the radio to be dynamically programmed to the radio environment, by adjusting operating parameters for the transmission on the fly without any modifications to the hardware components. Thus, CR is able to adapt to changes in the radio environment and dynamically reconfigure transmission parameters, such as transmission power, encoding scheme, frequency carrier and so on.

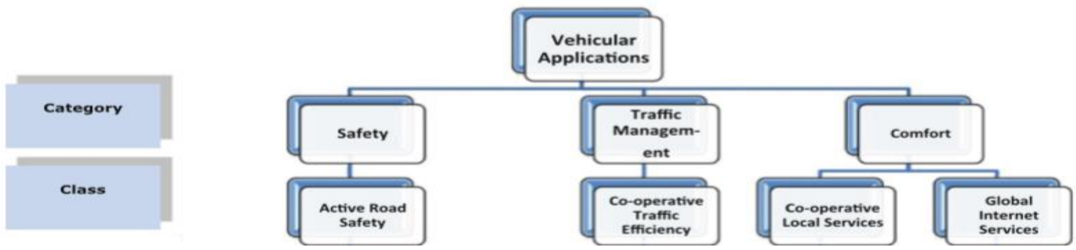
Cognitive radio spectrum sharing approaches fall into three classes according the knowledge needed to coexist with the primary network: Interweave, Underlay and Overlay (GOLDSMITH et al., 2009). The interweave approach enables joint spectrum use by PUs and

SUs in an opportunistic transmission process through sensing operations that should find the spectrum gaps or identify PU activity. The interweave category highly depends on spectrum sensing results but may also use geographical data to improve its effectiveness. Since spectrum sensing may be unreliable, the Overlay approach allows PUs to share spectrum knowledge with the cognitive users. Hence, the PUs may assist the SU transmission rather than competing for access. Differently from the two previous approaches, the underlay category admits simultaneous PU and SU transmissions as long as the interference level at the PU remains acceptable, not exceeding a predefined interference threshold. Spread spectrum (TSUI; CLARKSON, 1994) and Beamforming (VAN VEEN; BUCKLEY, 1988) techniques are some examples of how this category works so as not to degrade the PU's communication.

2.1.2 QoS Provisioning and Admission Control Mechanisms

Some types of networks naturally face QoS provisioning challenges such as in vehicular environments where highly unstable connection and diverse interference sources may undermine communication. In particular, these were created with the main purpose for security message passing, helping drivers to avoid accidents by reporting unusual road, weather or traffic conditions (NAJA, 2013). Nonetheless, this is not the only type of flow that these networks are expected to support (see Fig. 2.2). But because of the clear priority given to security data, researchers have been proposing QoS provisioning mechanisms for vehicular networks such as in the IEEE 802.11p standard (IEEE 1609, 2013) that may also be applied to other kinds of wireless networks, once considering that these should experience highly dense usage in a near future.

Fig. 2.2. Vehicular application classification

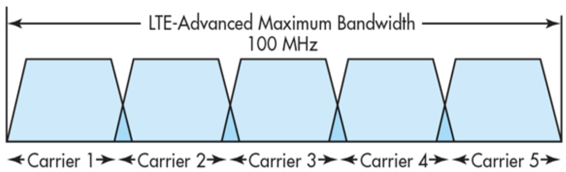


Source: (NAJA, 2013)

In a real scenario, the network traffic can be prioritized based on the QoS requirements (e.g., real-time traffic has higher priority than non real-time traffic). Besides that, we are motivated by the safety application provisioning in wireless vehicular networks, the natural user division caused by CR and the highly dense scenario in future wireless networks, which makes us believe that user/service differentiation will be an important issue to tackle. Offering different QoS guarantees, such as in the IEEE 802.11p standard for example, could enhance channel access. In that case, eight different priority levels are established and differentiated by specific values for a medium access control (MAC) parameter (NAJA, 2013).

If service prioritization is beneficial to higher priority level applications, it can also undermine lower priority ones. Channel aggregation may be a useful approach so as to increase the speed rates for low priority delay-tolerant applications. This mechanism has recently been incorporated to the LTE-A standard. In this case, it allows bonding up to five 20MHz channels totalizing a bandwidth of 100MHz as depicted in Fig. 2.3 (3GPP SPECIFICATION, 2013). Current deployment is still limited by contiguous aggregation but non-contiguous intra and inter-band aggregation or even channel aggregation and fragmentation (implies the previous characteristics) is also envisioned.

Fig. 2.3. Illustration of contiguous carrier aggregation based on the LTE-A standard



Source: (LEE et al., 2014)

2.2 Queueing Theory

This section addresses a review on discrete-event simulation, queueing theory examples and a brief paragraph out-lining the Continuous-Time Markov process.

2.2.1 Discrete-Event Simulation

A discrete-event simulator is able to model a system's operation such as the behavior of a centralized CRN based on a discrete sequence of events in time. Each event occurs at a particular instant and may cause a change of state in the system (ROBINSON, 2004). As opposed to real-time simulations, no change in the system is assumed to occur between consecutive events and the simulation can jump to the next event. The simulation must also keep track of the current simulation time; in whatever measurement units are suitable for the system being modeled.

In a simulation model, the performance metrics are not analytically derived from probability distributions, but rather as averages from different runs of that model. For this reason, it is important to correctly dimension the total number of runs that is necessary to guarantee reasonable statistical significance. In addition, the simulation designer must decide the ending condition for a “run”. Common examples are: at a specific time instant or after processing a pre-defined number of events.

Moreover, depending on the system to be modeled, a priority discipline can be adopted. There are prioritized queueing systems in which customers with higher priority are selected for service ahead of those with lower priorities, independent of their arrival time arrival into the system. In our system, we have considered the service discipline as being the typical FIFO, where the customers are serviced according to their arrival order. The arrival order is obtained after merging our independent event lists due to the afore-mentioned Poisson-exponential property.

2.3 Queueing Theory

This section addresses the most common queues that are frequently used to model real systems.

2.3.1 Concept, Characteristics and Notation

Queueing theory is derived from probability theory and its object of study is the phenomenon of waiting in queues. Although the term is often used to describe the queue’s mathematical behavior, it may also be applied to models based on systems where queues are not allowed to form (COOPER, 1981).

Queueing theory has been widely adopted to study communication features in the early days of telecommunication and, although the term cognitive radio was firstly introduced in 1999, only recently the literature has devoted attention to applied queueing theory regarding CR functionalities. According to our research, the first references concerning CR and queueing theory appeared in 2007 and since than, the number of publications has become larger, reaching up to 143 references by the end of 2013 (SUN et al., 2014). However, it is hard to count the recent numbers once the works that use both concepts may appear under many names. A general queueing system can be usually described in terms of the following characteristics:

- The arrival process: An arrival process is generally characterized by a distribution that described the amount of costumers that arrive in queue per unit time and the distribution that describes the times between successive customer arrivals (inter-arrival times) (GROSS et al., 2008). The most common Markovian arrival process is a Poisson process where the time between each arrival is exponentially distributed.

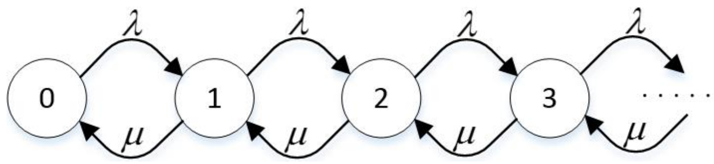
- The service time: describes the time a customer spends being served.
- The number of servers: indicates how many servers are considered in the system to attend the customers.
- The system capacity: this parameter specifies the maximum number of customers allowed to stay in the system. This includes customers that may be waiting for service (in a buffer) and those customers that are currently being served.
- Population size: is the total amount of customers that can enter the system. This parameter can be finite or infinite.
- The service discipline: this parameter defines the policy for service order. The most common disciplines are the First Come, First Served (FCFS), Last Come, First Served (LCFS) and the Static Priorities (SP). The latter selects customers to be served based on pre-defined priorities.
- The preemption discipline can be used in conjunction with the LCFS or Static Priorities. This discipline interrupts or preempts the customer currently being served if there is a higher priority customer in the queue (BOLCH et al, 2006).

Kendall's notation has been widely used to represent queueing systems. It uses the symbols A/S/m/N/K/SD to describe them, where A indicates the inter-arrival times distribution, S is the service time distribution, m is the number of servers, N is the system capacity, K is the population size and SD is the service discipline (COOPER, 1981). The letter M (for Markovian) is used to denote that the inter-arrival times and service times are exponentially distributed. A queueing system can also be represented in a shorter version considering that the system capacity is infinite, the population size is infinite or the service discipline is FIFO. The main queueing systems are briefly described in the sequence.

2.3.2 Types of Queues

- M/M/1 queue: Is a single-server queue, widely used to model systems where a single server provides the service to the customers. In this type of queue, the inter-arrival times and service times are exponentially distributed, there is no limitations toward either the population size or the system capacity and the adopted service discipline is FCFS (JAIN, 1991). The number of customers in the system denotes its state, and a state transition diagram is shown in Fig 2.5, where the two main parameters are the mean arrival rate of customers λ and the mean service rate μ. In this diagram, we observe that the arrival and service rates are fixed, regardless the number of users in the system. Furthermore, the quantity $\frac{\lambda}{\mu}$ is called traffic intensity, denoted as ρ (GROSS et al., 2008).

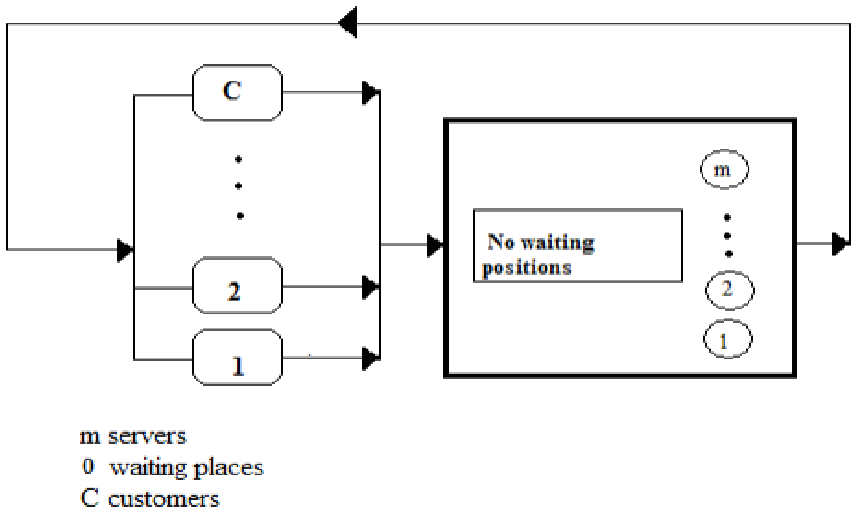
Fig. 2.4. State transition diagram for a M/M/1 queue



Source: (BALIEIRO, 2015)

- M/M/m queue: The M/M/m queue is a multi-server model where the arrival rate distribution is Poisson, the service times have exponential distributions, and there are *m* identical servers, where each one has the same service capacity. In this system, if there is at least one idle server, the arriving customer is serviced immediately. Otherwise, the customer may wait in a buffer to be served. The buffer is of infinite size, which implies that there is no limit on the number of customers it can handle (JAIN, 1991).
- M/M/m/N queue: This system is similar to the previous one (M/M/m), but it has a limited amount of users denoted by N, that is, the system capacity is limited (GROSS et al., 2008). If *m* and *N* have the same value, thus becoming M/M/N/N, it is assumed that the system has no buffer to hold blocked or interrupted users. Other sources apply a similar notation to indicate this special case (WHITT, 2012), for instance, the M/M/m/0 queue that is called the Erlang's loss system (see Fig. 2.6). They are characterized by having *m* identical servers, Poisson input, Exponential service times, no waiting positions $N = 0$ (no buffer) and unlimited number of customers. This means that after the system reaches full capacity, all new arrivals are blocked. In such cases, the authors use the effective arrival rate as the difference between the total and the blocked arrival rates.

Fig. 2.5. State transition diagram for a M/M/m/0 queue



2.3.3 Continuous-Time Markov Process

In probability theory, a continuous-time Markov chain or process (CTMC or CTMP) is a collection of variables generally indexed by time, which is a continuous quantity. It follows the Markovian property that future variable distribution is independent of the historical behavior, depending solely on its current state. Moreover, the stationary analysis for a CTMC gives the probability distribution to which the process converges for large values of time units. In brief, a CTMC may be a powerful tool for forecasting a system's stationary state probability. Most CTMCs properties follow directly from the results about its discrete version, the Poisson process and the exponential distribution. For in-depth view on the terminology and definitions on the matter, this author recommends (NORRIS, 1998).

2.4 Chapter Summary

This Chapter described the technical background in order to best guide the reader throughout this dissertation. First, we have reviewed the importance of the role played by cognitive radio in wireless network's future and its main concepts. Then, the main QoS provisioning mechanisms that will be detailed further in this work were briefly described and exemplified. Moreover, we have shown a technical review on queuing theory outlining the main variable distributions and queueing types.

Chapter 3 - Related Work

This Chapter provides a literature review in order to present the reader the current state-of-the-art of spectrum access models regarding CRN. It also highlights the main drawbacks and strengths of each investigated work and moreover, we propose a qualitative classification to clarify our contributions.

3.1 Introduction to Spectrum Access modeling through Continuous-Time Markov Chains

The CRN spectrum access is often studied in terms of analytical models that help to better understand different access strategies. In this work, we will be focused on the evaluation of CRNs modeled as loss systems, which will not include buffered queue approaches. A number of buffered queueing types can be used to model the same CRNs, but they will usually add a number of non-essential performance metrics (e.g. mean queue time, mean queue occupation). In this field of study, each author formulates their own access strategies with unique characteristics and different performance metrics that are described as follows:

- System Priority Levels – A CRN modeled as a loss system may assign priority to its users. In such case, a higher priority user may take the resources from a lower priority user, forcing them either to handoff to another idle set of resources or to quit service. The literature commonly refers to the PU as the highest priority user and the SU as the lowest, that is, characterizing a two-priority system. However, there can be as many priorities as one likes; recently, some authors have introduced three priority users, with the PU being the highest and the secondary network being divided into two layers with intermediate and low priorities. Slicing the secondary network is particularly useful for managing applications with different requirements (e.g., in the 802.11p standard).
- Secondary Traffic Type – Often, the primary user is considered homogeneous whereas the secondary traffic has been reproduced as heterogeneous traffic. Since delay-tolerant services such as file transferring and web browsing rely on the network's throughput, they should normally benefit from the amount of bandwidth resources that are made available, whereas delay-sensitive

applications (e.g. video streaming, VoIP) will not necessarily decrease their service times once they are able to assemble bandwidth resources. In brief, this division becomes interesting for better resource utilization, as the CRN will ideally provide the proper resource amounts to each application type.

- Primary and Secondary Resource Occupation – Until now, resource occupation has been modeled either having the minimum PU bandwidth as being always equal or smaller than the minimum required SU bandwidth, or the opposite, where the PU bandwidth is strictly greater than or equal to the SU minimum bandwidth.
- The resource occupation may vary from author to author; for example, an author may consider a PU to occupy a single channel unit or propose a flexible formulation that allows a PU to occupy any amount of channels. The same idea is applied to one or more SU kinds.
- Aggregation Techniques(s) – CA/CF are applied as aggregation techniques under different conditions and may be used statically, that is to say, they assign a fixed number of resources that a user can assemble, from the beginning of the experiment until the end. They can also be used dynamically; where the number of assembled channel resources may vary between a lower and an upper bound during operation.
- Performance Metrics – Loss systems are more commonly evaluated in terms of blocking probability and forced termination probability. But solely analyzing the two previous metrics might hide interesting details. Therefore, less commonly, some works also provide analysis on spectrum utilization and throughput.
    - Blocking probability: Probability that a user arrival will not be served.
    - Forced Termination (dropping) probability: Probability of service being interrupted (e.g., due to higher priority arrivals).
    - Spectrum utilization: The amount of occupied spectrum resources per unit time regarding individual user priority or joint utilization.
    - Throughput: The amount of completed services per unit time also regarding a particular user priority or group of user priorities.
- Dynamic Access Control (DAC) – The most common type of DAC is the addition of buffers to hold blocked/dropped users in the CRN. The main

benefit is associated to blocking and dropping rate reduction. But DAC may also be applied through channel reservation schemes that can act in single or multiple priority levels. The main benefit here is a considerable reduction in forced termination probability at the expense of having higher blocking rates. This strategy aims at regulating the number of available channels for a particular user. Once fewer channels are made available, higher blocking might be experienced, but because fewer users are also being served, there is a smaller probability that they might be interrupted.

Although buffering is commonly used in DAC, this dissertation will not address related works to that matter. The use of buffers belongs to another queueing class and it would be hard to compare results from such different formulations. Also, the system's transitions drastically change with buffer addition, and therefore, building a buffer compliant queueing system would take the focus of the loss system approach. Some example works using Markovian buffered queues as CRNs can be found in (CHU et al., 2013), (WU et al., 2013) and (KHEDUN; BASSOO, 2015).

3.2 Overview of spectrum access mechanisms for CRNs

A cognitive radio system is often divided in at least two user classes. But because one of its main envisioned applications is to be applied to the TV White Spaces (TVWS), where licensed transmissions generally use large bandwidths, many authors have modeled their CRN restricting the resource occupation to the case where the PU minimum bandwidth is strictly greater than or equal to the SU minimum bandwidth, as in works (ZHU et al., 2007), (JIAO et al., 2014), (CHU et al., 2014) and (CHU et al., 2015). An example can be found in Fig. 3.1, where the secondary bandwidth is three times smaller than the primary bandwidth. Other authors propose the opposite (see Fig. 3.2), that is, works that model the minimum PU bandwidth as being always equal or smaller than the minimum required SU bandwidth as in (JIAO et al., 2010), (JIAO et al., 2012) and (LI et al., 2012). Obviously, regarding this subject, each work has its own particularities, such as in (LI et al., 2012) where the PU bandwidth is statically equal to one unit channel. But, for the purpose of this work, this aspect can be summarized into those two groups. Furthermore, the afore-mentioned works will be briefly described and analyzed.

Fig. 3.1 Resource occupation example where the PU minimum bandwidth is strictly greater than the SU minimum bandwidth

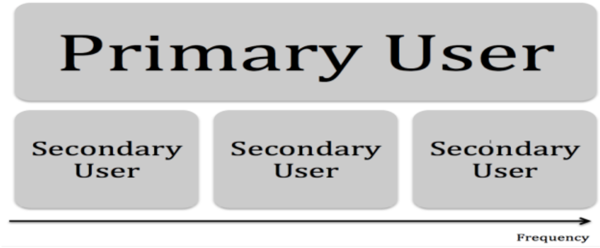
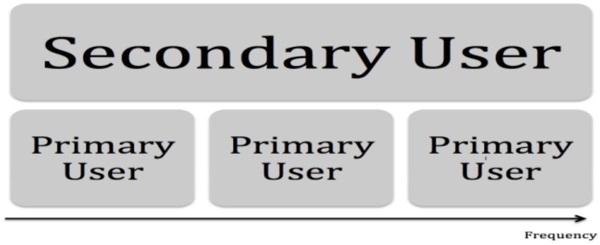


Fig. 3.2 Resource occupation example where the SU minimum bandwidth is strictly greater than the PU minimum bandwidth



1. The model developed in (ZHU et al., 2007) represents a homogeneous two-priority level CRN (primary and secondary network). The authors evaluated the system's throughput, other than blocking and dropping probabilities through analytical expressions. Besides, their CRN is compliant with a single-level channel reservation strategy in the lowest priority layer, i.e., the secondary network. In addition, they have proposed an iterative algorithm to find the channel reservation value that maximizes the secondary throughput.

    Comments on (1): The work in (ZHU et al., 2007) inspired many authors in the spectrum access field, including (CHU et al., 2014) and (CHU et al., 2015). It is clear that both nomenclature and ideas are similar, except that (ZHU et al., 2007) has not provided a simulation model to compare their analytical results. Although the work was considered a breakthrough at that time, different researchers have found incorrect expressions and conclusions, publishing their own findings in (MARTINEZ et al., 2010) and (AHMED et al., 2009). Furthermore, although the author has formulated both blocking and dropping metrics, their algorithm only takes the throughput into account, that is, there is no concern in limiting these other performance metrics, making the solution improper.

2. In (JIAO et al., 2010), the authors describe a Markovian-modeled two level priority CRN with static channel aggregation, similar to (ZHU et al., 2007), but without channel reservation or secondary handoff. They have analyzed the system's capacity, blocking and forced termination probabilities in different load situations.

   Comments on (2): Such work shows an extensively obvious analysis as they unnecessarily varied both arrival and service rates creating similar scenarios. They have evaluated the system's capacity (throughput), blocking and forced termination probabilities, using solely their analytic expressions, i.e., not presenting a simulated version for result comparison. They also give inappropriate conclusions as it states that channel aggregation does not increase the system capacity (throughput) and leads to higher blocking probability. Although they have used static aggregation, this statement may not be true for every scenario; the authors mistakenly lead the reader to believe that channel aggregation worsens blocking probability because the number of aggregated channels is increased in each test. Since their aggregation strategy is not flexible, logically, if the number of statically aggregated channels is too high, one secondary user will occupy considerable amounts of resources, producing higher blocking effects on new arrivals. However, this could not be true since it highly depends both in the PU load and the secondary service and arrival rates. Their results indicate that their testing scenarios were not properly designed for attending multiple network load situations.

3. In (JIAO et al., 2012) the authors extended their previous work (JIAO et al., 2010) modeling an heterogeneous CRN. Although there is traffic differentiation (e.g., delay sensitive and delay tolerant), they did not separate the two kinds of secondary traffic into different priorities, that is, one type of secondary user is not able to terminate the other, if there are not enough resources available. Instead, the secondary elastic user gives away extra resources, either until it accommodates a new secondary static user or it reaches a minimum amount of sub bands. They have applied channel

aggregation and fragmentation to part of their secondary network (delay tolerant traffic).

Comments on (3): The authors change the static channel aggregation strategy to a dynamic approach and provide a different conclusion compared to that on (2). They state that under appropriate system parameter configurations, channel aggregation can be beneficial. Interestingly, they use a two priority level approach to represent three types of traffic: static primary, static secondary and elastic secondary. Hence, this is a better piece of work than their last paper as more scenarios are investigated and more importantly, their model is validated through simulation, although we have found the nomenclature and transition disposal very hard to read, that is, some equations were compressed to fit in a smaller space.

4. The study (LI et al., 2012) focuses on the combined channel aggregation and fragmentation strategy for a two-priority level CRN with homogeneous secondary network. It compares this dynamic strategy to static channel aggregation through both simulated and analytical experiments. The benefit of the combined strategy is a two-fold: on one hand, aggregation provides high data rate and improves spectrum utilization. On the other hand, fragmentation is allowed such that multiple SUs may share a single channel (not considering guard band) when needed, hence, decreasing both blocking and dropping (forced termination) probabilities. Other than the three afore-mentioned metrics, they also provide a capacity analysis under the name: throughput, which is the number of completed services per unit time.

   Comments on (4): Differently from the other referenced works, (LI et al., 2012) shows a different nomenclature for the same problem, which for us was simpler to understand. Nonetheless; the results of (LI et al., 2012) were also too obvious as their flexible strategy will always be equal or more efficient than the static aggregation strategy used for performance comparison.

5. In (JIAO et al., 2014), both channel aggregation and fragmentation were investigated. However, differently from (4), the adopted aggregation strategy

allows SUs to use channel portions (under one channel unit). So they have explored the benefits of elastic traffic and derived a new formulation for the same performance metrics as in their previous work (capacity, blocking and forced termination probabilities), but separated the evaluation of real-time (static) traffic from file downloading (elastic) traffic in separate two-layered priority CRNs.

Comments on (5): This work addressed some of the previous issues highlighted in this section. Regarding their concluding remarks, the authors suggested that, as long as it meets the minimum QoS requirements, using smaller portions of channels makes the system more efficient. In fact, their performance metrics show better response but mostly because of the smaller QoS requirements that made more users to be served at the same time. Besides, it is important to note that high bandwidth granularity can be particularly difficult to deploy in real network.

6. The authors of (CHU et al., 2014) assume three priority levels in a strategy to coordinate different types of traffic (heterogeneous secondary system). Using a three-dimensional Markov chain, they have investigated the system's performance in terms of blocking and dropping probabilities for the secondary network. They have used static channel aggregation, under the nomenclature "sub band assembling".

   Comments (6): The work developed in (CHU et al., 2014) contains some similarities with our approach: three priority levels and an heterogeneous secondary network. However, we have noted some erroneous transitions in their proposed model, which may confuse the reader. These mistakes are described in appendix A.

7. In (CHU et al., 2015) the authors expanded their previous DSA formulation to become compliant with a channel reservation strategy, keeping the three-layered priority network and applying static channel aggregation in all levels. In addition, it was provided an iterative algorithm that finds the optimal

number of reserved sub bands for both SU types. Their results were also shown in terms of blocking and forced termination probability (dropping).

Comments on (7): The work developed by the author simply extended their previous paper (CHU et al., 2014), however, this one only presented the analytical results, i.e., this model was not properly validated with regard to channel reservation. Besides, the effects of channel reservation were not thoroughly investigated because only very low PU density scenarios were tested and no spectrum utilization or throughput metrics were provided. Hence, there was no concern on determining when channel reservation may in fact lead to better results and when it may undermine the system's performance.

8. This dissertation explores CRN modeling with three priority levels, two of them dedicated to the secondary network, similarly to the studies described in (6) and (7), but unlike (3) that had three user types and just only two priority levels. Because prioritization may cause starvation on the lower user levels, we provided aggregation approaches in order to boost their performance, adopting the strategy designed in (4). This approach enables better performance in all the four investigated metrics while not harassing higher layer's performances. Among the aggregation options, one can choose between the combined aggregation and fragmentation approach that uses an integer and fractionary number of assembled channels or just the integer number of aggregated channel units, i.e., excluding CF but keeping CA, for the lowest priority traffic. In our model, it is also possible to reduce the amount of resources to a specific secondary user type by means of multi-level channel reservation. In our formulation, there are more reservation options than in (7)

Comments on (8): Noticing that the literature lacked a formulation that provided both resource occupation options simultaneously, we have designed our model to allow such behavior, with an especial restriction: the chosen bandwidth values for each user should be multiples between them and the chosen amount of channels*. Furthermore, our CRN modeled to be compliant with three user layers, each of which having priority values, labeled as high (for PU), intermediate (for delay sensitive SUs) and low (for delay-sensitive or delay-tolerant SUs). In addition we have provided channel reservation for both secondary network layers and three different channel aggregations strategies for the lowest user layer. The system was evaluated in terms of four performance metrics: Blocking probability, forced termination probability, throughput and spectrum utilization through analytical and simulated results, differently from the majority of the previous studies. Besides technicalities, we have formulated both transitions and performance metrics in a comprehensive manner, providing examples and illustrations of our model.

We have outlined the main advantages and disadvantages of the previously described literature in Table 3.1.

Table 3.1. Drawbacks and strengths of each investigated work

Paper	Drawbacks	Strengths
1. '(ZHU et al., 2007)'	Formulation mistakes. No simulation analysis. Has formulated both blocking and dropping metrics, but their optimization algorithm only takes the throughput into account.	Evaluated three performance metrics. Uses channel reservation for achieving better forced termination probabilities.
2. '(JIAO et al., 2010)'	No new ideas, very similar to (ZHU et al., 2007) but without channel reservation. Obvious results. No simulation analysis. Too many similar testing scenarios. Limited conclusions toward channel aggregation.	Evaluated three performance metrics. Correct formulation.
3. '(JIAO et al., 2012)'	Very similar to old formulation. Non-intuitive formulation. The nomenclature and transition disposal are hard to read.	Applied dynamic channel aggregation. Used simulation analysis.
4. '(LI et al., 2012)'	Obvious conclusions. The analysis and formulation are restricted to the case were each PU occupies only one channel.	Dynamic channel aggregation and fragmentation formulated analytically. Evaluated four performance metrics. Intuitive formulation. Used simulation analysis.
5. '(JIAO et al., 2014)'	Non-intuitive formulation. The nomenclature and transition disposal are hard to read. Limited concluding remarks.	Proposes to study the influence of channel aggregation and fragmentation on three performance metrics. Used simulation analysis. Interesting testing scenarios.
6. '(CHU et al., 2014)	Formulation mistakes. Evaluated only two performance metrics. Utilization and Throughput were not investigated. Static CA was the only aggregation technique presented.	The nomenclature is partially understandable. Used simulation analysis.

7. ‘(CHU et al., 2015)’	Formulation mistakes. Evaluated two performance metrics. Utilization and Throughput were not investigated. No simulation analysis. Static CA was the only aggregation technique presented. Limited results: tradeoff between blocking and forced termination probability under low PU loads. Same optimization approach from ZHU et al. 2007. Uses the same amount of reservation channels in both secondary priority levels and it does not take into account higher SU forced termination probability.	The nomenclature is partially understandable. Uses channel reservation on the secondary network.
8. Present Dissertation	Formulation is limited to the case where the user’s bandwidths have to be multiples between each other and the total channel number. Channel aggregation is only possible in the lowest user priority.	Intuitive Formulation. Simulation analysis. Evaluated four performance metrics. Accepts both resource occupation options. Accepts channel reservation in two secondary levels.

3.3 Classification of the Related Works

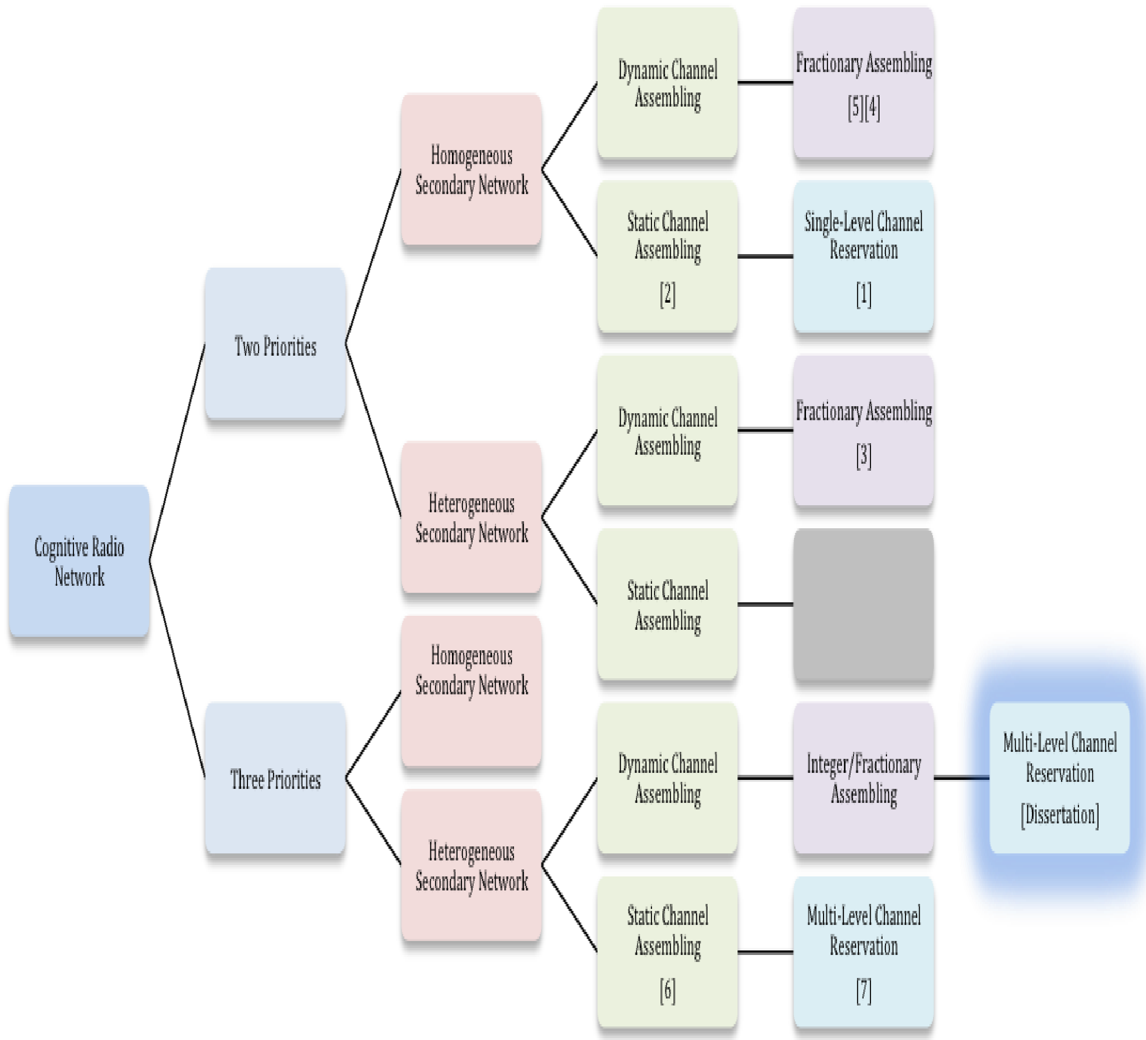
We have assessed the previous mentioned works in terms of ‘System Priority Levels’, ‘Secondary Traffic Type’, ‘Resource Occupation’, ‘Aggregation Technique(s)’, ‘Dynamic Access Control’ and ‘Evaluation Metrics’. These characteristics were summarized in Table 3.2 and part of them was used to illustrate a CRN classification diagram (Fig. 3.2). Each diagram node inherits the feature of its parent, in such a way that the work (LI et al. 2012), for example, numbered as ‘4’ presents a two priority system, a homogeneous secondary network and a dynamic aggregation and fragmentation approach towards the SUs. There is no dynamic access control implied in their formulation.

Table 3.2. Characteristics of the investigated works

Paper	System Priority Levels	Secondary Traffic Type	PU/SU* Resource Occupation	Aggregation Technique(s)	Dynamic Access Control	Evaluation Metrics
1. ‘(ZHU et al., 2007)’	Two	Homogeneous	Minimum PU bandwidth ≥ Minimum SU bandwidth	Static CA	Single-Level Channel Reservation	1) Blocking probability 2) Forced termination probability 3) Throughput
2. ‘(JIAO et al., 2010)’	Two	Homogeneous	Minimum SU bandwidth ≥ Minimum PU bandwidth	Static CA	No Channel Reservation	1) Blocking probability 2) Forced termination probability 3) Capacity

3. ‘(JIAO et al., 2012)’	Two	Heterogeneous	Minimum SU bandwidth ≥ Minimum PU bandwidth	Dynamic CA+CF	No Channel Reservation	1) Blocking probability 2) Forced termination probability 3) Capacity
4. ‘(LI et al. 2012)‘	Two	Homogeneous	Minimum SU bandwidth ≥ Minimum PU bandwidth	Dynamic CA+CF	No Channel Reservation	1) Blocking probability 2) Forced termination probability 3) Throughput 4) Spectrum Utilization
5. ‘(JIAO et al., 2014)’	Two	Homogeneous	Minimum PU bandwidth ≥ Minimum SU bandwidth	Dynamic CA+CF	No Channel Reservation	1) Blocking probability 2) Forced termination probability 3) Capacity
6. ‘(CHU et al., 2014)’	Three	Heterogeneous	Minimum PU bandwidth ≥ Minimum SU bandwidth	Static CA	No Channel Reservation	1) Blocking probability 2) Forced termination probability
7. ‘(CHU et al., 2015)’	Three	Heterogeneous	Minimum PU bandwidth ≥ Minimum SU bandwidth	Static CA	Multi-Level Channel Reservation	1) Blocking probability 2) Forced termination probability
8. Present Thesis	Three	Heterogeneous	Minimum PU bandwidth ≥ Minimum SU bandwidth or Minimum SU bandwidth ≥ Minimum PU bandwidth*	Dynamic CA+CF	Multi-Level Channel Reservation	1) Blocking probability 2) Dropping probability 3) Throughput 4) Spectrum Utilization

Fig. 3.3. Classification of the related works



3.4 Chapter Summary

In this chapter we have studied some characteristics concerning previous CRN models: ‘System Priority Levels’, ‘Secondary Traffic Type’, ‘Resource Occupation’, ‘Aggregation Technique(s)’, ‘Dynamic Access Control’ and ‘Evaluation Metrics’. We have also highlighted their main strengths and shortcomings, by performing a qualitative comparison in order to show where our proposal should be placed and to clarify its contributions and besides, taking into account some of those characteristics, we have proposed a taxonomic classification.

Chapter 4 - System Model

Our system was built to cope with the three-user priority levels, with the PU being the user with the highest priority, the SU_1 as the intermediate priority level and the SU_2 as the lowest user in the hierarchy. The secondary network has two different types of traffic, one that works with static bandwidth (SU_1) and therefore used for real-time applications and another that may dynamically assemble a number of channel resources (SU_2), that is, delay-tolerant applications. We have also modeled our CRN to allow a dynamic access control mechanism called channel reservation, in both levels of the secondary network. Thus, the model will specify two variables (*R1* and R_2) that will control the total amount of channel resources that are available for each secondary level.

4.1 Model Assumptions

The CRN addressed by our model uses a centralized overlay approach that utilizes a common control channel to map resources status along operation. Similar to the majority of the studies in the field, this work does not take into account the time overhead imposed by the spectrum sensing delay and the system's collision delay as they are too small compared to the transmission duration. For the latter, this means that not time is spent if a higher priority user drops a lower priority one, taking control of its resources. In our CRN, a total number of N channels are shared by PUs, SU_1s and SU_2s and each channel is assumed to have one unit bandwidth. In this work, bandwidth will not be expressed in Hertz but in channel units (e.g., one channel, two and a half channels, etc.). Our model allows a PU to statically occupy an integer amount of channels that is smaller or equal to N, i.e., $1 \leq B_{PU} \leq N$. Similarly, the SU_1s may statically occupy $1 \leq B_{SU1} \leq N$ channels (real-time applications). On the other hand, the SU_2 will be allowed to dynamically assemble channels, as long as the number of assembled resources is kept between a lower and an upper bound $[B_{SU2}^{min}, B_{SU2}^{max}]$, with $1 \leq B_{SU2}^{min} \leq B_{SU2}^{max} \leq N$ (delay-tolerant applications).

We have adopted a Continuous Time Markov Chain (CTMC), in which the user arrivals are assumed to be independent Poisson processes with arrival rates λ_{PU}, λ_{SU1}, λ_{SU2} for the PU, SU_1 and SU_2 respectively. The service rates for these users are exponentially distributed with service rates μ_{PU}, μ_{SU2} and μ_{SU2} for one unit channel. Naturally, for any described user, if M channels are assembled, the service rate becomes $M* \mu$.

4.2 DCAF, DCA and FCA Strategies

In this section, we will present the three aggregation approaches to be evaluated in the lowest layer users (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. It is implied that the PU has the highest priority, thus, if it arrives in a channel that is occupied by an SU_1, this will either handoff or be terminated. The ESA will be executed only for adjusting the SU_2 bandwidth, but for every arrival or service completion event detailed in the following:

- SU_2 arrival event: When a SU_2 arrives, it will try to occupy maximum bandwidth B_{SU2}^{max}, otherwise, the available channels will be equally shared by the ongoing SU_2s and by the newcomer, as long as every user gets at least B_{SU2}^{min} resources. If not, then the new SU_2 will be blocked and thus will not be served.
- PU/SU_1 arrival event: 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 case, these SU_2s will either need to handoff to another idle channel, or share the available spectrum by ESA, ensuring that every SU_2's bandwidth is kept between $[B_{SU2}^{min}, B_{SU2}^{max}]$, otherwise, the target SU_2(s) will be forcibly terminated.
- PU/SU_1/SU_2 service completion events: 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 not violated.

The equal sharing process has already been used in the literature of resource allocation in wireless communication (JENG, LIN, 1999) (HEREDIA-URETA et al., 2003) but, to the best of our knowledge, has only recently been introduced in CRNs by (LI et al., 2012) together with the channel aggregation and fragmentation strategy (CAF). However, their CRN model was much simpler that ours, as they considered only two priority levels with the PU occupying only one channel and no dynamic access control. Thus, because more recent and complex works (CHU et al., 2014) (JIAO et al., 2014) avoided CAF, we have incorporated it in one of our scenarios under the name dynamic channel aggregation and fragmentation (DCAF), which seems more proper, due to the fact that many of these works utilize CA and/or CF statically.

In DCAF, the SU_2's bandwidth can dynamically be adjusted according to the system load (see Fig. 4.1.). In DCAF, CA and CF are performed adaptively and the number o assembled resources may be a real positive number between B_{SU2}^{min} and B_{SU2}^{max}.

Even though DCAF and CAF are essentially the same strategy applied in different CRNs, the DCA is a derivation from DCAF that, as far as we can tell, has not been explored in the literature yet. In this strategy, the SU_2's bandwidth can also be dynamically adjusted according to availability. But, in DCA, only adaptive CA is allowed; hence, the number of assembled resources may be an integer positive number between B_{SU2}^{min} and B_{SU2}^{max} (see Fig. 4.2.).

The acronym FCA stands for fixed channel aggregation and is the most common strategy explored in the literature. The authors frequently use it to compare with a CAF derivative or as the main aggregation strategy itself, such as in (CHU et al., 2014). Here, a fixed integer number of channels are assembled during operation, if there are available resources (see Fig. 4.3.).

Fig. 4.1. DCAF example in a CRN with four channels

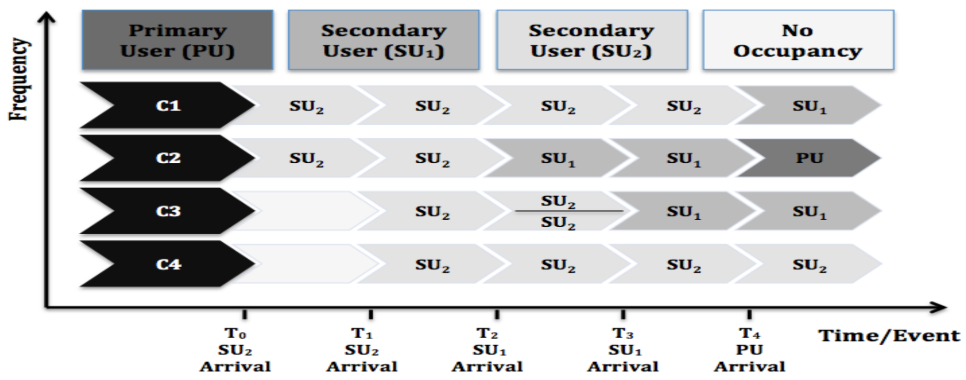


Fig. 4.2. DCA example in a CRN with four channels.

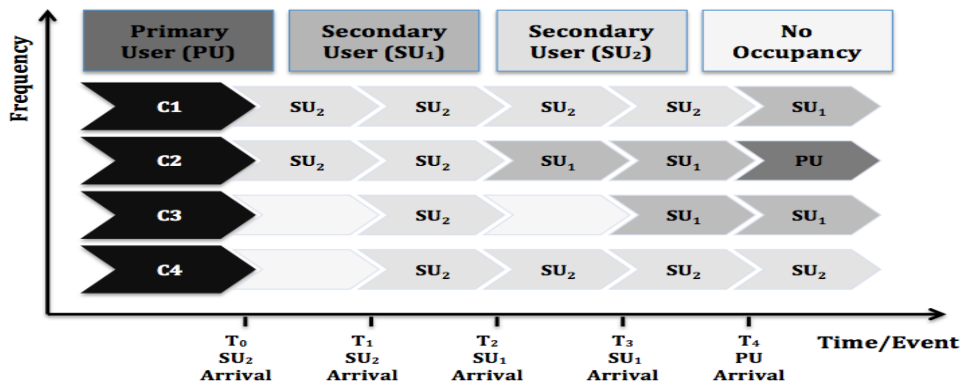
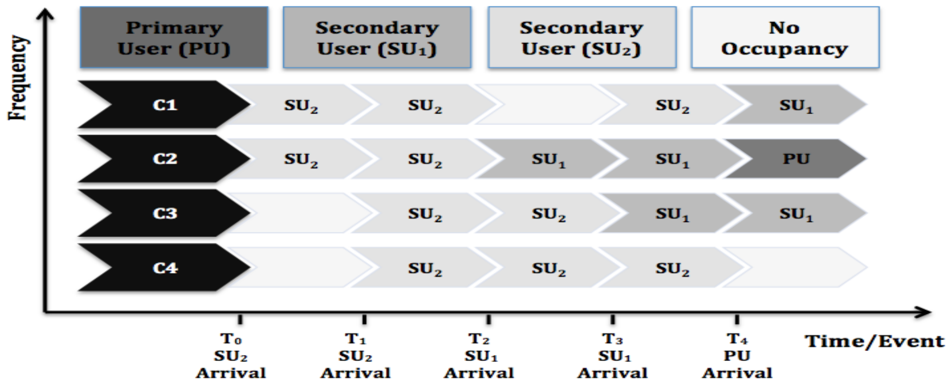


Fig. 4.3. FCA example in a CRN with four channels.



4.3 Analytical Model

We propose a DSA approach as a Markov chain with three-state variables i, j, k (a graphical model example will be shown further in this document), where i is the integer number of PUs, j is the integer number of SU_1s and k is the integer number of SU_2s in the system. The feasible state space (Ω) is generated according to: $\Omega = \{(i,j,k) | 0 \leq i \leq, \left\lfloor \frac{N}{B_{PU}} \right\rfloor, 0 \leq j \leq \left\lfloor \frac{N-R_1}{B_{SU1}} \right\rfloor, 0 \leq k \leq \left\lfloor \frac{N-R_2}{B_{SU2}^{min}} \right\rfloor$, with $(i * B_{PU}) + (j * B_{SU1}) + (k * B_{SU2}^{min}) \leq N\}$.

The SU_2's bandwidth can be calculated in different forms, depending on the adopted aggregation strategy by one of the following:

In DCAF, the bandwidth and service rate of each SU_2 in the state space are expressed by (Eq.1) and (Eq. 2) respectively:

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

(1)

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

(2)

In DCA, the bandwidth and service rate of each SU_2 in the state space are expressed by a variation of (Eq. 1) and (Eq. 2), respectively (Eq. 3) and (Eq. 4):

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

(3)

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

(4)

To avoid dynamicity in channel aggregation, it is possible to obtain FCA from either DCAF or DCA just by setting equal values to $B_{SU2}^{min}\ and\ B_{SU2}^{max}$. Thus, $B_{SU2,FCA}\,(i,j,k) = B_{SU2}^{min} = B_{SU2}^{max}$ and $\mu_{SU2,FCA} = B_{SU2,FCA}\,(i,j,k) * \mu_{SU2}$.

4.4 Analytical Model Transitions

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 user requests can be further divided into normal and dropping transitions, where the first stands for those arrivals that do not imply in user interruptions while the latter will necessarily involve user interruption.

Considering that the transitions $\gamma_{(i,j,k)}^{(i\prime,j\prime,k\prime)}$ occur from one feasible state (i,j,k) to another (i',j',k'), and that any afore-mentioned aggregation techniques may be applied for the SU$_2$ by replacing B_{SU2*} and μ_{SU2*} by the desired approach's bandwidth and service rate equivalent.

A. Transitions from state (i,j,k) to other states.

Consider the number of idle resources in state (i,j,k) to be: $idle = N - (\mathrm{i} * B_{PU}) - (\mathrm{j} * B_{SU1}) - \left(\mathrm{k} * B_{SU2}^{min}\right)$, then the transitions are the following:

1) Primary user requests service: Three situations may occur depending on the CRN's occupation (5, 6 and 7).
   a) If the number of idle resources is greater than or equal to the PU bandwidth, i.e., $B_{PU} \le idle$, the arrival PU will be assigned B_{PU} channels without any other user being forced to terminate.

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

(5)

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., $idle < B_{PU} \leq idle + \left(\mathrm{k} * B_{SU2}^{min}\right)$, then the arrival PU will be assigned B_{PU} channels and $z = \left\lceil(-idle + B_{PU})/B_{SU2}^{min}\right\rceil$ SU_2s will be terminated.

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

(6)

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

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

(7)

2) Class one secondary user requests service: Two situations may occur depending on the CRN's channel occupation (8 and 9).

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., $B_{SU1} + R_1 \leq$ idle, the arrival SU_1 will be assigned B_{SU1} channels without any SU_2 being forced to terminate.

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

(8)

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., $idle < B_{SU1} + R_1 \leq idle + \left(k * B_{SU2}^{min}\right)$, then the SU_1 arrival is assigned B_{1-SU} channels. Thus, $z = \left\lceil(-idle + B_{SU1})/B_{SU2}^{min}\right\rceil$ SU_2s are terminated.

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

(9)

3) Class two secondary user requests service (10): 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., $B_{SU2}^{min} + R_2 \leq idle$, the arrival SU_2 will be assigned $B_{SU2*}(i,j,k+1)$ channels where the output for $B_{SU2*}(i,j,k+1)$ should be a value in between of the pre-defined lower and upper bounds or, if these values are equal, the algorithm should output the same value as the lower and upper bound values.

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

(10)

4) Primary user completes service (11), implies $i > 0$.

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

(11)

5) Class one secondary user completes service (12), implies $j > 0$.

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

(12)

6) Class two secondary user completes service (13), implies $k > 0$.

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

(13)

Table 4.1. Transitions from state *i*,*j*,*k* to other states

User	Type	Transition	Value	Eq. nº
PU	Normal	$\gamma_{(i,j,k)}^{(i+1,j,k)}$	λ_{PU}	5
PU	Dropping	$\gamma_{(i,j,k)}^{(i+1,j,k-z)}$	λ_{PU}	6
PU	Dropping	$\gamma_{(i,j,k)}^{(i+1,j-y,0)}$	λ_{PU}	7
SU_1	Normal	$\gamma_{(i,j,k)}^{(i,j+1,k)}$	λ_{SU1}	8
SU_1	Dropping	$\gamma_{(i,j,k)}^{(i,j+1,k-z)}$	λ_{SU1}	9
SU_2	Normal	$\gamma_{(i,j,k)}^{(i,j,k+1)}$	λ_{SU2}	10

PU	Normal	$\gamma_{(i,j,k)}^{(i-1,j,k)}$	$i * \mu_{PU}$	11
SU_1	Normal	$\gamma_{(i,j,k)}^{(i,j-1,k)}$	$j * \mu_{SU1}$	12
SU_2	Normal	$\gamma_{(i,j,k)}^{(i,j,k-1)}$	$k * \mu_{SU2_*}(i,j,k)$	13

Source: The author

B. Transitions from other states to State (i,j,k).

In this subsection, consider the number of free resources to be a function: $free\ (a,b,c) = N - (\mathrm{a} * B_{PU}) - (\mathrm{b} * B_{SU1}) - \left(\mathrm{c} * B_{SU2}^{min}\right)$, then the transitions are the following:

1) Primary user requests service: Three situations may occur depending on the CRN's occupation (14 - 16).

a) If the number of idle resources is greater than or equal to the PU bandwidth, i.e., $B_{PU} \leq free\ (i-1,j,k)$, the arrival 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).*

$$\gamma_{(i-1,j,k)}^{(i,j,k)} = \lambda_{PU} \tag{14}$$

b) If the number of free resources is less than the PU's bandwidth, i.e., $B_{PU} >$ *free (i-1, j, k+z')*, *k > 0*, then two situations might occur (15):

i. $If\ B_{PU} \leq B_{SU2}^{min}$, then only one SU_2 will be dropped, i.e., *z'=1.* Because the PU's bandwidth is less than the SU_2's, only one SU_2 will be dropped, that is, vacating only one SU_2 will free enough resources for accommodating a PU.

ii. If $B_{PU} > B_{SU2}^{min}$, then up to $z' = B_{PU}/B_{SU2}^{min}$ SU_2 will be removed.

$$\gamma_{(i-1,j,k+z\prime)}^{(i,j,k)} = \lambda_{PU} \tag{15}$$

c) If the number of free resources is less than the PU's bandwidth, i.e., $B_{PU} >$ *free (i-1, j, k+z')* and *k = 0*, then four situations might occur (16):

i. If $B_{PU} < B_{SU2}^{min}$ and $If\ B_{PU} \geq B_{SU1}$

1. $k + z' > 0$, then only one SU_2 will be dropped, i.e., $z'=1$.
2. $k + z'= 0$, then there are no SU_2s to be removed. Thus, according to our pre-defined priority, if there are SU_1s in the system, up to $y' = B_{PU}/B_{SU1}$ SU_1 will be removed.

ii. $If\ B_{PU} < B_{SU2}^{min}$ and $If\ B_{PU} < B_{SU1}$

1. $k + z' > 0$, then only one SU_2 will be dropped, i.e., $z'=1$ *and* $y'=0$.
2. $k + z'= 0$, then only one SU_1 will be dropped, i.e., $z'=0$ *and* $y'=1$.

iii. If $B_{PU} \geq B_{SU2}^{min}\ If\ B_{PU} \geq B_{SU1}$

y' may vary from zero up to B_{PU}/B_{SU1}, whereas z' varies from zero to B_{PU}/B_{SU2}^{min}, as long as, $y' * B_{SU1} + z' * B_{SU2}^{min}$ is equal to $B_{PU} - free(i-1, j+y', k+z')$.

iv. $If\ B_{PU} \geq B_{SU2}^{min}\ If\ B_{PU} < B_{SU1}$

1. $If\ B_{PU} \leq free(i-1, j+y', k+z') + (k+z') * B_{SU2}^{min}$, then z' varies from zero to B_{PU}/B_{SU2}^{min}, as long as, $z' * B_{SU2}^{min}$ is equal to $B_{PU} - free(i-1, j+y', k+z')$ and $y'= 0$.
2. $If\ B_{PU} > free(i-1, j+y', k+z') + (k+z') * B_{SU2}^{min}$, then not enough SU_2s can be dropped in order to free resources for the arriving PU, so z' equals k and $y' = 1$.

$$\gamma_{(i-1,j+y',k+z')}^{(i,j,k)} = \lambda_{PU} \tag{16}$$

2) Class one secondary user requests a service: Two situations may occur depending on the CRN's channel occupation (17 - 18).

a) If the number of free resources is greater than or equal to the sum of the SU_1's bandwidth and their number of reserved channels, i.e., $B_{SU1} + R_1 \leq free\ (i, j-1, k)$, the arrival SU_1 will be assigned B_{SU1} channels without any SU_2 being forced to terminate.

$$\gamma_{(i,j-1,k)}^{(i,j,k)} = \lambda_{1-SU} \tag{17}$$

b) If there are no free resources for a SU_1 arrival but, by dropping SU_2s, enough space can be made available, i.e., $free(i-1, j+y', k+z') + (k+z') * B_{SU2}^{min} \geq B_{SU1} + R_1 > free\ (i, j-1, k+z')$, then there might be more than one situation from which the system changes to state *(i, j, k).*

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

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

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

(18)

3) Class two secondary user requests a service (19): 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., $B_{SU2}^{min} + R_2 \leq free\ (i, j, k-1)$, the arrival SU_2 will be assigned $B_{SU2*}(i, j, k+1)$ channels.

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

(19)

4) Primary user completes service (20).

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

(20)

5) Class one secondary user completes service (21).

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

(21)

6) Class two secondary user completes service (22).

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

(22)

Table 4.2. Transitions from other states to state *i,j,k*

User	**Type**	**Transition**	**Value**	**Eq. nº**
PU	Normal	$\gamma_{(i-1,j,k)}^{(i,j,k)}$	λ_{PU}	14
PU	Dropping	$\gamma_{(i-1,j,k+z')}^{(i,j,k)}$	λ_{PU}	15
PU	Dropping	$\gamma_{(i-1,j+y',k+z')}^{(i,j,k)}$	λ_{PU}	16

SU1	Normal	$\gamma_{(i,j-1,k)}^{(i,j,k)}$	λ_{SU1}	17
SU1	Dropping	$\gamma_{(i,j-1,k+z\prime)}^{(i,j,k)}$	λ_{SU1}	18
SU2	Normal	$\gamma_{(i,j,k-1)}^{(i,j,k)}$	λ_{SU2}	19
PU	Normal	$\gamma_{(i+1,j,k)}^{(i,j,k)}$	$(i+1)*\mu_{PU}$	20
SU1	Normal	$\gamma_{(i,j+1,k)}^{(i,j,k)}$	$(j+1)*\mu_{SU1}$	21
SU2	Normal	$\gamma_{(i,j,k+1)}^{(i,j,k)}$	$(k+1)*\mu_{SU2*}(i,j,k)$	22

Source: The author

C. Blocking Situations.

With regard to loss systems, blocking transitions are not valid. This subsection depicts them for calculating the blocking probability further on, using a similar notation to those in subsections A and B. Hence, the reader should note that the following blocking situations (23 and 24) should not be implemented, differently from the aforementioned normal and dropping transitions.

1) A SU_1 blocking event happens when there are insufficient spectrum resources upon a SU_1 arrival, which is represented by equation (23). Depending on the number of reserved channels for this layer (R_1), there may be free channels upon a SU_1 arrival, but these will not be able to use it.

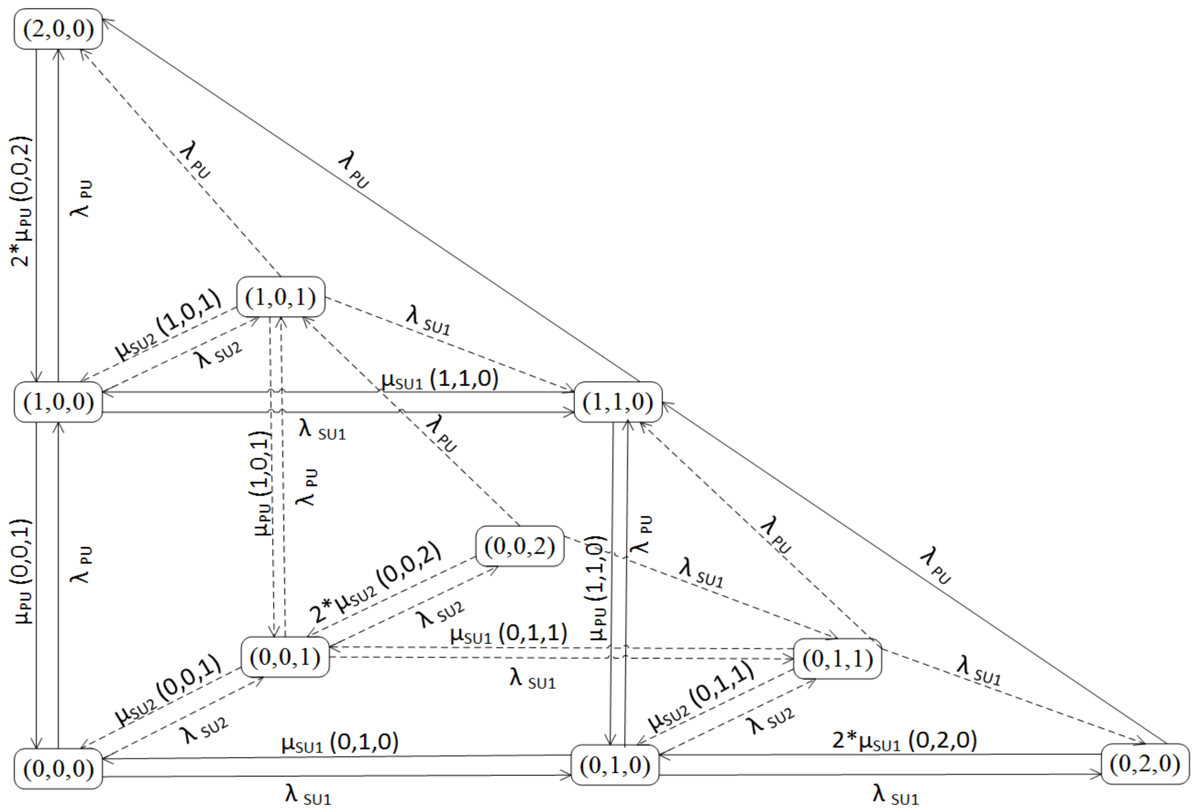
$$\gamma_{b1}{}_{(i,j,k)}^{(i,j,k)} = \lambda_{SU1}, if \ \ (i * B_{PU}) + ((j+1) * B_{SU1}) > N - R_1. \tag{23}$$

2) A SU_2 blocking event happens when there are insufficient spectrum resources upon a SU_2 arrival, even after the equal sharing process, represented by equation (24). Depending on the number of reserved channels for this layer (R_2), there may be free channels upon a SU_2 arrival, but these will not be able to use it.

$$\gamma_{b2}{}_{(i,j,k)}^{(i,j,k)} = \lambda_{SU2}, if \ \ (i * B_{PU}) + (j * B_{SU1}) + ((k+1) * B_{SU2}^{min}) > N - R_2. \tag{24}$$

Note that all states involved above must be feasible; otherwise, the corresponding transition rate is set to zero. An example of the state transition diagram when $N = 2$, $B_{PU} = 1$, $B_{SU1} = 1$, $B_{SU2}^{min} = 1$, $R_1 = 0$ *and* $R_2 = 0$ is depicted in Fig. 4.4. In this example, B_{SU2}^{max} can assume the values one or two and the diagram remains the same for any dynamic aggregation strategy. If FCA is chosen, then B_{SU2}^{max} will necessarily assume the value of B_{SU2}^{min} that is equal to one bandwidth unit. In addition, no channel reservation is assumed, hence, R_1 *and* R_2 variables are set to zero. A detailed study on the total number of states using this example is shown in Appendix C.

Fig. 4.4 FCA example in a CRN with no channel reservation.



The same example can also be used for representing situations when channel reservation is taken into account, that is, setting positive integers for the variables R_1 and R_2. Fig. 4.5, for instance, admits R_1 = *1 and* R_2 = *0*, while Fig. 4.6 takes the opposite R_1 = *0 and* R_2 = *1;* finally, Fig. 4.7 provides a diagram in which channel reservation is used in both secondary levels, i.e. R_1 = *1 and* R_2 = *1* simultaneously. As the values for each variable increases, fewer channel are made available for the specified user layer. For instance, in Fig. 4.5 note that the horizontal axis's state (0,2,0) is suppressed due to the value set to $R_1 = 1$, which implies $N - R_1$ resources for the SU1. This feature allows to individually diminishing the number of accepted users in the network, being an important mechanism for performance control.

Fig. 4.5. FCA example in a CRN with channel reservation applied to SU_1.

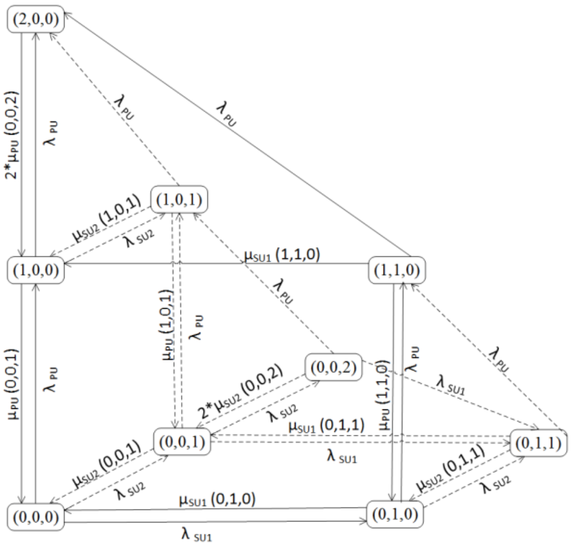


Fig. 4.6. FCA example in a CRN with channel reservation applied to SU_2.

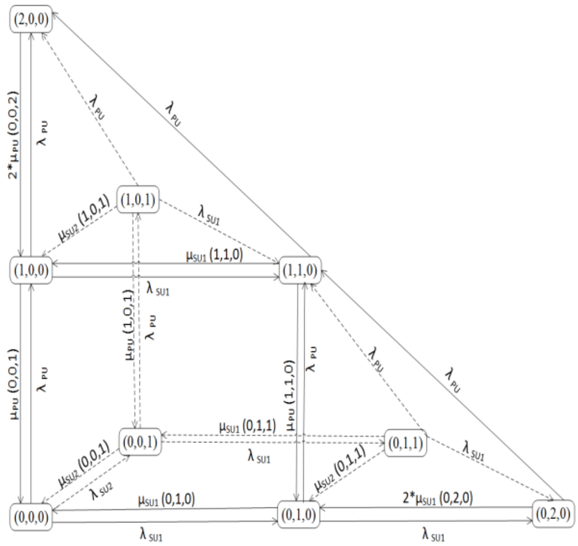
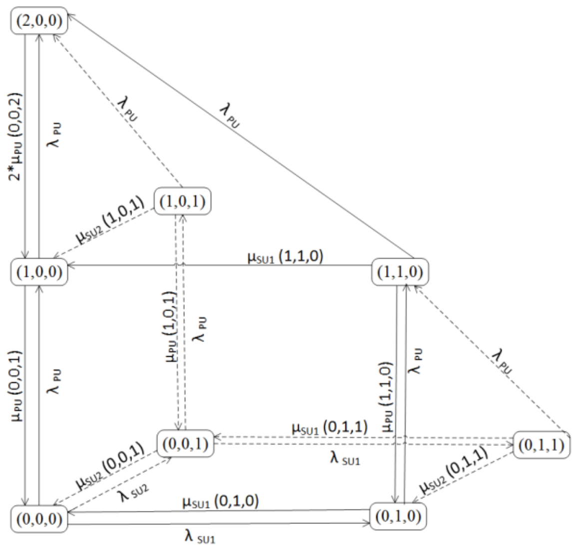


Fig. 4.7. FCA example in a CRN with channel reservation applied to SU_1 and SU_2.



4.5 Steady State Distribution

Let $\pi(i,j,k)$ be the steady-state probability of state *(i,j,k)* and $I\,(i,j,k) = \begin{cases} 1, if\ (i,j,k) \in \Omega; \\ 0, otherwise. \end{cases}$ an indication function for correcting possible unfeasible states. In order to obtain the generator matrix (*Q*), we must consider the conservation law in the queueing network, that is, the flow out of a state is equal to the flow into that state. Particularly in our CRN, the balance equations are described by (25).

$$\sum_{i\prime=0}^{\left\lfloor\frac{N}{B_{PU}}\right\rfloor}\sum_{j\prime=0}^{\left\lfloor\frac{N-R_1}{B_{SU1}}\right\rfloor}\sum_{k\prime=0}^{\left\lfloor\frac{N-R_2}{B_{SU2}^{min}}\right\rfloor}\pi(i,j,k)*\gamma_{(i,j,k)}^{(i',j',k')}*I(i,j,k)*I(i',j',k') =$$

$$\sum_{i\prime=0}^{\left\lfloor\frac{N}{B_{PU}}\right\rfloor}\sum_{j\prime=0}^{\left\lfloor\frac{N-R_1}{B_{SU1}}\right\rfloor}\sum_{k\prime=0}^{\left\lfloor\frac{N-R_2}{B_{SU2}^{min}}\right\rfloor}\pi(i',j',k')*\gamma_{(i',j',k')}^{(i,j,k)}*I(i,j,k)*I(i',j',k')$$

$$with\ 0\le i\le,\left\lfloor\frac{N}{B_{PU}}\right\rfloor,0\le j\le\left\lfloor\frac{N-R_1}{B_{SU1}}\right\rfloor,0\le k\le\left\lfloor\frac{N-R_2}{B_{SU2}^{min}}\right\rfloor\ and$$

$$(i',j',k')\neq\ (i,j,k).$$

(25)

Further, we know that the sum of each state probability in π equals to one. So, the normalized condition is expressed as:

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

(26)

Consider *n* to be the total number of states, i.e., $n = |\Omega|$. In order to obtain the steady-state probabilities it is necessary to derive a linear equation system, which includes *n* balance equations and the normalized expression. In this way, we have used an infinitesimal generator matrix method whose elements are the transition rates between different states. Then, by putting (25) and (26) into a compact matrix equation, all steady state probabilities can be uniquely solved. In other words, there will be *n+1* equations and *n* variables. The detailed process may be found in appendix B. Once the vector π (steady-state probabilities) is obtained, various metrics that may characterize the system performance can be calculated.

4.6 Performance Metrics

In this section, we will provide the necessary formulation so to calculate four performance metrics for the secondary layer of our CRN. We will focus our analysis on the SU's performance for two reasons: Firstly, most related works have already explored the static bandwidth PU's performance. Secondly, it becomes simple to derive the PU's performance since tis behavior is similar to the SU_1 (static channel aggregation is common between them), differing only because there is no other layer on top of the PU. The following metrics are derived in this sequence: SU_1's blocking probability (27), SU_2's blocking probability (28), SU_1's forced termination probability (29), SU_2's forced termination probability (30), SU_1's spectrum utilization (31), SU_2's spectrum utilization (32), SU_1's throughput (33) and SU_2's throughput (34).

1. Blocking Probability

Let BP_{SU1}denote the SU_1's blocking probability. It symbolizes the probability that an arriving SU_1 will not be served by the CRN. In such a way, the product of the blocking transition (23) by the vector π and the indication function $I(i,j,k)$ divided by the mean number of SU_1s arrivals (λ_{1-SU}) results in BP_{SU1}. In other words, the following expressions denote the situation where the network is either full of PUs, SU_1s or both. The simplification adopted here means that the resulting blocking probability is equal to the sum of the stationary probabilities of the system being in states that characterize a full network (27) from the SU_1's perspective.

$$BP_{SU1} = \frac{1}{\lambda_{SU1}} \sum_{i=0}^{\left\lfloor \frac{N}{B_{PU}} \right\rfloor} \sum_{j=0}^{\left\lfloor \frac{N-R_1}{B_{SU1}} \right\rfloor} \sum_{k=0}^{\left\lfloor \frac{N-R_2}{B_{SU2}^{min}} \right\rfloor} \gamma_{b1}{}_{(i,j,k)}^{(i,j,k)} * \pi(i,j,k) * I(i,j,k).$$

Applying all conditions from equation (23):

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

(27)

Similarly, BP_{SU2} stands for the SU_2's blocking probability and is calculated in the same way as for the SU_1. Note that by replacing the value and conditions of $\gamma_{b1}{}_{(i,j,k)}^{(i,j,k)}$, the resulting expression will be (28), in other words, the blocking probability will be equal to the sum of the steady probabilities of all states that characterize a full network.

$$BP_{SU2} = \sum_{i=0}^{\left\lfloor \frac{N}{B_{PU}} \right\rfloor} \sum_{j=\left\lceil \frac{N-R_1-(i*B_{PU})}{B_{SU1}} \right\rceil}^{\left\lfloor \frac{N-R_1}{B_{SU1}} \right\rfloor} \sum_{k=\left\lceil \frac{N-R_2-(i*B_{PU})-(j*B_{1-SU})}{B_{SU2}^{min}} \right\rceil}^{\left\lfloor \frac{N-R_2}{B_{SU2}^{min}} \right\rfloor} \pi(i,j,k) * I(i,j,k). \quad (28)$$

2. Forced Termination Probability

The forced termination probability is the probability that a service is interrupted due to a higher priority user arrival. For the SU_1 this shall occur if the there are no available resources for an arrival PU, i.e., $idle + \left(k * B_{SU2}^{min}\right) < B_{PU}$, considering the system state is *(i,j,k)* and using *idle* to denote the number of available resources in state *(i,j,k).* As a result, the number of forcibly terminated SU_1s is $y = \lceil ((i * B_{PU}) + (j * B_{SU1}) - N + B_{PU})/B_{SU1} \rceil$. Then, the forced termination rate of SU_1s will be $\lambda_{PU} * \pi(i,j,k)$. However, we must also eliminate the percentage of users that have never been in the CRN, thus, we have: $(1 - BP_{SU1}) * \lambda_{1-SU}$. In other words, the stationary probability value for each feasible state *i,j,k* indicated by the feasibility function I(*i,j,k*) that has a transition for another state $I(i+1, j-y, 0)$, which implies an SU_1 dropping event will be summed up and multiplied by λ_{PU}, that is, the PU arrival rate. Finally, this value will be divided by the total amount of SU_1 that have been served by the CRN, i.e., excluding the ones that have never entered the network (the blocked individuals).

$$FTP_{SU1} = \frac{\lambda_{PU}}{(1 - BP_{SU1}) * \lambda_{SU1}} \sum_{i=0}^{\left\lfloor \frac{N}{B_{PU}} \right\rfloor} \sum_{j=0}^{\left\lfloor \frac{N-R_1}{B_{SU1}} \right\rfloor} \sum_{k=0}^{\left\lfloor \frac{N-R_2}{B_{SU2}^{min}} \right\rfloor} \pi(i,j,k) * y(i,j,k) * I(i,j,k) * I(i+1, j-y, 0). \quad (29)$$

The lowest priority class of users is subject to forced termination events in the following two situations: First (pt1), 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 (pt2), where the condition is $idle < B_{SU1} + R_1$, which considers channel reservation. The number of SU_2s to be dropped is $z_1 = \left\lceil (-idle + B_{PU})/B_{SU2}^{min} \right\rceil$ in the first case and $z_2 = \left\lceil (-idle + B_{SU1})/B_{SU2}^{min} \right\rceil$ in the latter. Thereof, the SU_2 forced termination probability is denoted as (30).

$$FTP_{SU2,pt1} = \sum_{i=0}^{\left\lfloor \frac{N}{B_{PU}} \right\rfloor} \sum_{j=0}^{\left\lfloor \frac{N-R_1}{B_{SU1}} \right\rfloor} \sum_{k=0}^{\left\lfloor \frac{N-R_2}{B_{SU2}^{min}} \right\rfloor} \lambda_{PU} * \pi(i,j,k) * z_1 * I(i,j,k) * I(i+1,j,k-z);$$

$$FTP_{SU2,pt2} \sum_{i=0}^{\left\lfloor \frac{N}{B_{PU}} \right\rfloor} \sum_{j=0}^{\left\lfloor \frac{N-R_1}{B_{SU1}} \right\rfloor} \sum_{k=0}^{\left\lfloor \frac{N-R_2}{B_{SU2}^{min}} \right\rfloor} \lambda_{SU1} * \pi(i,j,k) * z_2 * I(i,j,k) * I(i,j+1,k-z);$$

$$FTP_{SU2} = \frac{1}{(1 - BP_{SU2}) * \lambda_{SU2}} * \left(FTP_{SU2,pt1} + FTP_{SU2,pt2}\right). \tag{30}$$

3. Spectrum Utilization

The spectrum utilization *(U)* is defined as the ratio of the mean channel occupancy by each SU class to the total number of channels. Considering that every channel is assumed to have a unit bandwidth, the utilization can be calculated by (31) for the SU_1s and by (32) for SU_2s. An important remark in (32) is that B_{SU2*} can be replaced by $B_{SU2,DCAF}$, $B_{SU2,DCA}$ or $B_{SU2,FCA}$, depending on the adopted aggregation technique.

$$U_{SU1} = \frac{1}{N} \sum_{i=0}^{\left\lfloor \frac{N}{B_{PU}} \right\rfloor} \sum_{j=0}^{\left\lfloor \frac{N-R_1}{B_{SU1}} \right\rfloor} \sum_{k=0}^{\left\lfloor \frac{N-R_2}{B_{SU2}^{min}} \right\rfloor} j * B_{SU1} * \pi(i,j,k) * I(i,j,k). \tag{31}$$

$$U_{SU2} = \frac{1}{N} \sum_{i=0}^{\left\lfloor \frac{N}{B_{PU}} \right\rfloor} \sum_{j=0}^{\left\lfloor \frac{N-R_1}{B_{SU1}} \right\rfloor} \sum_{k=0}^{\left\lfloor \frac{N-R_2}{B_{SU2}^{min}} \right\rfloor} k * B_{SU2*}(i,j,k) * \pi(i,j,k) * I(i,j,k). \tag{32}$$

4. Throughput

In this work, the throughput is defined as the mean number of service completions per unit time *(T)*. It is expressed by (33) for the SU_1s and by (34) for the SU_2s. The term $\mu_{2-SU*}(i,j,k)$ will also depend on the chosen aggregation technique.

$$T_{SU1} = \sum_{i=0}^{\left\lfloor \frac{N}{B_{PU}} \right\rfloor} \sum_{j=0}^{\left\lfloor \frac{N-R_1}{B_{SU1}} \right\rfloor} \sum_{k=0}^{\left\lfloor \frac{N-R_2}{B_{SU2}^{min}} \right\rfloor} \mu_{1-SU} * j * \pi(i,j,k) * I(i,j,k)$$

(33)

$$T_{SU2} = \sum_{i=0}^{\left\lfloor \frac{N}{B_{PU}} \right\rfloor} \sum_{j=0}^{\left\lfloor \frac{N-R_1}{B_{SU1}} \right\rfloor} \sum_{k=0}^{\left\lfloor \frac{N-R_2}{B_{SU2}^{min}} \right\rfloor} \mu_{SU2*}(i,j,k) * k * \pi(i,j,k) * I(i,j,k)$$

(34)

4.6 Chapter Summary

Our system was built to cope with three-user priority levels. A CTMC was utilized to model our CRN considering the user arrivals to be independent Poisson processes, the service rates to be exponentially distributed and the bandwidth unit to be one unit channel. The model assumes an equal sharing resource strategy, in which idle channels are equally shared among ongoing SU_2s. ESA is executed whenever the number of ongoing users is altered, either by an arrival or departure event, independently of the chosen aggregation technique.

Our model presents three alternatives for channel aggregation in the lowest user layer: DCAF, DCA and FCA. The first option allows a SU_2 to dynamically assemble a fractionary number of resources within an interval; similarly, the second enables aggregation an integer number of resources between the bounds and lastly, the user is allowed to assemble an integer fixed number of channels. Moreover, the transitions from our analytic model were introduced in a practical manner, with the transitions from state (i,j,k) to other states being summarized in Table 4.1 and the transitions from other states to state (i,j,k) in Table 4.2. The transitions $\gamma_{(i,j,k)}^{(i\prime,j\prime,k\prime)}$ occur from one feasible state (i,j,k) to another (i',j',k') and here are divided in normal (no forced termination) and dropping (forced termination) cases. In addition, the blocking transitions are also depicted in equations (23) and (24).

Aided by Appendix B, the last part of the chapter explores didactically the process of obtaining the steady-state probability vector and from this information how to derive four performance metrics: blocking probability, forced termination probability, spectrum utilization and throughput for both user types from the secondary network.

Chapter 5 -Model Validation and Analysis

This section describes the experiments and metrics adopted to analyze the proposed model outlined in the System Model Chapter. It shows the results obtained by the proposed analytical model and compares them with a simulation model. Then, an analysis is performed highlighting the main improvements of this model in comparison to the existing ones in the literature.

5.1 Evaluated Metrics

Four metrics were employed to evaluate the secondary network, which are as follows: blocking probability, forced termination probability, spectrum utilization and throughput. Since our secondary network is divided into two layers, there will be a total amount of eight performance metrics, which were defined in Chapter 4 System Model.

5.2 Evaluation Cases

In this section, a total amount of three experiments were defined for validating the proposed model. Each experiment is meant to test the three basic model features, namely: bandwidth flexibility, multi-level channel reservation and channel aggregation. We have varied the PU arrival rate within the interval $[\lambda_{PU\,min}, \lambda_{PU\,max}]$, with $\lambda_{PU\,min} \leq \lambda_{PU\,max}$, for each case. Combining these with the other arrival and service rates, we have simulated low and high amounts of PU traffic and therefore observed the CRN behavior in different conditions. The first experiment explored many different input configurations regarding the bandwidth size of each user (PU, SU_1 and SU_2); the second tested the multi-level channel reservation mechanism, and the last experiment compared the performance of three different aggregation approaches applied only to the lowest level users (SU_2), one of which is based on the *equal sharing algorithm (DCA)*, differing by the fact that only integer channel aggregation is allowed (no float sizes are possible). All three experiments were conceived using MATLAB 2015b software and the results were composed of 100 simulation instances performed for each evaluated point (input), with simulation time set to 10,000 time units. The average results are presented considering a 95% confidence level. The interval bars were not plotted because there were many evaluation points, and thus, the images would become very loaded. Together with the mean simulation values, we have plotted our analytical model outcome. The code for all experiments can be found in the following link: https://github.com/mfalcaojr/CHANNEL-RESERVATION-AND-SPECTRUM-

ADAPTATION-STRATEGIES-IN-A-MULTI-LEVEL-PRIORITIZED-COGNITIVE-RADIO-.git

In this Chapter, except when indicated in the following figures, the 'theoretical' (continuous line) expresses the analytical model outputs and the markers (square, triangle, etc.) should stand for the mean simulation value.

5.2.1 Description for the First Experiment

The switchover to digital television is a phenomenon that frees up large and valuable spectrum chunks, but still presents active PUs that usually use wideband transmissions. So, because CR is expected to allow the coexistence between PUs and SUs, the literature usually considers the PUs to have larger or equal bandwidths than the opportunistic users (CHU et al., 2015), (ZHU et al., 2007). However, narrowband primary applications may also exist in those frequency bands, e.g., IEEE 802.22 is a standard for wireless regional area network that uses CR to avoid interference with incumbent TV broadcasting (wideband primary application) and low power licensed devices such as wireless microphone operation (narrowband primary applications) (MISHRA; JOHNSON, 2015).

So previous works have limitations related to the bandwidth size of each user, whereas our approach allows any combination of such parameters, as long as they are multiples between them. Hence, we chose six different input configurations that represent the following situations: $B_{PU} > B_{SU1} > B_{SU2}^{min}$ (configuration 1), $B_{PU} > B_{SU2}^{min} > B_{SU1}$ (configuration 2), $B_{SU1} > B_{SU2}^{min} > B_{PU}$ (configuration 3), $B_{SU1} > B_{PU} > B_{SU2}^{min}$ (configuration 4), $B_{SU2}^{min} > B_{PU} > B_{SU1}$ (configuration 5) and $B_{SU2}^{min} > B_{SU1} > B_{PU}$ (configuration 6). The three values chosen for each configuration were one, two and four; besides, the total number of channels in this CRN was set also set to four. In addition, we have configured all three layers to use static channel aggregation, not allowing the elastic behavior, thus $B_{SU2}^{min}$was equal to B_{SU2}^{max}. Two curves for each configuration were plotted for our four performance metrics, totalizing twelve curves (six from the simulation and six from the analytical model). Table 5.1 summarizes this information.

Table 5.1. Bandwidth configurations for the first experiment

Configuration nº	**Total nº of channels**	**PU bandwidth (B_{PU})**	**SU_1 bandwidth (B_{SU1})**	**SU_2 min bandwidth (B_{SU2}^{min})**	**SU_2 max bandwidth (B_{SU2}^{max})**
1	4	4	2	1	1
2	4	4	1	2	2
3	4	1	4	2	2
4	4	2	4	1	1
5	4	2	1	4	4
6	4	1	2	4	4

The PU arrival rate varied from 1 to 4 with a step of 0.5. The aim here is to show the behavior of the secondary network under different PU loads. The remaining arrival and service rates, i.e., SU_1 arrival rate, SU_2 arrival rate, PU service rate, SU_1 service rate and SU_2 arrival rate were fixed with a single unit value. Furthermore, no channel reservation was considered, implying that SU_1s and SU_2s have full access to the networks resources. These inputs were summarized in Table 5.2. The goal in the first scenario is to demonstrate that, for any given configuration, both analytical and simulated models should output approximately the same results.

Table 5.2. Arrival and service rates for each user layer in the first experiment.

PU arrival rates	**SU_1 arrival rate**	**SU2 arrival rate**	**PU service rate**	**SU_1 service rate**	**SU2 service rate**
1/ 1.5/ 2/ 2.5/ 3/ 3.5/ 4	1	1	1	1	1

Little's Law states that, under steady state conditions, the average number of items in the system should be equal to the average rate at which the items arrive multiplied by the average time that they spend in the system (CHHAJED D.; LOWE, T, 2007). Replacing the arrival/service rate values from Table 5.2 we get the active PU average number in the queueing system (Eq. 35) described in Table 5.3. Moreover, the mean PU network occupation can be found dividing the average number of active PUs by the total number of channels.

$$\lambda_{PU} \cdot \frac{1}{\mu_{PU}} \tag{35}$$

Table 5.3. Arrival and service rates impact in network occupation

PU arrival rate λ_{PU}	PU service rate μ_{PU}	Average number of active PUs	Total nº of channels	PU network occupation
1	1	1	4	25%
1.5	1	1.5	4	37.5%
2	1	2	4	50%
2.5	1	2.5	4	62.5%
3	1	3	4	75%
3.5	1	3.5	4	87.5%
4	1	4	4	100%

5.2.2 Results and Analysis for the First Experiment

The results from the first experiment can be further divided into two groups: SU_1 and SU_2 blocking probabilities together with their forced termination probabilities evaluated separately from the remaining two metrics: SU_1 and SU_2 spectrum utilization and throughput.

5.2.2.1 Blocking and Forced Termination Probabilities

The strictly increasing nature of this section's measures is what characterizes the difference between these and the other two performance metrics. As the network becomes crowded with PUs, it was expected that the secondary blocking and forced termination probabilities would also rise (see Figs. 5.1 to 5.4). This is an obvious remark because only the PU arrival rate was varied along the experiment; therefore, the remaining space for secondary usage was progressively reduced, producing higher blocking and forced termination. In addition, our input configurations provided a large variety of output values as it can be viewed in Fig. 5.1 where, for example, the blocking value for configuration 5 ($B_{SU2}^{min} > B_{PU} > B_{SU1}$) was nearly 0% (PU arrival rate = 1) while configuration 3 ($B_{SU1} > B_{SU2}^{min} > B_{PU}$) reached up to 100% of blocked SU_1s (PU arrival rate = 4).

A non-obvious remark is that one could expect that a configuration that outputs high blocking probability values would also show high forced termination probability values. Analyzing the SU_1, for instance, and taking the highest configuration values from Fig. 5.1 configuration 3 ($B_{SU1} > B_{SU2}^{min} > B_{PU}$) and Fig. 5.3 configuration 2 ($B_{PU} > B_{SU2}^{min} > B_{SU1}$) as examples, at first, their results may seem intriguing. However, because a user type that has a high blocking probability is likely to be less served, low forced termination probability should happen. In other words, there will be fewer users to be dropped, hence a low forced termination probability, and so the results are correct. Table 5.4 outlines this behavior for configurations 2 and 3 for the SU_1.

Table 5.4. Relationship between blocking and forced termination for SU_1 using approximated values.

Configuration nº	SU_1 Blocking Probability [min, max]	SU_1 Forced Termination Probability [min, max]
2	[~20%, ~50%]	[~50%, ~80%]
3	[~70%, ~100%]	[~20%, ~50%]

In Figs. 5.3 and 5.4 two curves overlap for similar reasons. In Fig 5.3 for example, configurations 3 ($B_{SU1} > B_{SU2}^{min} > B_{PU}$) and 4 ($B_{SU1} > B_{PU} > B_{SU2}^{min}$), both with B_{SU1} = 4, present the same results. Although B_{PU} varies from one configuration to another, one SU_1 occupies the whole network when active (because the total channel amount is set to four), which implies that any value for B_{PU} will cause the same effects on the SU_1 performance. The same occurs in the SU_2's forced termination probabilities for configurations 5 ($B_{SU2}^{min} > B_{PU} > B_{SU1}$) and 6 ($B_{SU2}^{min} > B_{SU1} > B_{PU}$), see Fig. 5.4.

Fig. 5.1. SU_1 Blocking Probability

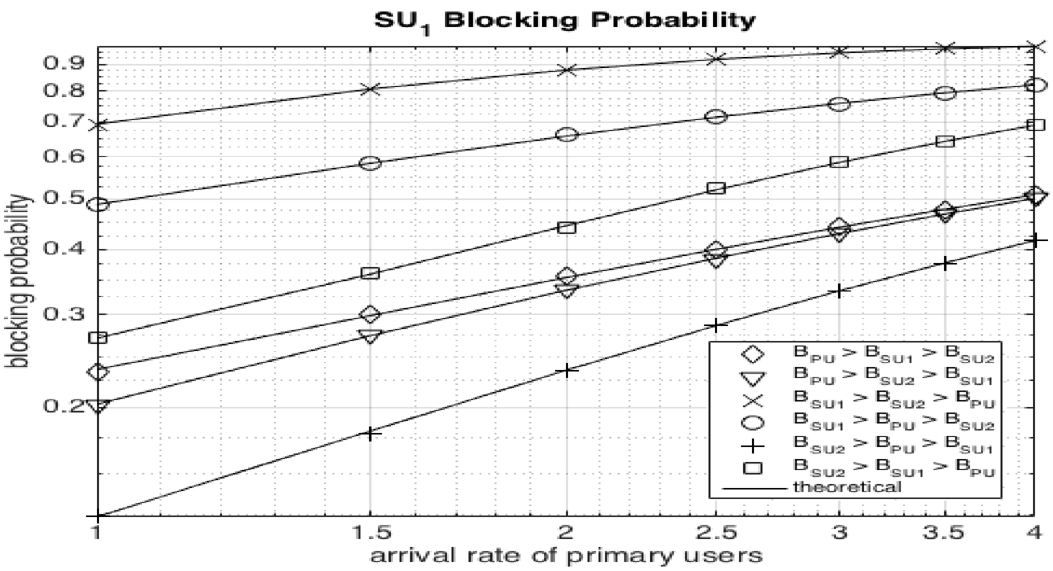


Fig. 5.2. SU_2 Blocking Probability

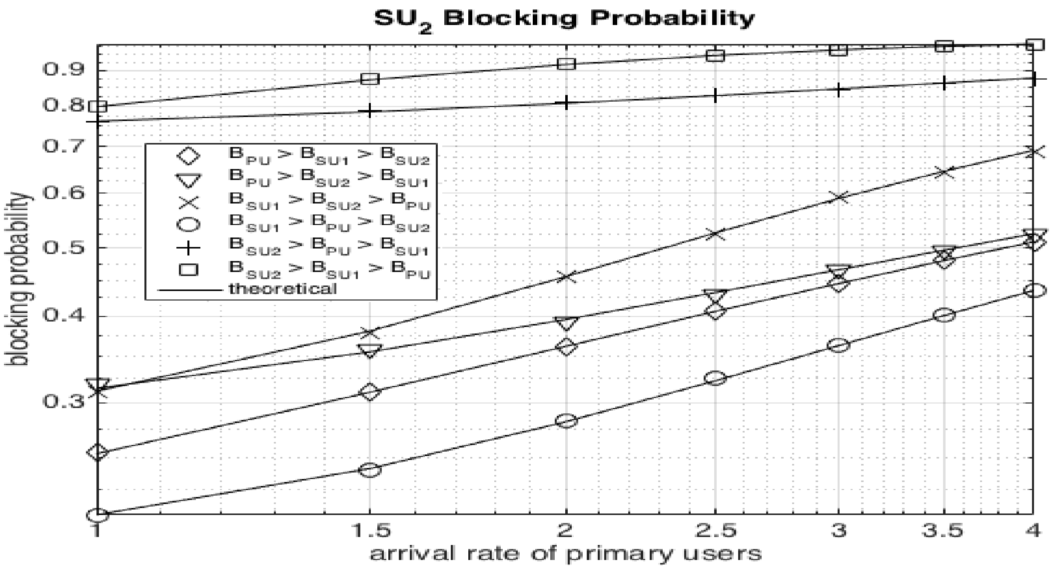


Fig. 5.3. SU_1 Forced Termination Probability

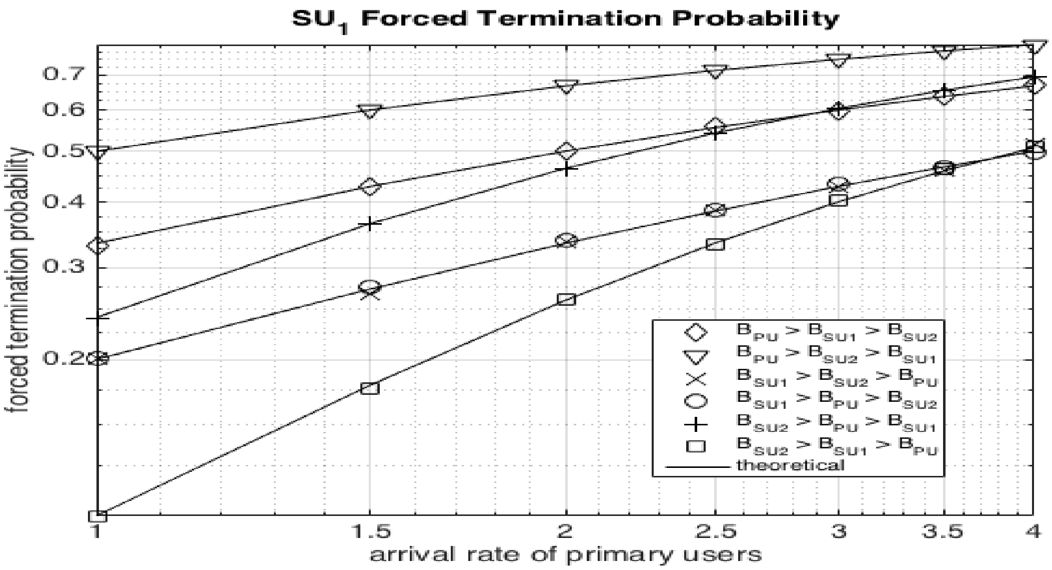
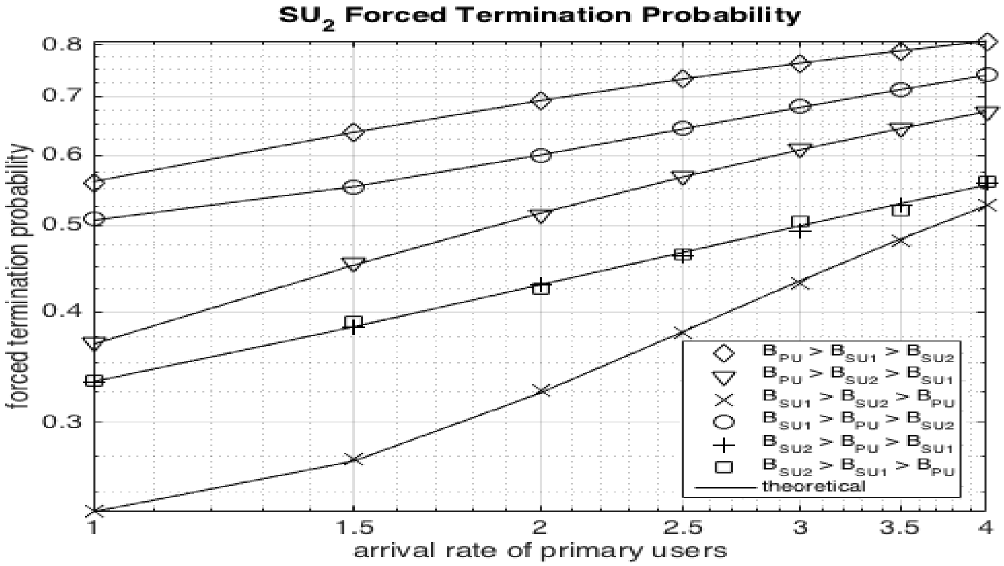


Fig. 5.4. SU_2 Forced Termination Probability



5.2.2.2 Spectrum Utilization and Throughput

In this set of performance metrics, a strictly decreasing behavior can be observed as the network becomes loaded with PUs. Again, this was expected because all arriving/service rates were fixed, but the PU arrival rate always increases. So, for the same reason as in the previous metrics, the remaining space for secondary usage was progressively reduced, lowering spectrum utilization and throughput. Nevertheless, differently to what happened in the previous set of metrics, there is a relationship between the mean configuration performances towards the spectrum utilization and the throughput in both user layers. By comparing Fig. 5.5 with Fig. 5.7 and Fig. 5.6 with Fig. 5.8, it is noticeable that they present the same performance order (mean values), which is summarized in Table 5.5.

Table 5.5. Relationship between spectrum utilization and throughput

	SU_1 Spectrum Utilization	**SU_1 Throughput**	**SU_2 Spectrum Utilization**	**SU_2 Throughput**
Configuration performance order (Worst to best)	[3,2,4,1,6,5]	[3,2,4,1,6,5]	[6,5,1,4,2,3]	[6,5,1,4,2,3]

However, some configurations such as 5 and 6 (cross and square markers respectively) almost overlap (Fig. 5.5). It is noticeable that for part of the figure where the PU load is low, both inputs perform similarly, whereas when $\lambda_{PU} = 2.5$ both curves diverge. In a loaded network, it becomes easier to fit smaller bandwidth SUs (Configuration 5 has $B_{PU} = 2\ and\ B_{SU1} = 1$), so its utilization should be greater than configuration 6, that has $B_{PU} = 1\ and\ B_{1-SU} = 2$.

Similarly to the previous comparison, another interesting behavior regarding the SU_1's spectrum utilization (Fig. 5.5) comprises configurations 2 ($B_{PU} > B_{SU2}^{min} > B_{SU1}$) and 4 ($B_{SU1} > B_{PU} > B_{SU2}^{min}$). Note that the SU_1's bandwidth is one unit channel for configuration 2 (inverted triangle marker) and four channels for the other input (circle marker) and although there is a considerable difference in these values, they perform almost equally, but intersecting and changing the performance order in the point where $\lambda_{PU} = 2.5$. For this example, when $\lambda_{PU} < 2.5$, higher SU_1 bandwidth configurations will likely have higher occupancy than smaller bandwidth inputs and for $\lambda_{PU} > 2.5$ the opposite occurs.

In the previous metrics, particularly at Figs. 5.3 and 5.4, we have noted that two inputs completely overlap. Taking these same configurations 3 ($B_{SU1} > B_{SU2}^{min} > B_{PU}$) and 4 ($B_{SU1} > B_{PU} > B_{SU2}^{min}$), that have both with $B_{SU1} = 4$, in Fig. 5.5 a curious phenomenon happens. Differently from previous evaluation metrics where they performed equally, with regard to the SU_1's spectrum utilization, configuration 4 outperforms input 3 by almost 2.5% (mean value), although for this configuration, B_{PU} is higher. One could expect that by occupying more channels, the PUs would push the SUs out of the network as λ_{PU} increases. Instead, higher B_{PU} also makes them being serviced faster, freeing up spectrum resources and benefiting the SU_1s, in this case. For the SU_2s, this phenomenon also happens but in a smaller scale. Taking inputs 5 ($B_{SU2}^{min} > B_{PU} > B_{SU1}$) and 6 ($B_{SU2}^{min} > B_{SU1} > B_{PU}$) in Fig. 5.6, for example, we note that their mean difference is only about 1%. In such case, two user types alternate their bandwidths and due to the priority order, the performance difference of these two inputs is not as much as the mean difference value for the SU_1's spectrum utilization.

This experiment explored many configuration/network load possibilities and, as it can be ascertained, for every input the correspondent simulation output (markers) was accompanied by a theoretical value (line) originated from our analytical model, proving its correctness for these instance sets.

Fig. 5.5. SU_1 Spectrum Utilization

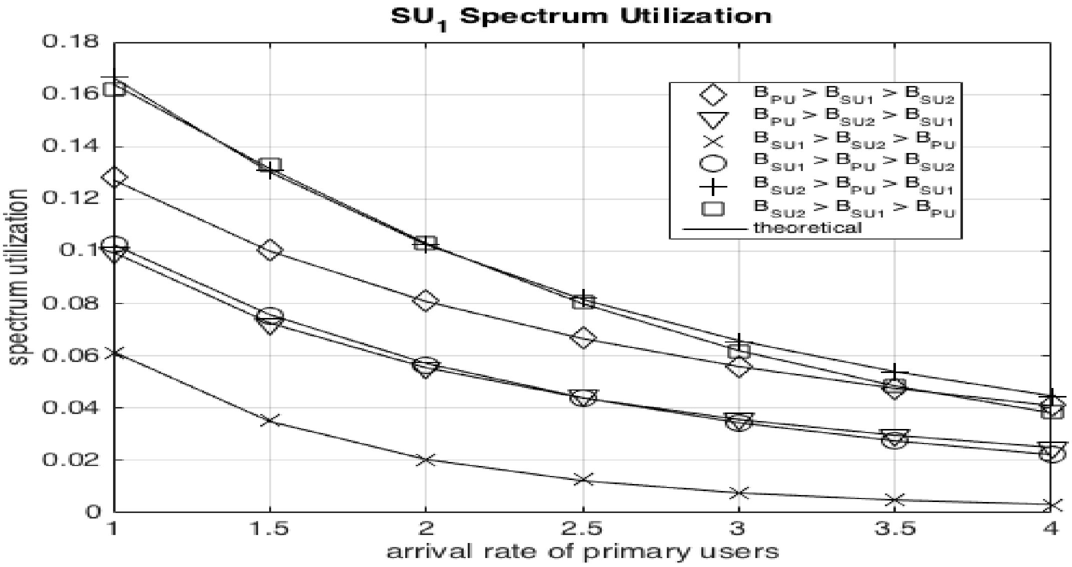


Fig. 5.6. SU_2 Spectrum Utilization

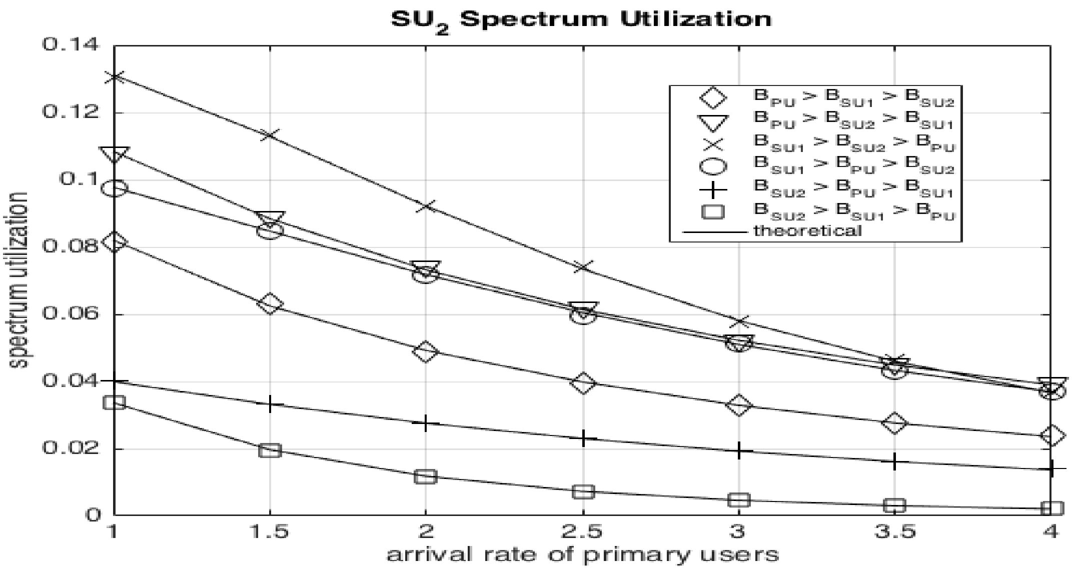


Fig. 5.7. SU_1 Throughput

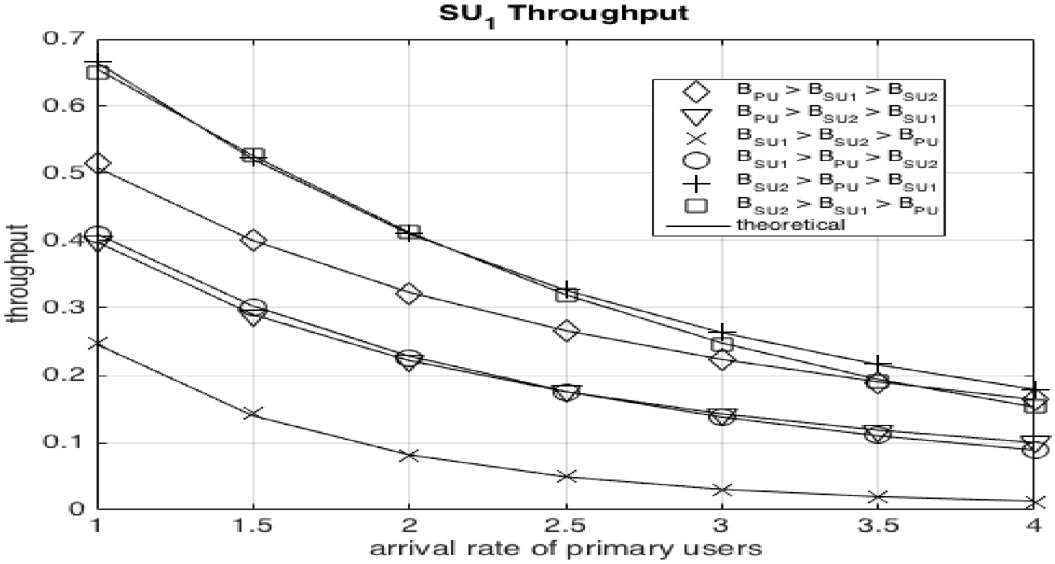
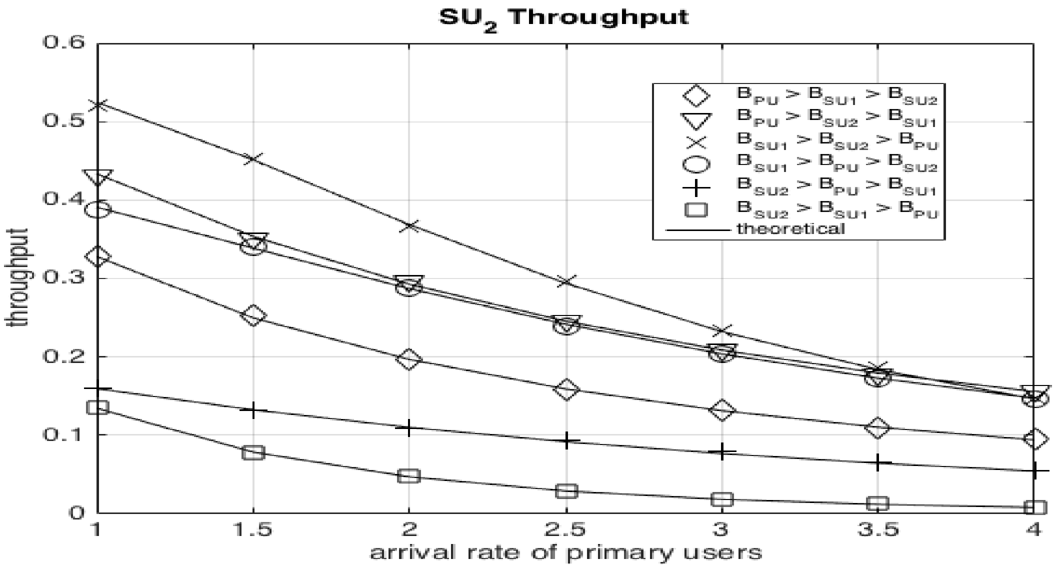


Fig. 5.8. SU_2 Throughput



5.2.3 Description for the Second Experiment

The first experiment tested different bandwidth configuration examples in order to validate part of the proposed model. In this section, another model feature named “Channel Reservation” will be tested. Channel reservation is a prioritization mechanism proposed in the literature for QoS provisioning in CRNs (YAFENG et al., 2015) (ZHU et al., 2007) (CHU et al., 2015). Allowing channel reservation means blocking access to new secondary arrivals even if there are enough available channels, lessening the number of secondary sessions to be forcibly terminated. Thus, it enables a tradeoff between forced termination and blocking probabilities, according to the QoS requirements of the secondary traffic.

The problem in most works is that they are strictly concerned in optimizing the tradeoff between blocking probability and forced termination probability by means of an iterative algorithm that tests many input possibilities and compares the model's answer in terms of those performance metrics to some pre-defined threshold (YAFENG et al., 2015) (ZHU et al., 2007) (CHU et al., 2015). However, it is noticeable that these authors always perform their experiments under low PU loads. For example, in (CHU et al., 2015), the author uses PU arrival rates ranging from 0.1 to 1 and a PU service rate equal to 30. In other words, according to (Eq. 35), this gives an extremely low PU load (considering the number of channels adopted), which poorly tests their proposals. For this reason, we have opted for higher PU loads in order to compute the afore-mentioned tradeoff under a more diverse scenario. Also, differently from these authors, we have provided two different channel reservation variables R_2 and R_1 one for each secondary layer, which causes different effects from those approaches in the literature.

Denoting R_1 and R_2 as the number of reserved channels from each secondary layer (SU_1 and SU_2) and hence, each user layer may access only $N - R_1$ and $N - R_2$ channels, we have chosen four different input configurations that represent the following situations: $R_1 = R_2 = 0$ (configuration 1), $R_2 > R_1$ (configuration 2), $R_1 > R_2$ (configuration 3) and $R_1 = R_2 = 1$ (configuration 4). The values chosen for configurations two and three were zero and one; besides, the total number of channels in this CRN was set to four. Moreover, again we have configured all three layers to use static channel aggregation, not allowing the elastic behavior. Two curves for each configuration were plotted for our four performance metrics. In this experiment, all user bandwidths were set to one unit. Table 5.6 summarizes this information.

Table 5.6. Bandwidth and reserved channels configurations for the second experiment

Configuration nº	**Total nº of channels**	B_{PU}	B_{SU1}	$B_{SU2}^{min} = B_{SU2}^{max}$	R_1	R_2
1	4	1	1	1	0	0
2	4	1	1	1	0	1
3	4	1	1	1	1	0
4	4	1	1	1	1	1

The PU arrival rate varied from 0.2 to 1.4 with a step of 0.2. The remaining arrival and service rates, that is, SU_1 arrival rate, SU_2 arrival rate, PU service rate, SU_1 service rate and SU_2 arrival rate were fixed with a single unit value. These inputs are summarized in Table 5.7.

Table 5.7. Arrival and service rates for each user layer in the second experiment

PU arrival rates	SU_1 arrival rate	SU_2 arrival rate	PU service rate	SU_1 service rate	SU_2 service rate
0.2/ 0.4/ 0.6/ 0.8/ 1/ 1.2/ 1.4	1	1	1	1	1

5.2.4 Results and Analysis for the Second Experiment

The results from the second experiment are divided into two groups as in the previous experiment, but instead, it presents SU_1 metrics separately from the SU_2 metrics. This experiment aims the demonstration that our model supports multi-level channel reservation.

5.2.4.1 SU_1 Performance Metrics

This set of performance metrics is composed by four different configurations and should provide us four different curves (eight if counting theoretical and simulation). Although Figs. 5.9, 5.10, 5.13 and 5.14 seem to present only half number of outputs, they correctly show the expected number of curves, but some of them overlap. The explanation for this behavior is simple. Because only one user layer, in our case the PUs, pressures the SU_1s then, configurations one and two ($R_1 = 0$) will naturally provide the same values while configurations three and four ($R_1 = 1$) another set of equal values. As expected, the number of reserved channels for the SU_2s does not affect the results for the SU_1s.

Fig. 5.9. SU_1 Blocking Probability

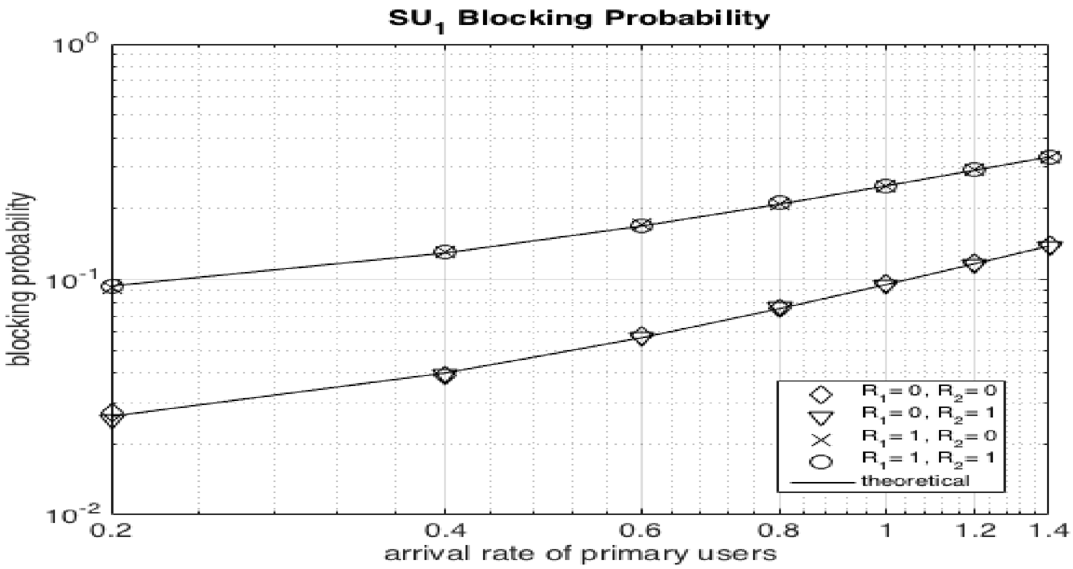
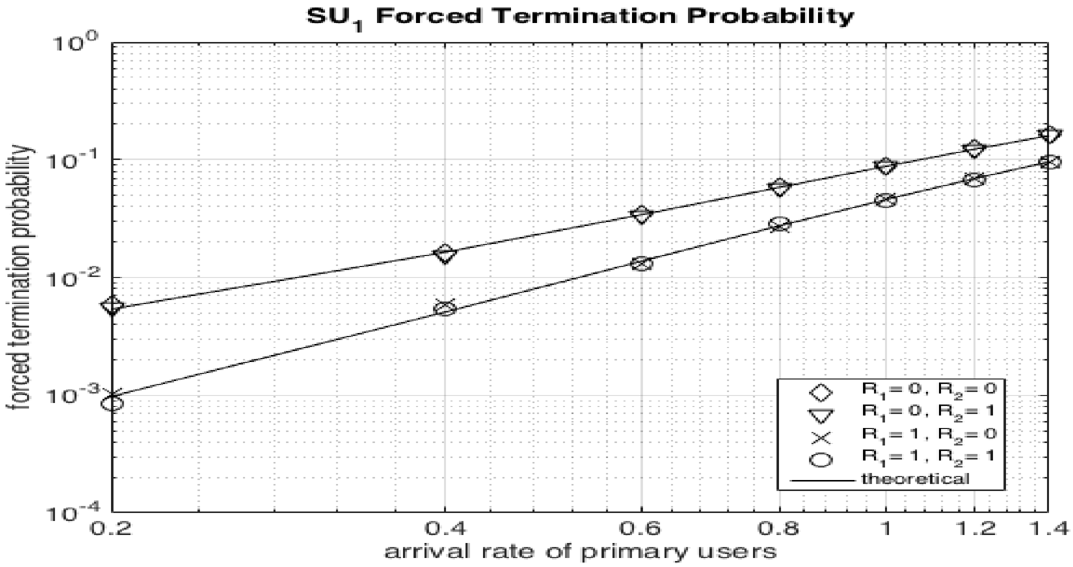
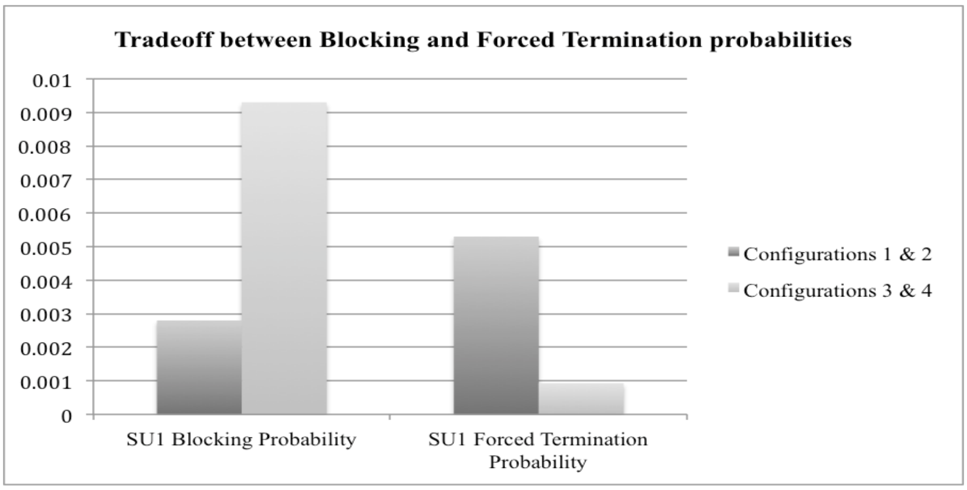


Fig. 5.10. SU1 Forced Termination Probability



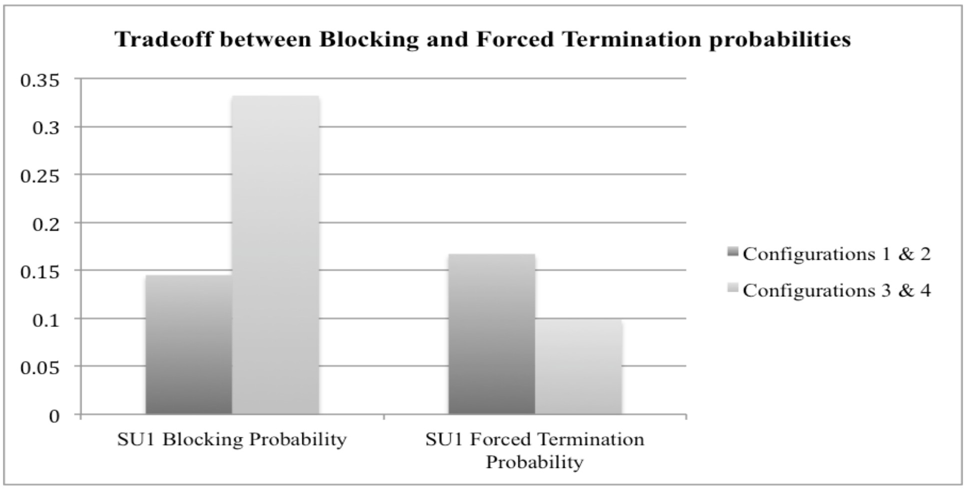
We have chosen a low PU load compared to our previous experiment but a high PU load if compared to other works (YAFENG et al., 2015) (ZHU et al., 2007) (CHU et al., 2015). Taking the point where the PU arrival rate equals 0.2 (lowest PU load), for instance, we have computed the tradeoff between the SU_1's blocking (Fig. 5.9) and forced termination (Fig. 5.10) probabilities in Fig. 5.11. The evaluated point gives the largest tradeoff value, that is, when channel reservation is set to one unit (configurations 3 & 4) the blocking probability is approximately three times higher than when channel reservation is not considered (configurations 1 & 2). On the other hand, the forced termination probability is reduced by a factor of five comparing the same approaches, proving that the tradeoff between these metrics exists, especially considering a very low PU load.

Fig. 5.11. Tradeoff between blocking and forced termination probabilities for PU arrival rate equal to 0.2



Other studies in the literature might show even better tradeoff values, as in (CHU et al., 2015). Because they are usually interested in testing search algorithms for finding the optimal or near optimal number of reserved channels, although they usually lack analyzing the same tradeoff value in higher PU loads. In Fig. 5.12, we evaluate the other extreme point of Figs. 5.9 and 5.10 (PU arrival rate equals 1.4) and it is noticeable that the previous tradeoff suffers an inversion. For this evaluation point, when channel reservation is triggered, we note that the blocking probability doubles its value whereas the forced termination probability reduces by a factor of only 1.5. Besides this fact, it is important to highlight that in crowded networks, the secondary users will naturally experience higher blocking, making channel reservation unfeasible as it will always tune the blocking probability up. In this case for instance, configurations one and two show a blocking probability of almost 15%.

Fig. 5.12. Tradeoff between blocking and forced termination probabilities for PU arrival rate equal to 1.4.



Previous works would also offer only blocking and forced termination probabilities in their analysis. In this experiment though, we have also provided the outputs for both spectrum utilization (Fig. 13) and throughput (Fig. 14), which are not heavily influenced by channel reservation under the tested configurations. For example, in the worst case (PU arrival rate equal to 1.4), the spectrum utilization reduces only by about 2% when channel reservation is taken into account. Looks as if the network was accepting many SU_1s but dropping them rapidly, so even considerably restricting SU_1s admission, the effects on both utilization and throughput were negligible.

Fig. 5.13. SU_1 Spectrum Utilization

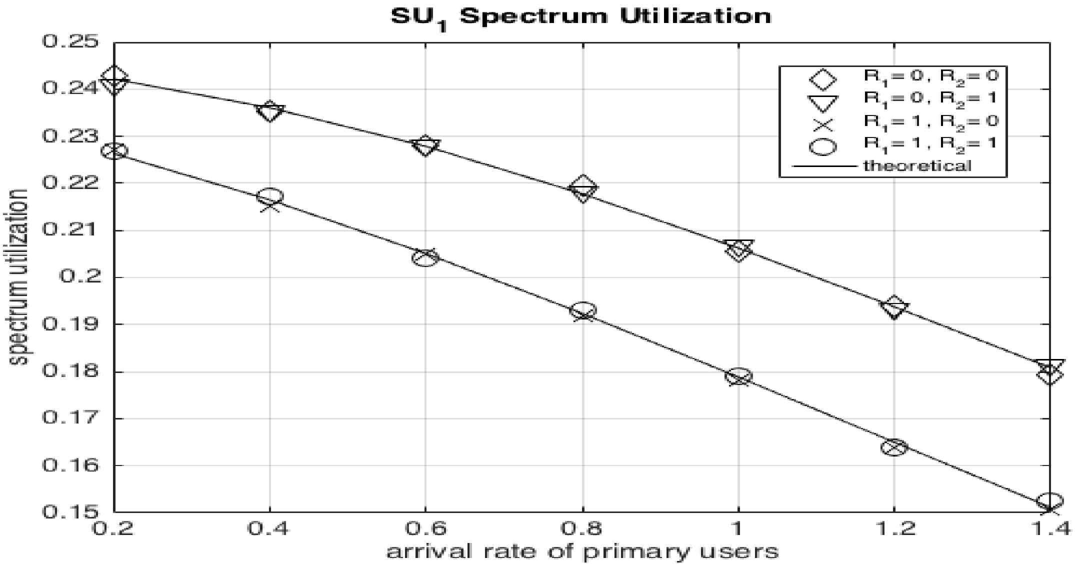
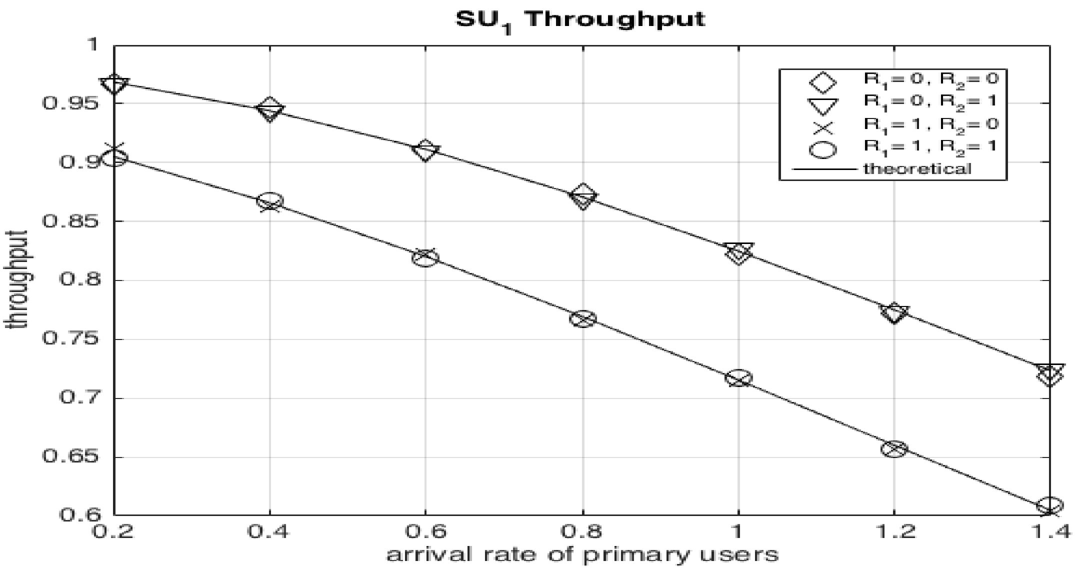


Fig. 5.14. SU_1 Throughput



5.2.4.2 SU_2 Performance Metrics

For this section, the results are obtained considering the multi-level channel reservation, which shows a more complex behavior. Differently from the previous user layer, the four configurations result in four different curves for each proposed performance metric (Fig. 5.15, Fig. 5.16, Fig. 5.17 and Fig. 5.18).

Analyzing the blocking probability in Fig. 5.15 for instance, we note that configuration three $(R_1 = 1, R_2 = 0)$ hits 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 $(N - R_2 = N)$ and consequently, fewer users harass the SU_2's performance. Curiously, regarding the forced termination probability (Fig. 5.16), we have observed that configuration three provides the third worse performance. This might seem contradicting at a first sight, but again, because configurations two and four have $R_2 = 1$, it implies fewer SU_2s being admitted in the CRN, which lessens the probability of a SU_2 being forcibly terminated.

Fig. 5.15. SU_2 Blocking Probability

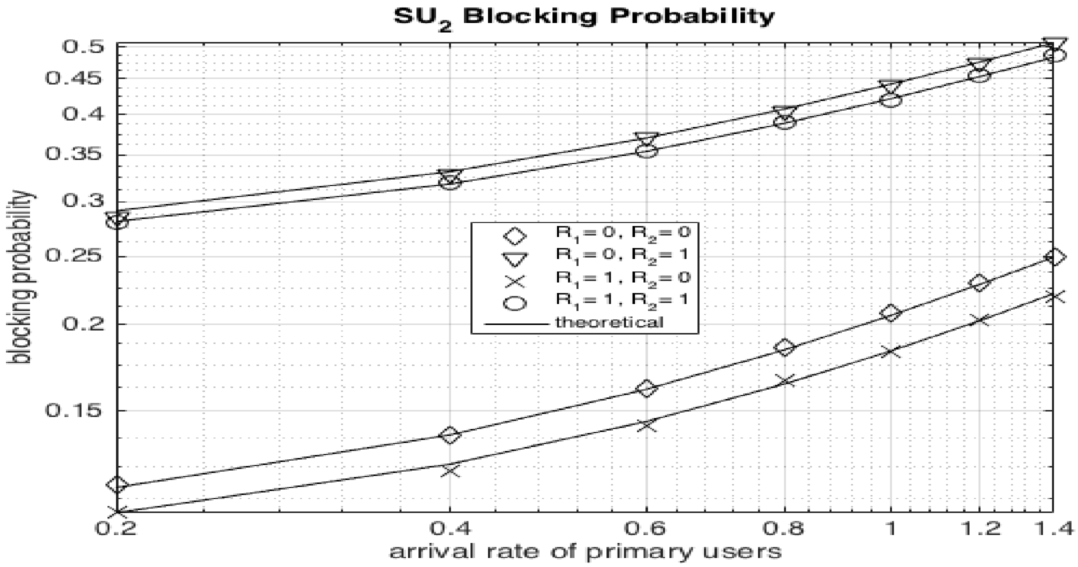
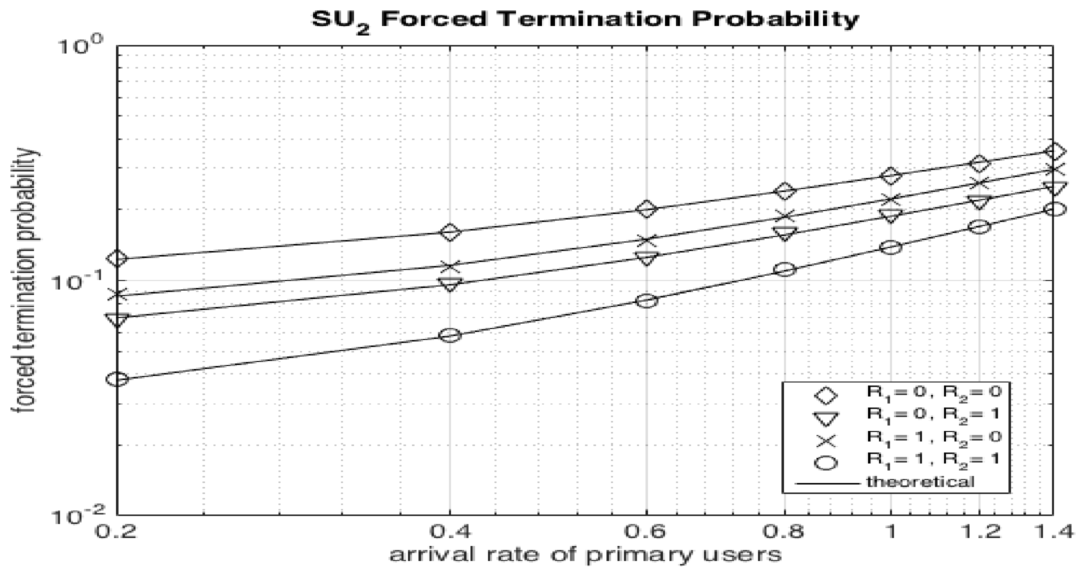


Fig. 5.16. SU_2 Forced Termination Probability



On the spectrum utilization (Fig. 5.17) and throughput (Fig. 5.18) side, the performance order from each configuration is maintained in both metrics. As expected, configuration three outperforms the other inputs as the network becomes sparser for the SU_2s. In his configuration, less SU_1s are accepted while the SU_2's flow is maintained, which gives, for example, about 3% spectrum utilization advantage throughout the experiment compared to when no channel reservation is considered (configuration 1).

Fig. 5.17. SU_2 Spectrum Utilization

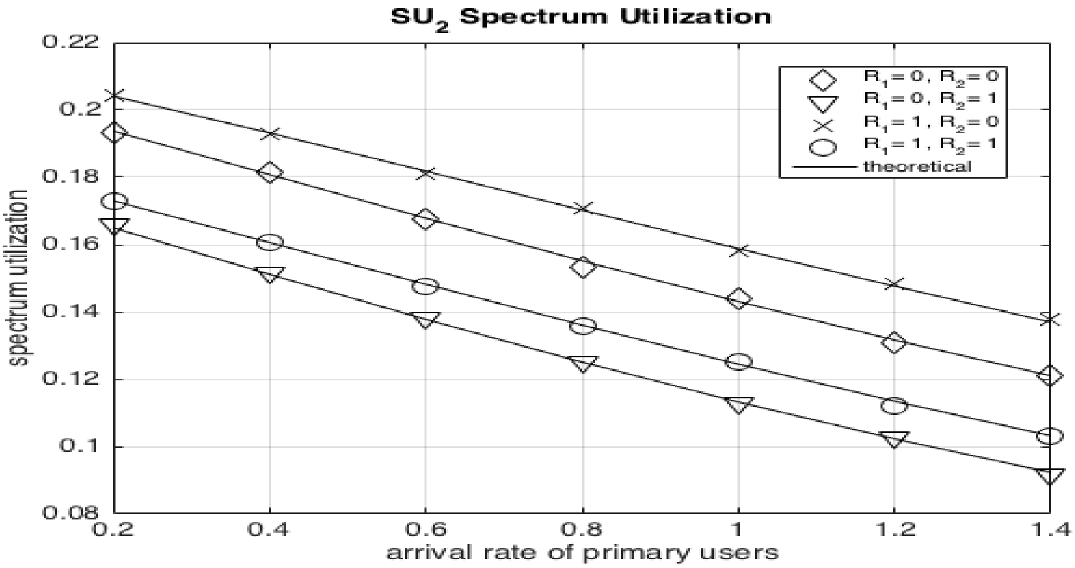
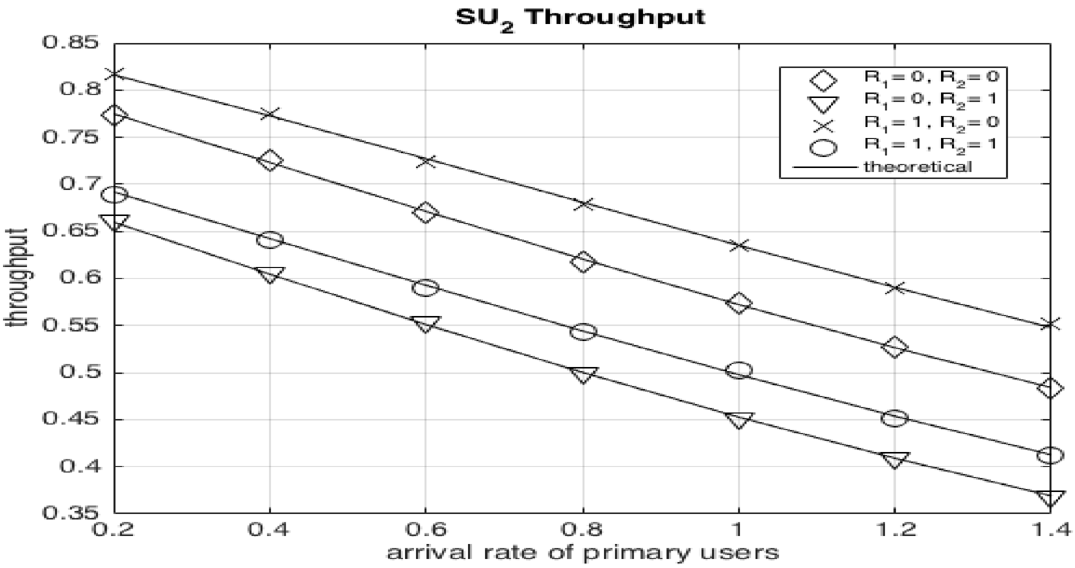


Fig. 5.18. SU_2 Throughput



It is noticeable how the SU_2 layer's performance might vary according to the chosen channel reservation numbers. According to the related works section, the most similar model to ours (CHU et al., 2015) also provides multi-level channel reservation but uses the same variable for reserving channels in their secondary layers. In other words, their model is only able to reproduce part of our configurations, where $R_1 = R_2 = 0,1,2 \dots N$, and not different values for R_1 and R_2. In addition, similarly to other authors, they perform an iterative algorithm that varies R_1 and R_2 searching for the best metric values, regardless the PU load.

Knowing that for larger networks (total channel number) and more priority levels, such algorithms may take too long to respond, we believe that other approaches may lower the problem's complexity. As a suggestion, it should be indispensable to check the PU occupation before proceeding to an iterative algorithm. After all, the PU load might be too high for channel reservation and so the algorithm would not waste time running possibilities for $R_1$and R_2.

The other downside of the channel reservation approach is that, depending on the inputs, the spectrum might become underutilized because as fewer clients will be served and the active users are unable to use those reserved spectrum portions, if static channel assembling is being currently used. The next experiment will test how to boost utilization regardless of the number of reserved channels, by means of dynamic channel assembling, making these two features complementary.

5.2.5 Description for the Third Experiment

The third experiment comprises the comparison between three distinct aggregation techniques for the lowest user layer (SU_2). Provided that this experiment is focused on the SU_2's performance, we have fixed the same parameters for the two other user layers and varied the SU_2's aggregation mechanism, allowing the elastic behavior, which means different values for B_{SU2}^{min} and B_{SU2}^{max}. Besides the common fixed channel aggregation approach (FCA), two other strategies will be tested, namely: dynamic channel aggregation (DCA) and dynamic channel aggregation and fragmentation (DCAF). The first allows assembling an integer number of channels while the latter, a fractionary number, which implies channel sharing among many SU_2s. For this reason, a total amount of three curves should be plotted for each performance metric (three from the simulation and three from the analytical model). Additionally, we have updated the total channel number to twelve for better viewing the aggregation effects, but have not considered channel reservation as it has been already tested in the previous experiment. For simplicity, we have set all user bandwidth values to one unit, except on DCA and DCAF that have both $B_{SU2}^{min} = 1$ and $B_{SU2}^{max} = 5$. Table 5.8 summarizes this information.

Table 5.8. Bandwidth and reserved channels configurations for the third experiment

Configuration nº	Total nº of channels	B_{PU}	B_{SU1}	$R_1 = R_2$	$[B_{SU2}^{min}, B_{SU2}^{max}]$	Aggregation Strategy
1	12	1	1	0	[1,1]	FCA
2	12	1	1	0	[1,5]	DCA
3	12	1	1	0	[1,5]	DCAF

The PU arrival rate varied from 1 to 4 with a step of 0.5. The remaining arrival and service rates were set to the values depicted in Table 5.9. Again, these values were chosen according to Little's law, providing low and high PU loads. Furthermore, this input is similar to that used in (LI et al., 2012).

Table 5.9. Arrival and service rates for each user layer in the third experiment

PU arrival rates	SU_1 arrival rate	SU_2 arrival rate	PU service rate	SU_1 service rate	SU_2 service rate
1/ 1.5/ 2/ 2.5/ 3/ 3.5/ 4	4.6	4.6	0.45	1	1

5.2.6 Results and Analysis for the Third Experiment

In (LI et al., 2012), the authors explored the fixed channel aggregation under two configurations: a maximum and a minimum aggregation rule, aside from the channel aggregation and fragmentation strategy (CAF). Their aim was to prove that CAF would outperform both rules for every performance metric. In this experiment though, we have provided an equivalent form of their minimum aggregation rule (configuration 1) that adopts the minimum bandwidth value. Similarly, the maximum aggregation rule would set the maximum bandwidth value, which would result in a configuration $B_{SU2}^{min} = B_{SU2}^{max} = 5$. However, we have not considered this last option for our experiments. Instead, we have provided an alternative form of aggregation approach called DCA.

Intuitively, the DCA strategy seems to be a compromise of the two other approaches such that the system performance should fall in between them, although its lower and upper bandwidth bounds are the same as in DCAF ($B_{SU2}^{min} = 1$ and $B_{SU2}^{max} = 5$). By analyzing Figs. 5.19 to 5.22, this hypothesis is confirmed. Both dynamic aggregation strategies outperform the fixed aggregation approach in the both extremes of these figures. Because DCA and DCAF are able to assemble up to five channels, we would expect a larger difference from

FCA when the PU load is low (PU arrival rate equal to one). When the network is crowded (PU arrival rate equal to four), DCA and DCAF converge while there is still a considerable difference to FCA. It is likely that for a more crowded network, all three strategies would converge, as the load would pressure the SU_2's bandwidth to the minimum value, which in this case is one channel unit.

Fig. 5.19. SU_2 Blocking Probability

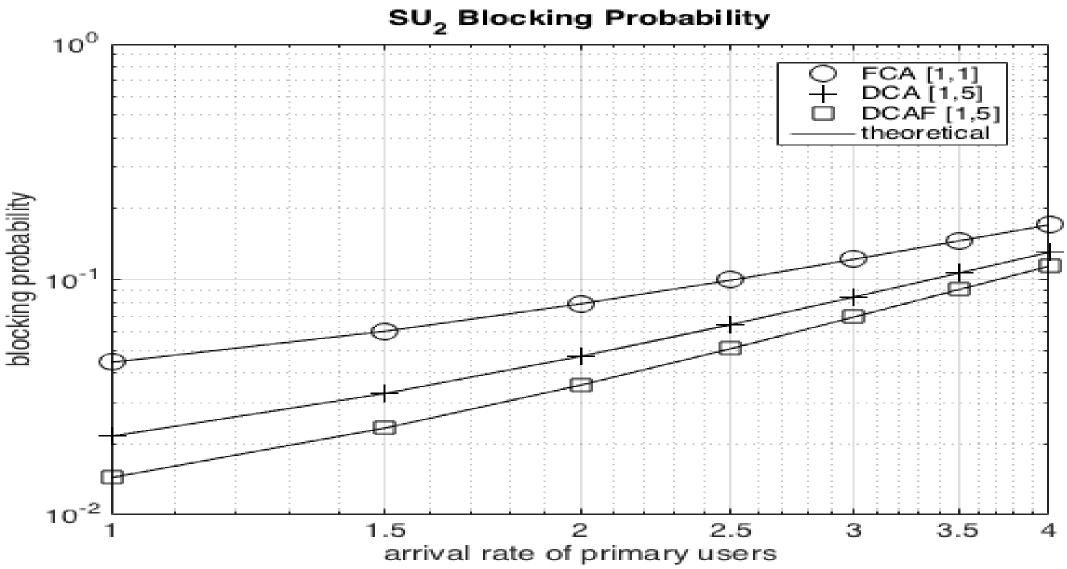


Fig. 5.20. SU_2 Forced Termination Probability

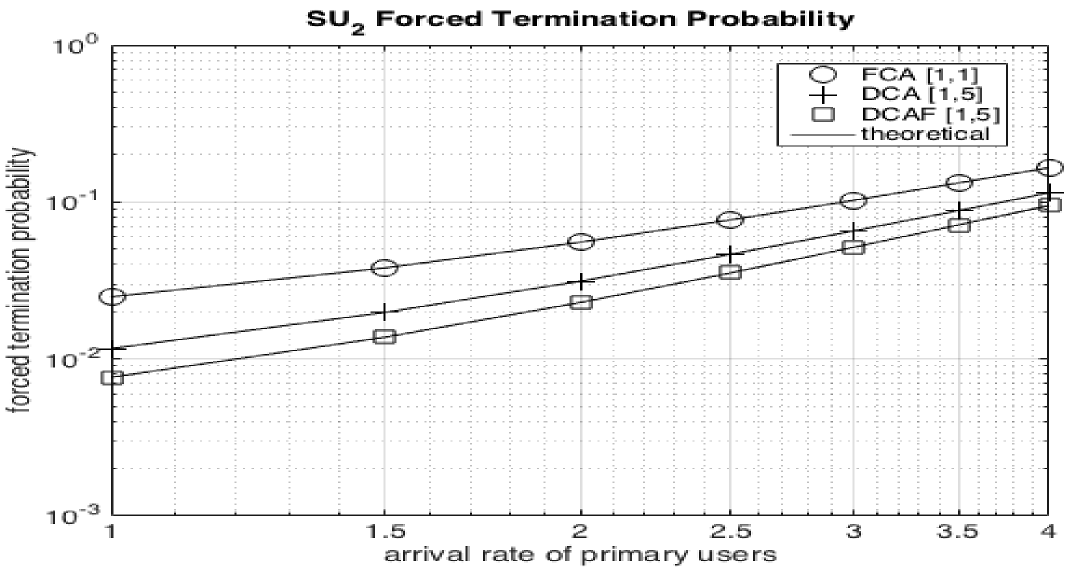


Fig. 5.21. SU_2 Spectrum Utilization

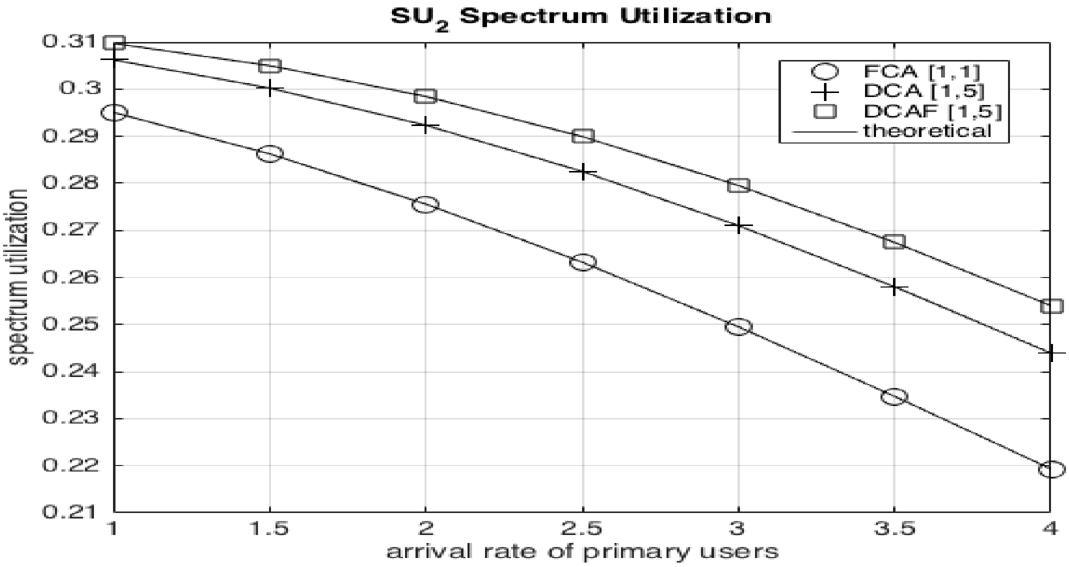
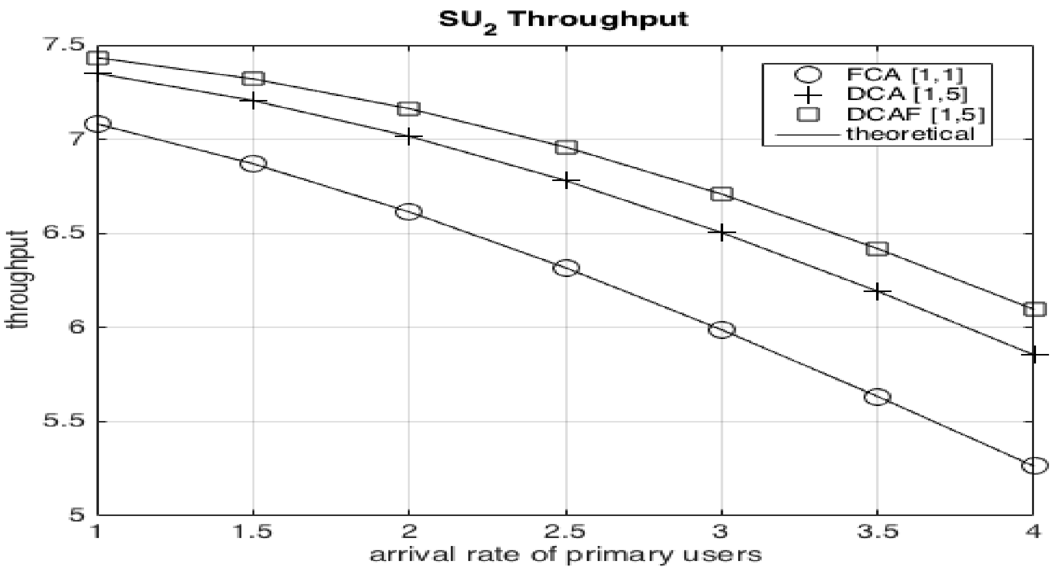
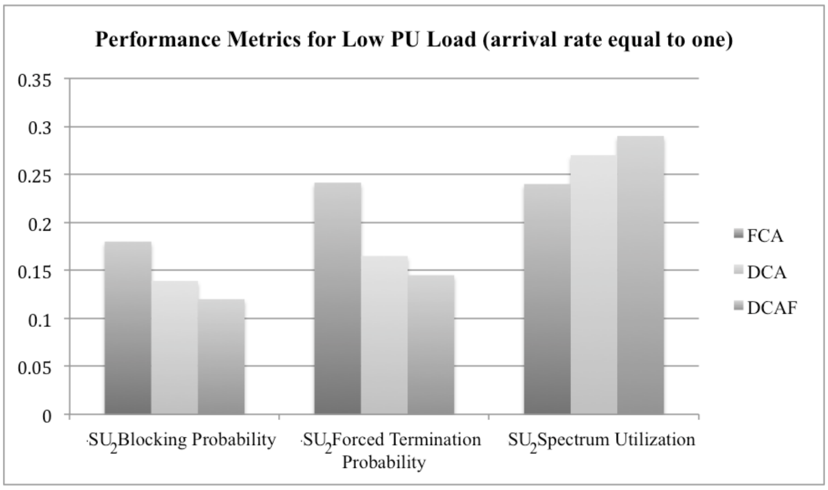


Fig. 5.22. SU_2 Throughput



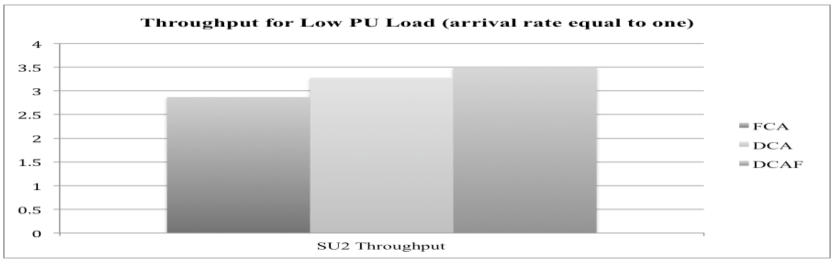
In Fig. 5.23 it becomes clear that upon strategy variation, the SU_2's blocking probability decrease is soft whereas the forced termination probability experiences a more steep reduction. This happens because the blocking probability is mainly affected by B_{SU2}^{min}, which in all three cases is set to one channel unit. Higher values for B_{SU2}^{max} will slightly decrease ongoing user's service times, which will make them finish earlier and thus release spectrum for new arrivals. Differently from the previous metric, the impact of B_{SU2}^{max} is noticeable in SU_2's forced termination probability. Clearly, tuning B_{SU2}^{max} highly benefits ongoing secondary users, as these are less likely to be terminated once their service times become much shorter.

Fig. 5.23. Blocking probability, forced termination probability and spectrum utilization for PU arrival rate equal to 1.



With regard to the spectrum utilization, again the difference becomes linear, almost as for the blocking probability. This specific metric is highly dependent of B_{SU2}^{max} but it also depends on the SU_2s arrival rate, which was set to 4.6 users per time unit. The utilization, although improved, is less affected because there are still too many users arriving and being accepted by the CRN, which does not means that they will complete service. For this reason, by evaluating the throughput in Fig. 5.24 (different scales), it is possible to view that DCAF is able to provide almost 0.7 more service completions per unit time than FCA, meaning that is much more efficient.

Fig. 5.24. Throughput for PU arrival rate equal to 1



The results for this section followed our expectations towards the afore-mentioned aggregation strategies. We knew beforehand that DCAF would outperform DCA and FCA. Even though our work is focused on a theoretical evaluation, we have proposed DCA because we have not found such a proposal in our literature review. Besides, we are aware that there is a concern regarding the feasibility of channel aggregation in real experimentation (ALKHANSA et al., 2014) and that DCAF is still distant from being conceived. Moreover, Figs 5.19 to 5.22 prove that DCAF and DCA perform similarly. So driven mainly by the network speed demands, DCA shows itself as a more realistic option for real-life implementation, and although we have used an analytical model in order to provide the results, these are frequently reasonable alternatives for estimating the benefits of each proposed technique and can be used to endorse further efforts on a real testing environment.

5.3 Evaluating Channel Reservation and Channel Aggregation Simultaneously

The present section aims at demonstrating the performance improvement that the system might achieve when both channel aggregation and channel reservation are tuned simultaneously. As previously discussed, channel reservation might not be an interesting 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 in order to mitigate the negative effects that may be caused by channel reservation. The results will be shown only for the SU_2s.

5.3.1 Description for the Experiment

Again, provided that this experiment is focused on the SU_2's performance when applying together channel reservation and aggregation (DCA), we have fixed the same parameters for the two other user layers and varied the SU_2's reservation R_1 and R_2 and the bandwidths B_{SU2}^{min} and B_{SU2}^{max} values. Four configurations will be tested and therefore, a total amount of four curves should be plotted for each performance metric (four from the analytical model) and the total amount of channels was set to twelve. Once the previous experiment simulations have demonstrated our analytical model precision, the results for this section will be based solely on our analytical model outputs. Table 5.10 outlines the four configurations to be tested.

Table 5.10. Bandwidth and reserved channels configurations for the fourth experiment

Configuration nº	**Total nº of channels**	B_{PU}	B_{SU1}	R_1	R_2	$[B_{SU2}^{min}, B_{SU2}^{max}]$	**Aggregation Strategy**
1	12	1	2	0	2	[4,4]	FCA
2	12	1	2	0	4	[4,4]	FCA
3	12	1	2	4	6	[4,4]	FCA
4	12	1	2	4	6	[1,4]	DCA

The PU arrival rate varied from 1 to 4 with a step of 0.5. The remaining arrival and service rates were set to the values depicted in Table 5.11. Again, these values were chosen according to Little's law, providing low and high PU loads.

Table 5.11. Arrival and service rates for each user layer in the fourth experiment

PU arrival rates	SU_1 arrival rate	SU_2 arrival rate	PU service rate	SU_1 service rate	SU_2 service rate
1/ 1.5/ 2/ 2.5/ 3/ 3.5/ 4	1	1	1	1	1

5.3.2 Results and Analysis for the Experiment

The configuration 1 ($R_1 = 0$ and $R_2 = 2$) enables channel reservation in the lowest priority user layer. For this configuration, it means that there are not *12 channels* (total) to be used, but $12 - 2 = 10$. Even though it is not capable of using the full network, it achieves the lowest blocking probability (Fig. 5.25) values since the other configurations 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 is able to use only $12 - 6 = 6$ channels (half of the total network resource). Moreover, configuration 4 achieves much lower blocking values compared to configuration 3. These inputs are differentiated just by the aggregation technique, which in configuration 3 is FCA [4,4] and in configuration 4 is DCA [1,4]. Again, we have noticed that the lower bandwidth bound of only 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.

Regarding the forced termination probability values in Fig. 5.26, we note a similar behavior as in the experiment 2 section, that is, as more channels are reserved from the SU_2, smaller values for the termination probability are experience. Naturally, the sequence from the greatest to the smallest values for this metric starts with input 1 followed by input 2 with $R_2 = 2$ *and* $R_2 = 4$, respectively. For inputs 3 and 4 much lower values can be observed, mainly because it uses multi-level channel reservation for denying some SU_1s but also because they have $R_2 = 6$. The interesting point here is when PU arrival is around 3.5, which enables a switch in performance, making the input 3 the configuration with the smallest forced termination probability value. This might seem unusual since input 4 uses DCA [1,4], but one should consider that input 4 has much lower blocking probability compared to input 3, therefore, much more SU_2s are accepted in the network, which at point 3.5 of Fig. 5.26 these users began to experience more service interruptions than those in input 3.

Fig. 5.25. SU_2 Blocking Probability

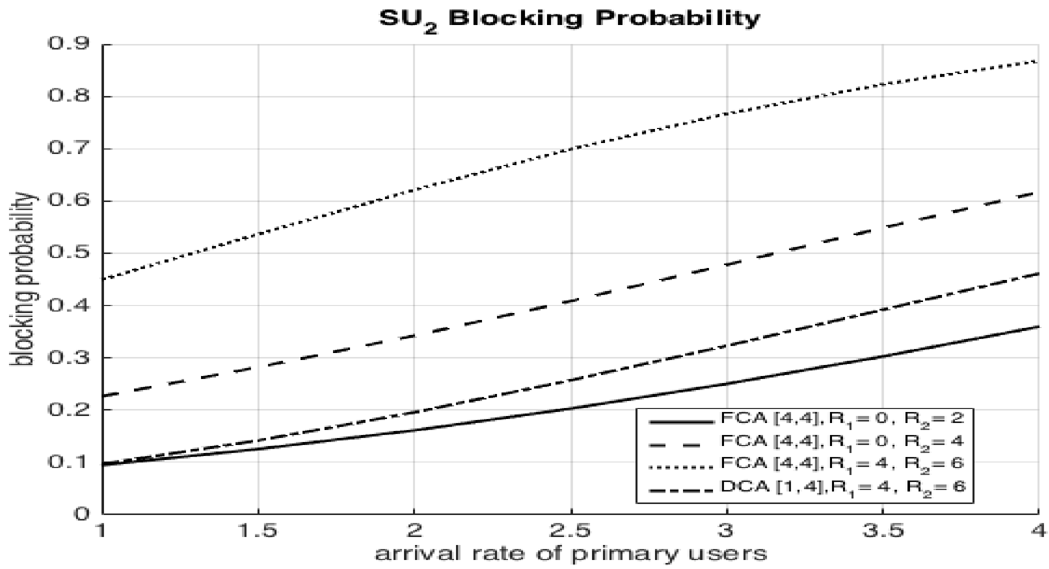
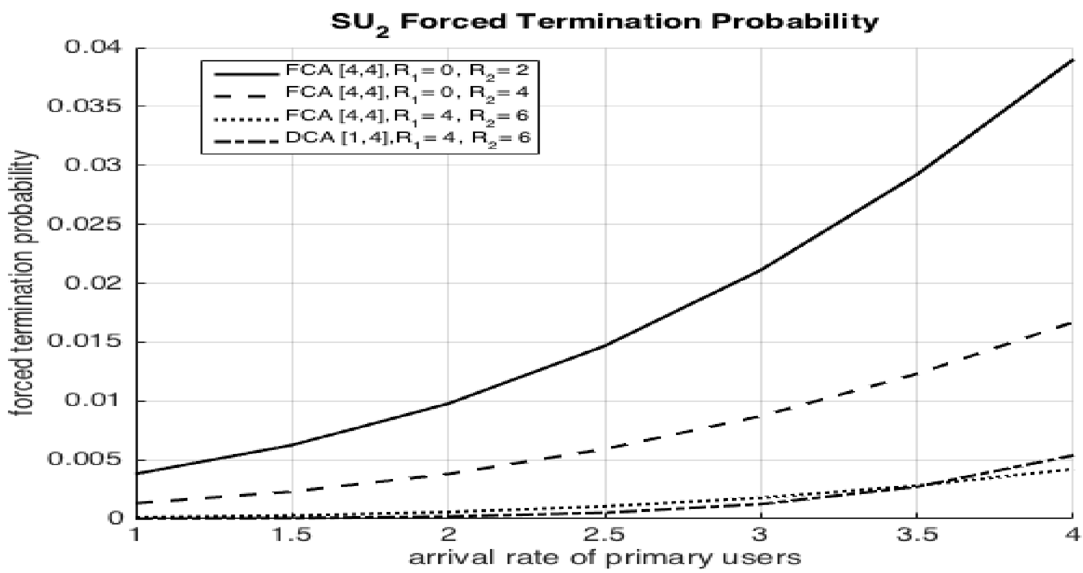


Fig. 5.26. SU_2 Forced Termination Probability



5.4 Chapter Summary

In this Chapter, the prioritized three-layered CRN model was tested in three experiments comparing the analytical and simulated results. The first aimed at showing that the model provided the correct outputs for any resource configuration, which comprises the different combinations of bandwidth sizes for the PU, SU_1 and SU_2. Then, we have demonstrated that the model is compliant with channel reservation and shown where it may or may not succeed to achieve reasonable performance while using this technique. Furthermore three channel aggregation options where compared and, based on our results, and in literature references, we have concluded that channel aggregation and fragmentation may be unfeasible, though, pure channel aggregation might be a reasonable option as their performances are similar. Finally, we have tested the joint utilization of channel reservation and aggregation in order to mitigate the negative effects of the first technique while benefiting from its properties.

Chapter 6 - Concluding Remarks

This chapter offers some final considerations about the work developed in this thesis by depicting its main contributions and proposing future studies.

6.1. Considerations

We have addressed the combination of cognitive radio, dynamic spectrum access techniques and traffic prioritization in a single three-layered heterogeneous cognitive radio network model in order to overcome the problem of resource underutilization likely to be the bottleneck for future wireless networks. Although this theme has drawn a great amount of attention, many authors seem to limit their models and analysis. Thus, we have outlined a more complete model from the resource allocation perspective, once our formulation allows multiple combinations of bandwidth sizes among primary and secondary users. Besides, we have proved by means of analytical modeling and simulation that the channel reservation approach not always provides reasonable performance tradeoffs for the secondary network, being required a study on the current network load before considering such approach. We have also tested an aggregation technique based on contiguous integer channel assembling that performed similarly to the channel aggregation and fragmentation, but is currently more feasible. Then, channel aggregation was applied together with channel reservation to prove how it can mitigate the negative blocking effects of this approach.

6.2. Future Works

As this work was developed, some ideas on how to expand its scope were discussed. In this section we will present three distinct future work lines.

The first ideas came from regular paper publications schemes such as some of those discussed in our related works chapter. For example:

- Provide an optimization alternative for finding the optimum number of reserved channels, maximum and minimum bandwidths considering the both PU and SU loads in order to obtain a better compromise between the performance metrics.
- Apply a different termination selection criteria, i.e., when user interruption is required, terminate those that are still far from finishing their services.
- Extend our model in order to contemplate many CRNs instead of only one.

The second involves the usage of newly acquired cognitive radio compliant lab equipment composed mainly by Universal Software Radio Peripherals (USRPs) to conceive part of the afore-mentioned experiments in a real test bed.

The third consists of modeling modifications to upgrade our three-layered prioritized CRN to an analytical framework, comprising the following features:

- To extend the model in order to cover any amount of user layers.
- Allow any combination of user bandwidths, completely without restrictions.
- Allow channel aggregation/fragmentation and channel reservation in any user layer (not only on the secondary network).

In the majority of the works (including this dissertation), it is assumed that the arrival flows of primary and secondary customers are stationary Poisson. The stationary Poisson arrival process is the simplest case of the well-known Markovian arrival process. Therefore, we might also upgrade our framework by means of allowing other probability distributions such as Hyper-Exponential and Erlang General, that can be considered to model HE/M/N/N and EG/M/N/N systems, in order to express different types of traffic, where the arrival process of users is not given by a Poisson distribution.

6.3 Summary of Contributions

The main contributions to the area of study provided by this thesis can be summarized as follows:

- Literature classification.
- Outlined and demonstrated formulation errors in two recent literature works (see Appendix A).
- Modeling of the three-layered prioritized CRN with a more flexible set of bandwidth inputs and formulation of important performance metrics related to the secondary communications.
- Formulation of a more flexible multi-level channel reservation technique for the secondary network (compared to the only solution found in (CHU et. al, 2015).

References

WEN, H.; TIWARY, P. K.; LE-NGOC, T. **Current trends and perspectives in wireless virtualization**, International Conference on Selected Topics in Mobile and Wireless Networking (MoWNeT), 2013a.

HASEGAWA, M. et. al. Optimization for Centralized and Decentralized Cognitive Radio Networks**, Proceedings of the IEEE**, v. 102, issue 4, p.574-584, 2014.

STAPLE, G. WERBACH, K. **The End of Spectrum Scarcity**. IEEE Spectr., v. 41, n. 3, p. 48-52, 2004.

AKYILDIZ, I. F. et. al. **Next generation/dynamic spectrum access/cognitive radio wireless networks: A survey**, Elsevier Computer Networks, v. 50, issue 13, p. 2127–2159, 2006.

MITOLA, J. **Cognitive radio architecture**. Hoboken, N.J.: Wiley-Interscience, 2006.

3GPP SPECIFICATION - **Requirements for further advancements for E-UTRA (LTE Advanced)**, 2013**.**

ZHU, X.; SHEN, L.; YUM, T. **Analysis of Cognitive Radio Spectrum Access with Optimal Channel Reservation**. IEEE Communications Letters, v. 11, n. 4, p. 304-306, 2007.

LI, L. et. al. **Combined channel aggregation and fragmentation strategy in cognitive radio networks**. arXiv preprint arXiv:1203.4913 (2012).

CHU, T.; ZEPERNICK, H.; PHAN, H. **Channel reservation for dynamic spectrum access of cognitive radio networks with prioritized traffic**. 2015 IEEE International Conference on Communication Workshop (ICCW), 2015

XIN, C.; SONG, M. **Dynamic Spectrum Access as a Service**, Proceedings IEEE INFOCOM, p.666-674, 2012

GOLDSMITH, A.; JAFAR, S.; MARIC, I.; SRINIVASA, S. **Breaking Spectrum Gridlock With Cognitive Radios: An Information Theoretic Perspective**. Proceedings of the IEEE, v. 97, n. 5, p. 894-914, 2009.

TSUI, T. CLARKSON, T. **Spread-spectrum communication techniques. Electronics & Communication Engineering Journal**, v. 6, n. 1, p. 3-12, 1994.

VAN VEEN, B. BUCKLEY, K. **Beamforming: a versatile approach to spatial filtering**. IEEE ASSP Magazine, v. 5, n. 2, p. 4-24, 1988.

NAJA, R. **Wireless Vehicular Networks for Car Collision Avoidance.** Springer, 2013.

IEEE 1609 - **Family of Standards for Wireless Access in Vehicular Environments (WAVE).**

LEE, H.; VAHID, S.; MOESSNER, K. **A Survey of Radio Resource Management for Spectrum Aggregation in LTE-Advanced**. IEEE Communications Surveys & Tutorials, v. 16, n. 2, p. 745-760, 2014.

STEWART, W. Probability, **Markov chains, queues, and simulation**. Princeton, N.J.: Princeton University Press, 2009.

BOLCH, G. et. al. **Queueing Networks and Markov Chains**, 2. ed, Wiley, 2006.

ROBINSON, S. **Simulation: the practice of model development and use**. Chichester, Eng.: Wiley, 2004.

COOPER, R. B. **Introduction to Queueing Theory**, 2. ed. North Holland, 1981.

SUN, B.; LEE, M.; DUDIN, S.; DUDIN, A. **Analysis of Multiserver Queueing System with Opportunistic Occupation and Reservation of Servers**. Mathematical Problems in Engineering, v. 2014, p. 1-13, 2014.

GROSS, D. et. al. **Fundamentals of Queueing Theory**, 4. Ed. Wiley, 2008.

BALIEIRO, A. **Cognitive radio virtual networks environment: definition, modeling and mapping of secondary virtual networks onto wireless substrate**. 2015. 105 p. Thesis (Doctorate in Computer Science) – Informatics Center, Federal University of Pernambuco, Recife. 2015.

JAIN, R. **The Art of Computer Systems Performance Analysis: Techniques for Experimental Design, Measurement, Simulation, and Modeling**, Wiley, 1991.

WHITT, W. **Continuous-Time Markov Chains**. 2012.

NORRIS, J. **Markov chains**. Cambridge, UK: Cambridge University Press, 1998.

CHU, T.; ZEPERNICK, H.; PHAN, H. **On the Performance of Underlay Cognitive Radio Networks Using M/G/1/K Queueing Model**. IEEE Communications Letters, v. 17, n. 5, p. 876-879, 2013.

WU, Y.; HU, F.; KUMAR, S.; GUO, M.; BAO, K. **Spectrum handoffs with mixed-priority queueing model over Cognitive Radio Networks**. 2013 IEEE Global Conference on Signal and Information Processing, 2013.

KHEDUN, N.; BASSOO, V. **Analysis of priority queueing with multichannel in cognitive radio network**. EUROCON, 2015.

JIAO, L.; BALAPUWADUGE, I.; LI, F.; PLA, V. **On the Performance of Channel Assembling and Fragmentation in Cognitive Radio Networks**. IEEE Transactions on Wireless Communications, v. 13, n. 10, p. 5661-5675, 2014.

CHU, T.; PHAN, H.; ZEPERNICK, H. **Dynamic Spectrum Access for Cognitive Radio Networks With Prioritized Traffics**. IEEE Communications Letters, v. 18, n. 7, p. 1218-1221, 2014.

JIAO, L.; PLA, V.; LI, F. **Analysis on channel bonding/aggregation for multi-channel cognitive radio networks**. 2010 European Wireless Conference (EW), 2010.

JIAO, L.; LI, F.; PLA, V. **Modeling and Performance Analysis of Channel Assembling in Multichannel Cognitive Radio Networks With Spectrum Adaptation**. IEEE Trans. Veh. Technol., v. 61, n. 6, p. 2686-2697, 2012.

MARTINEZ-BAUSET, J.; PLA, V.; PACHECO-PARAMO, D. **Comments on "analysis of cognitive radio spectrum access with optimal channel reservation"**. IEEE Communications Letters, 2010.

AHMED, W.; GAO, J.; SURAWEERA, H.; FAULKNER, M. **Comments on "Analysis of cognitive radio spectrum access with optimal channel reservation**. IEEE Transactions on Wireless Communications, v. 8, n. 9, p. 4488-4491, 2009.

JENG, J.; LIN, Y. **Equal Resource Sharing Scheduling for PCS data services**. Wireless Networks, v. 5, n. 1, p. 41-55, 1999.

HEREDIA-URETA, H.; CRUZ-PEREZ, F.; ORTIGOZA-GUERRERO, L. **Performance analysis of adaptive resource allocation strategies with service time dependence on the allocated bandwidth**. IEEE Wireless Communications and Networking, 2003.

MATLAB 2015b, **The MathWorks, Inc**., Natick, Massachusetts, United States. http://www.mathworks.com/products/matlab/ accessed on 2016.

CHHAJED D.; LOWE, T. **Building intuition**. New York: Springer, 2007.

YAFENG, W.; CHAO, L.; TIANWEI, W.; XIANG, W. **Dynamic Channel Reservation for Cognitive Radio Networks**. IEEE International Conference on Computational Intelligence & Communication Technology, 2015.

ALKHANSA, R.; ARTAIL, H.; GUTIERREZ-ESTEVEZ, D. **LTE-WiFi Carrier Aggregation for Future 5G Systems: A Feasibility Study and Research Challenges**. Computer Science, v. 34, p. 133-140, 2014.

Appendix A

In both works (CHU et al., 2014) and (CHU et al., 2015) there are erroneous transitions. In this section, we will demonstrate where are the mistakes and in Chapter 5 we will present an alternative form.

A.1 Transition Problems in (CHU et al., 2014)

In (CHU et al., 2014) there is a problem in part B item three of transition number one (Section Three - Markov Chain Model). In this transition, a number of j' highest priority SUs (SU_1s) and k' lowest priority SUs (SU_2s) should be forcibly terminated when the arrival PU enters the system. This situation may occur if there is no available free space for the arrival PU, even taking out all SU_2s (k=0). Therefore, it must drop all SU_2s and some SU_1s. The number of secondary users to be dropped are described as j' and k'.

The total number of sub bands is given by $M\ x\ N$. The bandwidth of each PU is N, while SU_1s and SU_2s have $N_1\ and\ N_2$ respectively. The error is in the formula for calculating j' that is: $0 \leq j' \leq\ ceil(N * (M - i) - j * N_1)/N_1)$.

By choosing $M = 2, N = 1, N_1 = 1\ and\ N_2 = 1$ we have a very small CRN (only two sub bands $M\ x\ N$) and all users occupy only one channel each. Therefore, there are many feasible states such as (0,0,0), (0,0,1), (0,0,2)… hence, for each state there should be one or more connections (transitions) to other feasible states. However, we have identified a problem in one of these connections, specifically in (1,1,0) → (2,0,0). This is an arrival PU event that encounters a full network and should drop one secondary user in order to free resources. In such case, it should use transition number one from item three (part B), but by replacing the values, j' does not assume the value of '1' as expected, instead it is: $0 \leq j' \leq\ (1 * (2 - 2) - 0 * 1)/1)$, which equals 0. Thus, because *i=2, j=0 and k =0,* by replacing *j' = 0* we get (1,0,0) → (2,0,0) and not the transition (1,1,0) → (2,0,0); so, we conclude that this part of the formulation is incorrect.

A.2 Transition Problems in (CHU et al., 2015)

A.2.1 First Problem in (CHU et al., 2015)

In (CHU et al., 2015) there is a problem in the same transition as in item (1) of this appendix. Because this work is an extension of (CHU et al., 2014), their formulations are very similar and consequently inherited the same problem. In the same section part B item three of transition number one (Section Three - Markov Chain Model) the calculations for j' are mistaken. Please refer to the previous item.

A.2.2 Second Problem in (CHU et al., 2015)

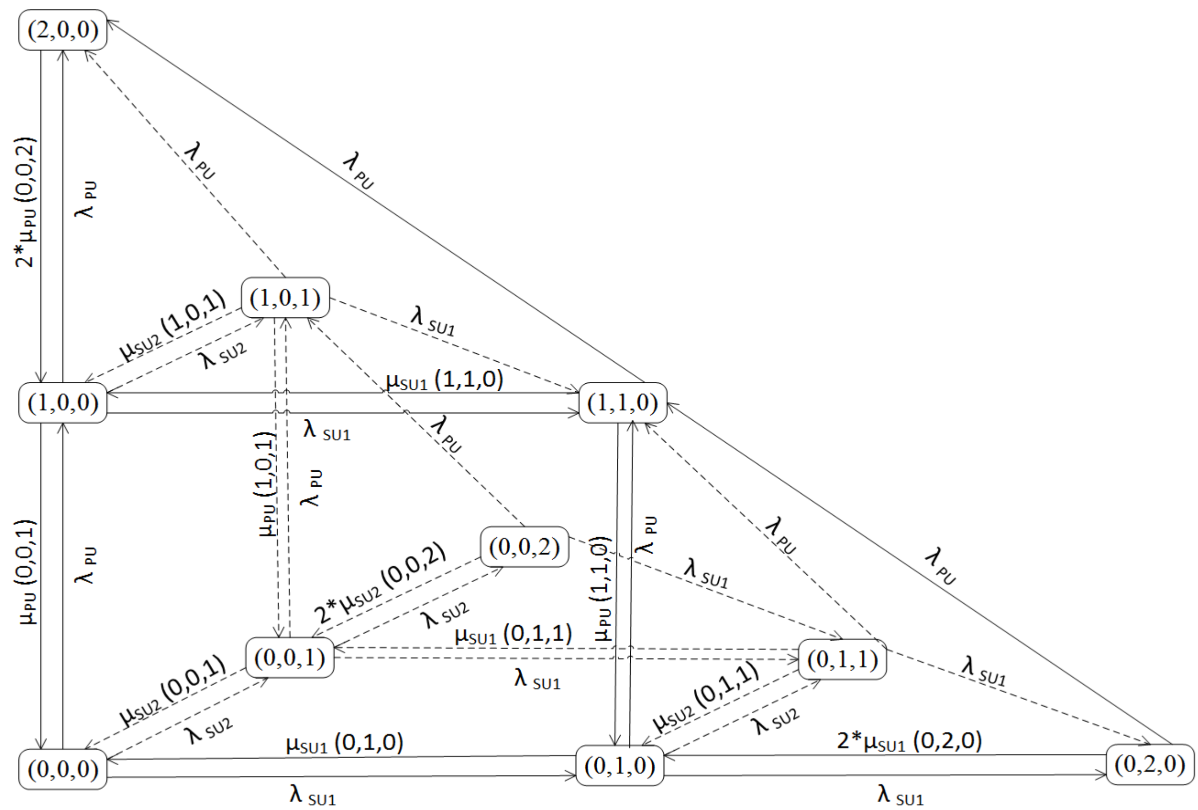
The last problem of (CHU et al. 2015) is not found in their previous work. This led us to conclude that there was a typo in the following transition: part B item two of transition number three (Section Three - Markov Chain Model), were a SU_1 drops a SU_2.

In this case, the number of dropped SU_2 (k') is the problem. The same parameters as in (1) may be chosen to prove this inconsistency, that is, $M = 2, N = 1, N_1 = 1 \, and \, N_2 = 1$. Let us chose the valid transition (0,0,2) → (0,1,1), which should be addressed by that part of the text. In such case, a number k' of SU_2s should be dropped. We know that k' should be equal to 1 in this case, but the wrong formulation calculates $0 \leq k' \leq$ min(ceil$\left(\frac{N1}{N2}\right), ceil$ $((N * (M - i) - j * N_1 - k * N_2)/N_1)$), which leads to $0 \leq k' \leq 0$. So, being *i=0, j=1 and k=1* the transition wrongly addressed is (0,0,1) → (0,1,1). We have corrected the afore-mentioned transitions in our model and written them in a more instructive manner, which can be found in Chapter 5 'System Model'.

Appendix B

This section offers a practical approach towards the solution of our system of equations, i.e., the steady state probability vector π. First, the example below will explore how to build the two parts that compose the system, namely: the balance equations and the normalized conditions; then, we will show how to join them in a single matrix structure and finally derive the vector π. We will utilize the same system as in Fig. 4.4 renamed here as Fig. B.1, which represents the state diagram for the input: $N = 2$, $B_{PU} = 1$, $B_{SU1} = 1$, $B_{SU2}^{min} = 1$, $R_1 = 0$ and $R_2 = 0$.

Fig.B.1 FCA example in a CRN with two channels and no channel reservation



1. Building the transition rate vector from state *(i,j,k)*

The first step of this example is to sum the transition rates departing from each state *(i,j,k)*. For example, considering the diagram Fig. B.4, there are three possible states departing from state *(0,0,0)*, therefore, there are three associated transition rates: $\lambda_{PU}, \lambda_{SU1}, \lambda_{SU2}$. This procedure is then repeated for every state and is contemplated by the first line of equation (25). Table B.1 expresses the result of this operation in a vector form named A. For better visualizing the next steps, we have used a variable to indicate each sum of rates.

$$\sum_{i'=0}^{\left\lfloor \frac{N}{B_{PU}} \right\rfloor} \sum_{j'=0}^{\left\lfloor \frac{N-R_1}{B_{SU1}} \right\rfloor} \sum_{k'=0}^{\left\lfloor \frac{N-R_2}{B_{SU2}^{min}} \right\rfloor} \pi(i,j,k) * \gamma_{(i,j,k)}^{(i',j',k')} * I(i,j,k) * I(i',j',k') =$$
$$\sum_{i'=0}^{\left\lfloor \frac{N}{B_{PU}} \right\rfloor} \sum_{j'=0}^{\left\lfloor \frac{N-R_1}{B_{SU1}} \right\rfloor} \sum_{k'=0}^{\left\lfloor \frac{N-R_2}{B_{SU2}^{min}} \right\rfloor} \pi(i',j',k') * \gamma_{(i',j',k')}^{(i,j,k)} * I(i,j,k) * I(i',j',k')$$
$$with\ 0 \le i \le, \left\lfloor \frac{N}{B_{PU}} \right\rfloor, 0 \le j \le \left\lfloor \frac{N-R_1}{B_{SU1}} \right\rfloor, 0 \le k \le \left\lfloor \frac{N-R_2}{B_{SU2}^{min}} \right\rfloor\ and$$
$$(i',j',k') \neq (i,j,k). \tag{25}$$

Table B.1 Vector A with the transition rates from state *(i,j,k)*

State	Transition Rates	Variable
(0,0,0)	$\lambda_{PU} + \lambda_{SU1} + \lambda_{SU2}$	a
(0,0,1)	$\lambda_{PU} + \lambda_{SU1} + \lambda_{SU2} + \mu_{SU2}$	b
(0,0,2)	$\lambda_{PU} + \lambda_{SU1} + 2 * \mu_{SU2}$	c
(0,1,0)	$\lambda_{PU} + \lambda_{SU1} + \lambda_{SU2} + \mu_{SU1}$	d
(0,1,1)	$\lambda_{PU} + \lambda_{SU1} + \mu_{SU1} + \mu_{SU2}$	e
(0,2,0)	$\lambda_{PU} + 2 * \mu_{SU1}$	f
(1,0,0)	$\lambda_{PU} + \lambda_{SU1} + \lambda_{SU2} + \mu_{PU}$	g
(1,0,1)	$\lambda_{PU} + \lambda_{SU1} + \mu_{PU} + \mu_{SU2}$	h
(1,1,0)	$\lambda_{PU} + \mu_{PU} + \mu_{SU1}$	i
(2,0,0)	$2 * \mu_{PU}$	j

2. Building the transition rate matrix arriving at state *(i,j,k)*

Now we will address the second line of equation (25) that represents all transition rates entering a state *(i,j,k)*. However, the transitions in use now are from Table 4.2 and thus cannot be expressed in by a vector, but in a matrix form. We have chosen the columns as the starting states and the lines as the ending states in our representation, for example, the transition departing state (0,0,1) and arriving at state (0,0,0) is located in line 1 column 2 and has the value μ_{2-SU}. This process is repeated for every state and the result is presented in Table B.2 and named matrix B.

Table B.2 Matrix B with the transition rates from other states to state *(i,j,k)*

	(0,0,0)	(0,0,1)	(0,0,2)	(0,1,0)	(0,1,1)	(0,2,0)	(1,0,0)	(1,0,1)	(1,1,0)	(2,0,0)
(0,0,0)	0	μ_{SU2}	0	μ_{SU1}	0	0	μ_{PU}	0	0	0
(0,0,1)	λ_{SU2}	0	2*μ_{SU2}	0	μ_{SU1}	0	0	μ_{PU}	0	0
(0,0,2)	0	λ_{SU2}	0	0	0	0	0	0	0	0
(0,1,0)	λ_{SU1}	0	0	0	μ_{SU2}	2* μ_{1-SU}	0	0	μ_{PU}	0
(0,1,1)	0	λ_{SU1}	λ_{SU1}	λ_{SU2}	0	0	0	0	0	0
(0,2,0)	0	0	0	λ_{SU1}	λ_{SU1}	0	0	0	0	0
(1,0,0)	λ_{PU}	0	0	0	0	0	0	μ_{SU2}	μ_{SU1}	2*μ_{PU}
(1,0,1)	0	λ_{PU}	λ_{PU}	0	0	0	λ_{SU2}	0	0	0
(1,1,0)	0	0	0	λ_{PU}	λ_{PU}	λ_{PU}	λ_{SU1}	λ_{SU1}	0	0
(2,0,0)	0	0	0	0	0	0	λ_{PU}	λ_{PU}	λ_{PU}	0

3. Building the generator matrix *Q*

Equation (25) says that for each state of a queueing network in equilibrium, the flux out of a state is equal to the flux into that state. This conservation of flow in steady state can also be written as $\pi * Q = 0$, where Q is the infinitesimal generator matrix of the CTMC, which in our example can be obtained by joining vector A and matrix B. Because π is multiplied by the vector and matrix respectively in lines one and two of equation (25), it is easy to view that an equation per row can be established, such as for row one indicated by Fig B.2.

Fig. B.2 First system equation considering equilibrium

$$\underbrace{\pi(0,0,0) * (\lambda_{PU} + \lambda_{1-SU} + \lambda_{2-SU})}_{\text{Flux out of state } (0,0,0)} = \underbrace{\pi(0,0,1) * \mu_{2-SU} + \pi(0,1,0) * \mu_{1-SU} + \pi(1,0,0) * \mu_{PU}}_{\text{Flux into state } (0,0,0)}$$

The process is then repeated for each state according to equation (25) and the result is a compact matrix *Q* depicted in Table B.3. We have moved each 'out of state' expression to the opposite side of (25), so, by using our vector variables (a, b, c…), we have placed them in the main diagonal of *Q*, with a negative sign.

Table B.3 Generator matrix *Q* in a compact form

(0,0,0)	(0,0,1)	(0,0,2)	(0,1,0)	(0,1,1)	(0,2,0)	(1,0,0)	(1,0,1)	(1,1,0)	(2,0,0)	-
-a	μ_{SU2}	0	μ_{SU1}	0	0	μ_{PU}	0	0	0	**0**
$\boldsymbol{\lambda_{SU2}}$	**-b**	2*μ_{SU2}	0	μ_{SU1}	0	0	μ_{PU}	0	0	**0**
0	λ_{SU2}	**-c**	0	0	0	0	0	0	0	**0**
$\boldsymbol{\lambda_{SU1}}$	0	0	**-d**	μ_{SU2}	2* μ_{SU1}	0	0	μ_{PU}	0	**0**
0	λ_{SU1}	λ_{SU1}	λ_{SU2}	**-e**	0	0	0	0	0	**0**
0	0	0	λ_{SU1}	λ_{SU1}	**-f**	0	0	0	0	**0**
$\boldsymbol{\lambda_{PU}}$	0	0	0	0	0	**-g**	μ_{SU2}	μ_{SU1}	2*μ_{PU}	**0**
0	λ_{PU}	λ_{PU}	0	0	0	λ_{SU2}	**-h**	0	0	**0**
0	0	0	λ_{PU}	λ_{PU}	λ_{PU}	λ_{SU2}	λ_{SU2}	**-i**	0	**0**
0	0	0	0	0	0	λ_{PU}	λ_{PU}	λ_{PU}	**-j**	**0**

4. Adding the normalized expression

The normalized expression indicates that the steady-state probability's sum equals to one, that is, $1 * \pi$ (0,0,0) + $1 * \pi$ (0,0,1) + … + $1 * \pi$ (2,0,0) = 1. This can be simply added to our system as a row (last row of Table B.4).

Table B.4 Matrix form system using the balance equations and the normalized expression

(0,0,0)	(0,0,1)	(0,0,2)	(0,1,0)	(0,1,1)	(0,2,0)	(1,0,0)	(1,0,1)	(1,1,0)	(2,0,0)	-
-a	μ_{SU2}	0	μ_{SU1}	0	0	μ_{PU}	0	0	0	**0**
$\boldsymbol{\lambda_{SU2}}$	**-b**	2*μ_{SU2}	0	μ_{SU1}	0	0	μ_{PU}	0	0	**0**
0	λ_{SU2}	**-c**	0	0	0	0	0	0	0	**0**
$\boldsymbol{\lambda_{SU1}}$	0	0	**-d**	μ_{SU2}	2* μ_{SU1}	0	0	μ_{PU}	0	**0**
0	λ_{SU1}	λ_{SU1}	λ_{SU2}	**-e**	0	0	0	0	0	**0**
0	0	0	λ_{SU1}	λ_{SU1}	**-f**	0	0	0	0	**0**
$\boldsymbol{\lambda_{PU}}$	0	0	0	0	0	**-g**	μ_{SU2}	μ_{SU1}	2*μ_{PU}	**0**
0	λ_{PU}	λ_{PU}	0	0	0	λ_{SU2}	**-h**	0	0	**0**
0	0	0	λ_{PU}	λ_{PU}	λ_{PU}	λ_{SU2}	λ_{SU2}	**-i**	0	**0**
0	0	0	0	0	0	λ_{PU}	λ_{PU}	λ_{PU}	**-j**	**0**
1	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**

5. Finding the steady-state probability vector π

Finally, MATLAB command 'linsolve' allows a quick solution to the aforementioned system using QR factorization with column pivoting. MATLAB uses the notation $Y * x = Z$, with Z being the last column of Table B.5, Y the rest of this table's content and x the steady-state probability vector π. Therefore, for each state there will be a number between zero and one, characterizing the probability of the system being in that state.

Table B.5 Matrix form of the system

(0,0,0)	(0,0,1)	(0,0,2)	(0,1,0)	(0,1,1)	(0,2,0)	(1,0,0)	(1,0,1)	(1,1,0)	(2,0,0)	-
-a	μ_{SU2}	0	μ_{SU1}	0	0	μ_{PU}	0	0	0	**0**
$\boldsymbol{\lambda_{SU2}}$	**-b**	2*μ_{SU2}	0	μ_{SU1}	0	0	μ_{PU}	0	0	**0**
0	λ_{SU2}	**-c**	0	0	0	0	0	0	0	**0**
$\boldsymbol{\lambda_{SU1}}$	0	0	**-d**	μ_{SU2}	2* μ_{SU1}	0	0	μ_{PU}	0	**0**
0	λ_{SU1}	λ_{SU1}	λ_{SU2}	**-e**	0	0	0	0	0	**0**
0	0	0	λ_{SU1}	λ_{SU1}	**-f**	0	0	0	0	**0**
$\boldsymbol{\lambda_{PU}}$	0	0	0	0	0	**-g**	μ_{SU2}	μ_{SU1}	2*μ_{PU}	**0**
0	λ_{PU}	λ_{PU}	0	0	0	λ_{SU2}	**-h**	0	0	**0**
0	0	0	λ_{PU}	λ_{PU}	λ_{PU}	λ_{SU2}	λ_{SU2}	**-i**	0	**0**
0	0	0	0	0	0	λ_{PU}	λ_{PU}	λ_{PU}	**-j**	**0**
1	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**

Appendix C

In this section, a brief study on the system's answers will be given based on the total number of states and the time to process the corresponding analytical model's output. Because the total number of states (*NoS*) is a function of the number of channels (*N*), the user's bandwidth (B_{PU}, B_{SU1} and B_{SU2}^{min}) and the reservation values (R_1 and R_2), there might be many ways to obtain its value that will depend on the relationship of these variables, e.g., if $B_{PU} > B_{SU1}$ or if $B_{SU1} > B_{PU}$. For this reason, we have opted for analyzing the *NoS* of the case studied in Appendix B, where $B_{PU} = B_{SU1} = B_{SU2}^{min} = 1$ and $R_1 = R_2 = 0$ (see Table C.1), but varying the total number of channels (*N*). Such example is the extreme case where the *NoS* achieves the highest value for our model as any other configuration where B_{PU}, B_{SU1} and $B_{SU2}^{min} > 1$ or R_1 and $R_2 > 0$ (maintaining the same number of channels) will result in fewer states (see Eq. 25 in Appendix B). A regular home machine was used for the following experiment (see Table C.2).

Table C.1 Input values for total state number demonstration

B_{PU}	B_{SU1}	R_1	R_2	$[B_{SU2}^{min}, B_{SU2}^{max}]$
1	1	0	0	[1,1]

Table C.2. Number of channels, states and total time for analytical outputs

Model Name:	MacBook Air
Model Identifier:	MacBookAir5,2
Processor Name:	Intel Core i5
Processor Speed:	1.8 GHz
Number of Processors:	1
Total Number of Cores:	2
L2 Cache (per Core):	256 KB
L3 Cache:	3 MB
Memory:	8 GB

For this example, the *NoS* is given by the expression 26.

$$NoS = 1 + \sum_{g=0}^{N-1} \sum_{h=0}^{N-g} (N - g - h + 1) \tag{26}$$

For the sake of simplicity, we evaluated six different values for the total channel number (5,10,15,20,25 and 30) and calculated the resulting number of states and the time for the system to process the analytical outputs in seconds as in Table C.3. Finally, we have plotted part of Table C.3 values in Fig. C.1 for better visualizing the exponential behavior of the total number of states as a function of the total channel number. Although the time to compute the outputs of the system also raises exponentially, this is not relevant for this model's context as its purpose is for network planning rather than operation, making this requirement more flexible.

Table C.3. Number of channels, states and total time for analytical outputs

Configuration nº	**Total nº of channels (*N*)**	**Total nº of states (*NoS*)**	**Time in (s)**
1	5	56	0.6764
2	10	286	1.8257
3	15	816	13.2067
4	20	1771	67.1957
5	25	3276	231.8704
6	30	5456	624.8751

Fig. C.1 Number of states as a function of the total channel number.