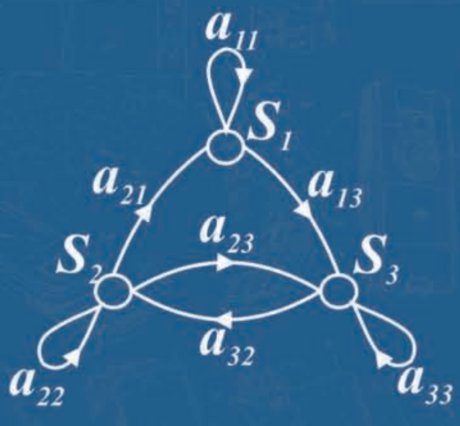
Departament de Teoria
del Senyal i Comunicacions

Contribution to Radio Resource and Spectrum Management Strategies in Wireless Access Networks: A Markov Modeling Approach



Doctoral Dissertation

Xavier Gelabert Doran

2010

Universitat Politècnica de Catalunya

UPC

Departament de Teoria
del Senyal i Comunicacions

UNIVERSITAT POLITÈCNICA DE CATALUNYA

GRCM

Grup de Recerca en Comunicacions Mòbils

(Mobile Communication Research Group)

Contribution to Radio Resource and Spectrum Management Strategies in Wireless Access Networks: A Markov Modeling Approach

A Thesis submitted in partial fulfillment of the requirements for the degree of doctor of Philosophy in the Departament de Teoria del Senyal i Comunicacions (TSC) Universitat Politècnica de Catalunya (UPC)

Xavier Gelabert Doran

Thesis Advisor

Dr. Oriol Sallent Roig

Barcelona, March 16, 2010

Contact information

✉ Xavier Gelabert Doran
Mobile Communication Research Group (GRCM)
Dept. Signal Theory and Communications (TSC)
Universitat Politècnica de Catalunya (UPC)
Jordi Girona 1-3, Campus Nord UPC
Building D4 Office 115
08034 – Barcelona, Spain

📧 xavier.gelabert@tsc.upc.edu

xavier.gelabert@ieee.org

🌍 http://www.tsc.upc.edu/xavier.gelabert.html

Andrei Andreyevich Markov (1856-1922)

To my family and friends.

This page intentionally left blank.

Preface

This thesis captures the work started back in September 2004, when I joined the Mobile Communication Research Group[1] (GRCM) as a PhD student, until the end of 2009. After an initial stage of PhD courses and preliminary assessment of radio resource management notions in wireless networks, the targeted scenario was set on heterogeneous wireless networks. A *hot-topic* then, and still of major relevance today as revealed by the large amount of research activity devoted to this matter. A first approach addressing the problems involved in this scenario was carried out using a simulation approach. After some time, in order to differentiate our work from existing and similar approaches, it was agreed that an analytical course (as opposed to simulation studies) would indeed bring substantial value to the research community. Accordingly, the use of Markov chains and stochastic modeling was adopted as a steering element throughout the research activity leading to the outcome of this dissertation. In addition, the increasing relevance of spectrum management strategies, through the growing adoption of cognitive radio concepts and methodologies, motivated to drive efforts towards this direction, again, by means of a Markovian analysis.

Acknowledgments

First and foremost I would like to thank my thesis advisor Dr. Oriol Sallent for his guidance along this research. In addition, I am thankful to Dr. Jordi Pérez-Romero and Prof. Ramon Agustí for their valuable and helpful assistance. During the course of my doctoral studies I had the chance to visit the Broadband Wireless Networking Lab. (BWN-Lab) at Georgia Institute of Technology. I am very thankful to Prof. Ian F. Akyildiz for his guidance and supportive attitude making my stay at the BWN-Lab an unforgettable experience. I had also the chance

[1] `http://www.tsc.upc.edu/grcm/`

to visit the Mobile Communications Group at the Instituto de Telecomunicaciones y Aplicaciones Multimedia (iTEAM) within the Universitat Politècnica de València (UPV). I am in deep gratitude with Prof. Narcís Cardona for his support. I would like to extend my gratitude to all past and present colleagues at GRCM-UPC, BWN-Lab and iTEAM, for the good times we shared over the past five years. Last, but not least, I am thankful to my family and friends for their constant support and encouragement.

Contents

This page intentionally left blank.

Summary

Current wireless networks exhibit heterogeneous multi-access features by means of the coexisting and cooperative deployment of several Radio Access Technologies (RATs). In this scenario, the provision of multimedia services with ensured Quality of Service (QoS) is mandatory. The overall goal of heterogeneous wireless access networks is to enable the realization of the *Always Best Connected* concept in which a user is seamlessly connected to the RAT best suiting its service requirements anytime, anywhere, anyhow. In this sense, Common Radio Resource Management (CRRM) strategies are devoted to provide an efficient utilization of radio resources within the heterogeneous network offering improved performances as opposed to performing stand-alone RRM in each RAT. In addition, allocated spectrum resources to each RAT must be efficiently utilized since it is a scarce and expensive resource. In this respect, cognitive radio concepts and methodologies have been applied to spectrum management by enabling dynamic/opportunistic spectrum sharing. In these scenarios, licensed spectrum is *opened* towards unlicensed access provided a non-harmful operation is guaranteed. This dissertation discusses both radio resource and spectrum management strategies to provide an utmost and efficient use of scarce radio/spectrum resources with the overall goal of maximizing user capacity while guaranteeing QoS constraints.

Specifically, the thesis is first focused on how to select an appropriate RAT upon call/session initiation (henceforth, *initial RAT selection*) in a heterogeneous access network. A Markovian framework is developed to such extent supporting the allocation of multiple service-type users (multi-service) on multiple RATs (multi-access). Under this framework, several RAT selection policies are proposed and evaluated, broadly categorized into service-based (SB) and load-balancing (LB). In addition, the performance of RAT selection policies in access-limited scenarios due to poor radio coverage, non multi-mode terminal availability and RAT-service incompatibility is also evaluated. Specific guiding principles for the allocation of services on several RATs are provided in the abovementioned scenarios with the overall goal of increasing user capacity while guaranteeing minimum QoS requirements. Finally, radio access congestion is also addressed in this multi-access/multi-service scenario and the impact RAT selection assessed. Suitable allocation principles avoiding congestion are also provided.

Secondly, this dissertation investigates on how to efficiently maximize the use of licensed spectrum by means of dynamic/opportunistic unlicensed spectrum access. Hereof, a Markovian framework is also devised to capture the problem of licensed spectrum sharing towards unlicensed users. A sensing-based spectrum awareness model is proposed in order to detect unused spectrum (so-called *white spaces*) which may be accessed by unlicensed users while remaining unused. Under this framework, the benefits of spectrum sharing are investigated and the involved gains assessed. Specifically, the sensing-throughput tradeoff and the adjustment of the sensing mechanism's operating point, which tradeoffs missed-detection and false-alarm errors, is evaluated. Moreover, fixed vs. adaptive spectrum channelization schemes are proposed and analyzed under two different service disciplines considering time-based and volume-based content delivery.

Resumen

Las redes inalámbricas actuales exhiben características heterogéneas de acceso múltiple mediante el despliegue, la coexistencia y la cooperación de varias Tecnologías de Acceso Radio (RAT[2]). En este escenario, la prestación de servicios multimedia garantizando una cierta calidad de servicio (QoS[3]) es obligatoria. El objetivo global de las redes heterogéneas de acceso inalámbrico consiste en sustentar la realización del concepto ABC (del inglés *Always Best Connected*), en el que un usuario está siempre conectado a la RAT que mejor satisface sus necesidades de servicio en cualquier momento, en cualquier lugar, de cualquier modo. En este sentido, las estratégias de gestión de recursos radio comunes [del inglés, *Common Radio Resource Management* (CRRM)] se diseñan para proporcionar una utilización eficiente de los recursos radio y de espectro radioeléctrico dentro de la red heterogénea, ofreciendo un mejor rendimiento en comparación con la realización independiente de RRM en cada RAT. Además, los recursos de espectro asignados a cada una de las RATs deben ser utilizado de manera eficiente, ya que se trata de un recurso escaso y costoso. En este sentido, conceptos y metodologías de radio cognitiva (del inglés *Cognitive Radio* o CR) se han aplicado a la gestión del espectro, permitiendo una compartición dinamico-oportunista del mismo. En estos casos, el espectro sujeto a licencia se *abre* hacia el acceso de usuarios sin licencia siempre que no perjudiquen y que el funcionamiento libre de interferencias esté garantizado. Esta tesis analiza estrategias de gestión de recursos radio y de espectro para ofrecer un uso mayor y eficiente de los escasos recursos radio y de espectro con el objetivo final de aumentar al máximo la capacidad de usuario, garantizando los requerimientos de QoS.

En concreto, estas tesis se centra primero en como seleccionar una RAT al inicio de una llamada/sesión (en adelante, selección inicial de RAT) en una red de acceso heterogénea. Un modelo de Markov ha sido desarrollado para definir la asignación de múltiples servicios (multi-servicio) en múltiples RATs (multi-acceso).

[2] Del inglés, *Radio Access Technology.*

[3] Del inglés, *Quality of Service.*

En este marco, varias políticas de selección de RAT son propuestas y evaluadas, genéricamente clasificándose en políticas basadas en servicio (SB[4]) y basadas en balanceo de carga (LB[5]). Además, el rendimiento de las políticas de selección de RATs en escenarios de acceso limitado debido a la deficiente cobertura radio, la falta de disponibilidad de terminales multi-modo y la incompatibilidad entre RAT y servicios también es evaluada. Principios específicos para la asignación de servicios a RATs serán provistos en los escenarios antes mencionados con el objetivo general de aumentar la capacidad de usuarios, garantizando los requisitos mínimos de calidad de servicio. Finalmente, la congestión en el acceso radio también se trata en este escenario multi-acceso/multi-servicio y el impacto de la selección de RAT evaluado. Los principios para la asignación inicial de RAT con tal de evitar la congestión radio serán también proporcionados.

En segundo lugar, esta tesis investiga sobre la forma de maximizar el uso eficiente del espectro sujeto a licencia (o licenciado) por medio del acceso dinámico-oportunista de espectro a usuarios sin licencia. En este sentido, se concibe un modelo de Markov para captar el problema del uso compartido de espectro entre usuarios con y sin licencia. Un modelo basado en sensado de espectro se propone con el fin de detectar porciones de espectro no utilizados (en inglés *white spaces*) que pueden ser usados por los usuarios sin licencia mientras este siga libre. En este marco, los beneficios obtenidos de la compartición del espectro son investigados y las ventajas que implican evaluadas. En concreto, se evalúa el rendimiento obtenido al ajustar el punto de funcionamiento (en inglés *operating point*) del mecanismo de sensado, el cual determina los errores de no-detección y falsa-alarma. Por otra parte, sistemas de canalización de espectro fijos versus adaptativos serán propuestos y analizados bajo dos disciplinas de servicio diferentes, cuya duración (o tiempo de permanencia en el sistema) esta basada en tiempo y en contenido respectivamente.

[4] Del inglés, *Service-Based*

[5] Del inglés, *Load Balancing*

Chapter 1

Introduction

1.1 Next Generation Wireless Access Networks

It is widely acknowledged, and well-assessed today, that current and next generation wireless networks encompass, among other features, the notion of network heterogeneity. That is, a plethora of different radio access technologies (RATs), each of them with distinctive and complementary characteristics, offering ubiquitous availability of multimedia services and seamless experience to QoS[1]-demanding users. Legacy RATs will co-exist and cooperate with new emerging technologies enabling the so-called *Always Best Connected* (ABC) paradigm [1]. The realization of the ABC concept implies that a user is seamlessly served through the RAT that "best" suits its service demands, according to some predefined criteria, throughout the whole duration of the service request. In this sense, Fig. 1.1 illustrates the ABC concept with a simple example. In Fig. 1.1(a) the user initiates a data-session in a home environment and WLAN is initially selected as opposed to 3G. As the user moves from the home environment towards the office environment WLAN can no longer sustain the data-session and at point (i) it is seamlessly transferred to the 3G network. In (b), the mobile user benefits from improved mobility management by the 3G network. Finally in (c), when the user is in the range of the corporate WLAN at point (ii), the data-session is transferred to the corporate WLAN.

Heterogeneous wireless network systems may consist of Wireless Local Area Networks (WLANs), Wireless Metropolitan Access Networks (WMANs) along with Wireless Wide Area Networks (WWANs). As for WLANs, they are primarily represented by the IEEE[2] 802.11 standard family offering high rates (up to 54Mbps) at low cost within a limited area [2]. The Worldwide Interoperability for Microwave Access (WiMAX) is an example of a WMAN based on the IEEE 802.16 standard which is able to provide enhanced data rates to those of WLANs while covering

[1] Quality of Service

[2] Institute of Electrical and Electronics Engineers.

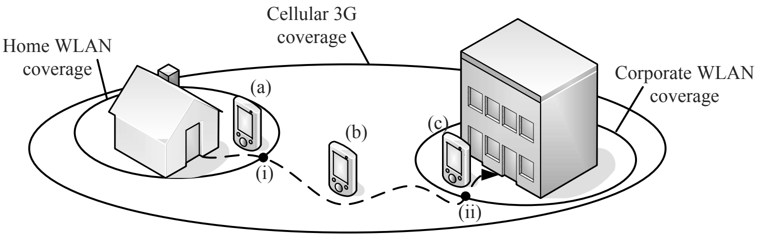


Figure 1.1: Mobility example illustrating the ABC concept.

larger areas [3]. Finally, WWANs comprise ubiquitous cellular mobile access technologies such as, for example, the Second Generation (2G) Global System for Mobile communications (GSM) [4], the 2.5G General Packet Radio Service (GPRS) [5] and the 3G Universal Mobile Telecommunications System (UMTS) [6]. In addition, enhancements to these systems have been developed such as the 2.75G Enhanced Data rates for GSM Evolution (EDGE) [7] and the 3.5G High Speed Packet Access (HSPA) [8]. Long-Term Evolution (LTE) [9–11] of these systems is expected to evolve into a 4G system providing up to 100 Mbit/s on the uplink and up to 1 Gbit/s on the downlink. These cellular systems provide wide coverage areas, full mobility and roaming, but traditionally offer low bandwidth connectivity and limited support for data traffic as compared to WLANs and WMANs (e.g. WiMAX). As a result, a heterogeneous network comprised of WLANs, WMANs and WWANs will provide complementary characteristics thus exhibiting higher flexibility, scalability, configurability, and interoperability, leading to an improvement with respect to traditional stand-alone communication systems.

The heterogeneous network concept becomes attractive for the network operator (NO) which can own several components of the network infrastructure (i.e., can own licenses for deploying and operating different RATs), and can also cooperate with affiliated NOs. Nonetheless, an NO can rely on several alternate radio networks and technologies for achieving the required capacity and QoS levels in a cost-efficient manner [12]. Then, the various RATs are thus used in a complementary manner rather than competing each other.

The deployment of wireless networks in general, and heterogeneous wireless networks in particular, has been traditionally facilitated and enhanced by means of concepts and solutions such as Radio Resource Management (RRM) and Spectrum Management (SM) in order to ensure efficient radio resource and spectrum utilization along with end-user service delivery with QoS constraints. Recently, the interest in RRM strategies has shifted towards SM methodologies propelled by the appearance of the Cognitive Radio (CR) concept [13].

In this scenario with multiplicity of RATs and new emerging technologies several challenging problems arise. Among others, the complexity involved in managing different radio resources corresponding to several RATs increases with the number of RATs. Thus, efficient RRM algorithms should be implemented in order to maximize and efficiently use the available resources offered by each RAT. The benefits of jointly considering all radio resources provided by all RATs, instead of managing each RAT as a stand-alone entity, is of particular interest as we will see next, and may lead to an improved system performance. There is also the concern that fixed (or licensed) spectrum allocation to each RAT results in a poor and ineffective utilization of scarce (and expensive) spectrum resources [14]. This spectrum underutilization and scarcity will be more evident as the number of RATs forming the heterogeneous network increases. In this case, SM algorithms and methodologies are required in order to improve the spectrum efficiency further. Among the SM techniques, CR methodologies have received much attention lately as a key technology enabling dynamic/flexible/opportunistic spectrum access between licensed and non-licensed users. Finally, the heterogeneous network scenario has motivated the need for multi-band/multi-standard devices so as to efficiently exploit the network heterogeneity by allowing connectivity to each of the available RATs. However, multi-standard devices offer only a short-term solution unable to cope with scalability issues and terminal re-design involved in these networks. In this sense, the notion of reconfigurability [15], which is an evolution of the Software-Defined Radio (SDR) concept [16], enables terminals and network elements to dynamically select and adapt to the most appropriate RAT for handling conditions rising in both time and space domains.

1.2 Managing Radio and Spectrum Resources in Wireless Access Networks: An Overview

As stated in the previous section, radio resource and spectrum management are key elements for the correct and efficient implementation and development of wireless access networks, in general, and heterogeneous wireless networks in particular. Next, some preliminary notions of both RRM and SM will be addressed.

1.2.1 Radio Resource Management

Radio Resource Management (RRM) involves strategies, algorithms and mechanisms for controlling radio transmission parameters in wireless communication systems in order to utilize the limited allocated radio spectrum resources and radio network infrastructure as efficiently as possible [17].

1.2.1.1 RRM in the Context of a Single RAT

An RRM framework is required for a wireless network to achieve the desired network and/or service requirements under the constraint on available radio resources. These radio resources may be in the form of frequency bands (or channels), time slots, spreading codes, transmission power, sub-carriers, etc. Some of the major functions of an RRM framework that arise in the context of a single RAT are [17], [18, §1.7]:

- Admission Control (AC): Decides whether a new connection should be accepted in order to guarantee minimum QoS requirements to both the new and ongoing connections. For a comprehensive survey on admission control schemes refer to [19, 20].
- Queue Management (QM): Refers to the actions taken over buffered packets prior to their transmission over the air interface. It is aimed at preventing packet dropping due to congestion situations which can be solved using active queueing management techniques [21, 22].
- Traffic Scheduling (TS): Responsible for selecting packets from the transmission queue for subsequent air interface delivery based on a pre-defined scheduling policy considering QoS and channel condition information [23].
- Medium Access Control (MAC): MAC protocols arbitrate the access of users to the shared radio channel. They target at detecting and avoiding packet collisions among users contending for channel access [24].
- Radio Resource Allocation (RRA): Aims at the efficient assignment of Radio Resource Units (RRUs), in the form of time-slots, frequency channels, spreading codes, modulation and coding schemes, power levels, etc., to a particular user so that some objective QoS constraint is met. Radio resource allocation principles are dependent on the underlying radio access scheme and may involve a broad number of physical parameters [25–28].
- Congestion/Traffic Control (CC): Manages situations in which the system has reached an overload status and therefore the QoS guarantees are at risk due to the evolution of system dynamics. Counter measures have to be taken to get the system back to a feasible load [29, 30].

- Handover Control (HOC): Handles the process of transferring an ongoing call or data session from one cell to another, that is, a horizontal handover (HHO) [31].

The above-listed RRM functionalities must not be regarded as individual and independent functions but, on the contrary, provide joint and overlapping functionalities. For example, HOC involves the AC functionality when a user moves from one cell to another; moreover, CC may imply resource re-allocation in order to lessen congestion.

1.2.1.2 RRM in the Context of a Heterogeneous Network: Common Radio Resource Management (CRRM)

The realization of the ABC concept in a heterogeneous network scenario where several RATs coexist calls for the introduction of new RRM strategies, operating from a common perspective, and taking into account the overall amount of resources in the available RATs. We refer to these as CRRM (Common[3] Radio Resource Management) algorithms and strategies [32]. In this scenario, the provision of radio resources can be seen as a problem with multiple dimensions, since every RAT is based on specific multiple access mechanisms exploiting in turn different orthogonal dimensions, such as, for example, frequency, time and code. Then, ”local” RRM mechanisms (as those bulleted in 1.2.1.1) are needed, see Fig. 1.2(a), in order to efficiently utilize the offered resources in each considered RAT. In addition to that, CRRM is based on the picture of a pool of radio resources, belonging to different RATs, but being managed in a coordinated way, see Fig. 1.2(b). The additional dimensions introduced by RAT multiplicity provides further flexibility in the way radio resources can be managed, leading to overall improvements provided by the resulting trunking gain [33]. In this case, we will foresee CRRM strategies, algorithms and mechanisms operating from a joint perspective on available resources. Notice that CRRM envisions the heterogeneous spatial distribution of radio resources yielding from different RAT spatial deployments. For example, GSM/EDGE tends to be the most widespread RAT, whereas HSPA or WiMAX may not be yet fully-deployed. In turn, WLAN *hotspots* provide a multiplicity of reduced coverage areas around specific locations.

As described in 1.2.1.1, the main RRM functions arising in the context of a single RAT are: admission control, congestion control, traffic scheduling, horizontal

[3] Sometimes referred to as Joint or Multiple.

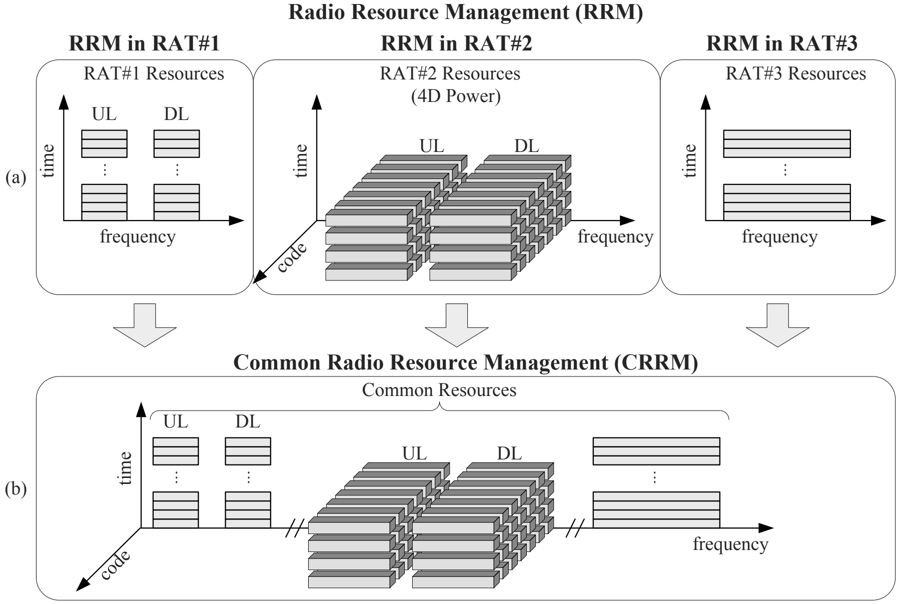


Figure 1.2: Management of radio resources. (a) stand-alone RRM for each RAT. (b) Common RRM (CRRM) over a pool of resources.

(intra-system) handover, etc. When these functionalities are coordinated between different RATs in a heterogeneous network scenario, they can be labeled as ”common” as long as algorithms take into account information about several RATs to make decisions. Then, we may have a Common Admission Control (CAC), Common Congestion Control (CCC), Common Traffic Scheduling (CTS), etc.

Additionally, when a heterogeneous scenario is considered, a specific functionality arises: the *RAT selection*[4] procedure. This functionality is devoted to decide the most appropriate RAT for a given demanding service at session initiation, known as initial RAT selection, or during an ongoing call/session, known as inter-system, inter-RAT or Vertical HandOver (VHO). Back to the simple example in Fig. 1.1, initial RAT selection at stage (a) will, for example, choose the home WLAN as opposed to 3G provided the enhanced data rates achieved by the former. When the user moves-out of the WLAN coverage region, at point (i) in Fig. 1.1, VHO mechanisms are triggered so that the call is transferred to 3G without service disruption. Once the user is in the range of the corporate WLAN, see Fig. 1.1 at point (ii), the network may decide to initiate another VHO in order to associate the call/session with the corporate WLAN.

[4] This term also appears in the literature, with the same meaning, as *network selection*, e.g. in [34–36], and *access selection*, e.g. in [37–39], among other term combinations.

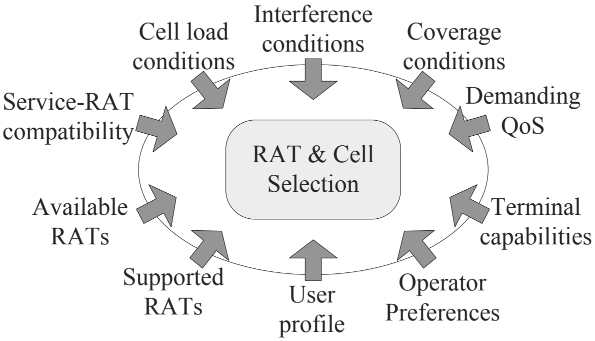


Figure 1.3: Factors influencing the RAT and cell selection.

It must be noted that RAT selection becomes a key CRRM element in order to exploit the flexibility offered by the available RATs in the heterogeneous network. The RAT selection procedure can be carried out considering different criteria (such as, service type, load conditions, etc.) with the final purpose of enhancing overall capacity, resource utilization and QoS. The heterogeneity scenario is also present in the customer side, where users with different multi-mode terminal capabilities co-exist providing connectivity to all or a subset of available RATs. In addition, different market segments can be identified, e.g. business vs consumer users, with their corresponding and distinctive QoS levels. Different RATs may co-exist in a given area, thus potentially exhibiting different overlapping coverage conditions and capabilities to support particular services. Then, selecting the proper RAT (and cell) is a complex problem due to the number of variables involved in the decision-making process, as reflected in Fig. 1.3 with some possible inputs. Furthermore, some of these variables may vary dynamically, which makes the process even more difficult to handle.

1.2.1.3 Architecture Supporting CRRM

A flexible architecture that allows interworking and cooperation across the considered RATs is needed to enable such coordinated use of resources. In this case, the level of interworking between network entities and the considered deployment topologies will, in great measure, impact the overall performance, as revealed in, e.g., [32, 40]. Depending on how data and signaling traffic is handled, if billing is commonly managed, and if radio resources are shared at the access layer, loosely or tightly coupled architectures arise. A range of characterizations of this kind have been proposed by 3GPP [41] and the European Telecommunications Standards Institute (ETSI) [42] among others. In general, the tighter the coupling, the

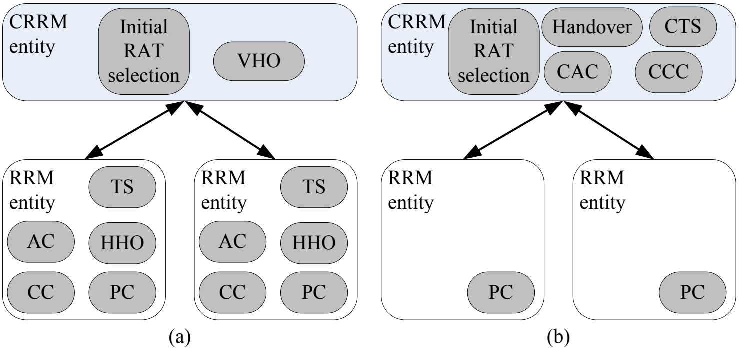


Figure 1.4: RRM/CRRM functionality split case: (a) CRRM having only long-term functionalities. (b) CRRM having long and short-term functionalities.

more flexible the possibilities of jointly exploiting the complementary features of composing RATs are. In [43, 44], the 3GPP considers two types of entities for the management of common radio resource pools, namely: the *RRM entity* and the *CRRM entity*. The RRM entity, which carries out the management of the resources in one radio resource pool of a certain RAT, usually resides in the Radio Network Controller (RNC) or the Base Station Controller (BSC). On the other hand, the CRRM entity executes the coordinated management of the resource pools controlled by different RRM entities, ensuring that the decisions of these RRM entities also take into account the resource availability in other RRM entities. The degree of interactions between the RRM and CRRM entities will determine the functionality split between them. Accordingly, Fig. 1.4 illustrates two possible functionality split cases between RRM and CRRM entities. In Fig. 1.4(a), so-called long-term functionalities reside at the CRRM entity, such as the initial RAT selection and the VHO functions. They are considered *long-term* given that, comparatively, they operate on a larger time-scale basis since initial RAT selection is carried out upon call/session establishment, and VHO is triggered on a RAN-boundary transit basis. Other RRM functions operating on a a shorter time-scale remain at the RRM entity. Alternatively, Fig. 1.4(b) reflects the case where short-term functionalities are implemented in the CRRM entity thus being re-defined as common admission control (CAC), common congestion control (CCC), common traffic scheduling (CTS), etc. In this case, *very-short-term* functionalities, such as power control (PC) still reside on the RRM entity. These interactions between RRM and CRRM entities mainly involve two types of functions: information reporting functions and RRM decision support functions, see Fig. 1.5. The interested reader is referred to [32, 45, 46] for a deeper explanation on the architectural procedures and practical implementations of CRRM.

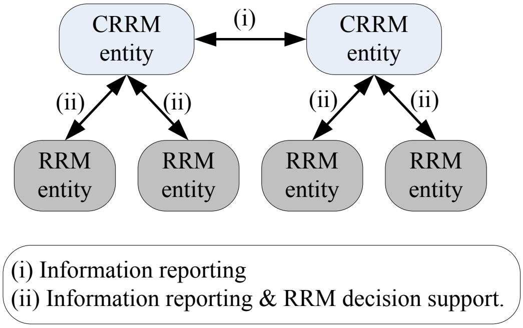


Figure 1.5: 3GPP View on RRM and CRRM functional entities: Interactions between RRM and CRRM entities.

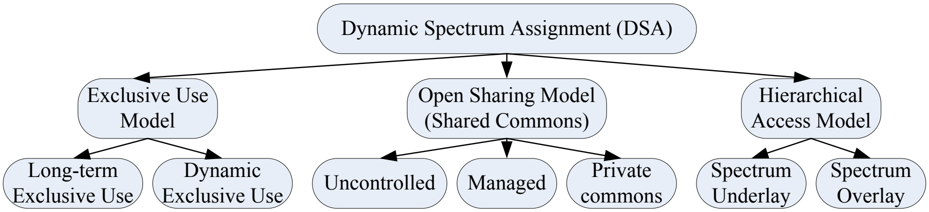


Figure 1.6: Dynamic Spectrum Access (DSA) Taxonomy.

1.2.2 Spectrum Management

In addition to RRM, Spectrum Management (SM) methodologies and techniques have been profusely investigated in the last years with Cognitive Radio (CR) receiving a huge attention. CR refers to a context-aware intelligent radio, built on a software radio platform, potentially capable of autonomous reconfiguration by learning from, and adapting to, the communication environment [47]. CR represents a broad paradigm stating that many aspects of communication systems can be improved via cognition. Among the many implementations of CR, Dynamic Spectrum Access (DSA) advocates for mechanisms and implementations providing a flexible access to spectrum resources leading to an improved efficiency as opposed to traditional Fixed Spectrum Access (FSA) [48]. DSA has also broad connotations that encompass various approaches to spectrum reform which can be generally categorized, see Fig. 1.6, under three different models [48, 49]: Exclusive Use Model, Open Sharing Model (also referred to as Shared Commons) and Hierarchical Access Model.

In the exclusive use model for spectrum access, the radio spectrum is licensed to a user/service to be exclusively used under a certain policy. In this model, the licenser (e.g. the government) allocates the spectrum to a licensee. If the licensee may not fully utilize the allocated spectrum in all times and in all locations it may grant spectrum access rights to cognitive radio users (i.e. unlicensed users). Constrained by the rules defined by the licensee, a cognitive radio user can optimize the spectrum usage to achieve the best performance. According to the time-scale granularity for spectrum trading, [48] identifies a *Long-term Exclusive Use* and a *Dynamic Exclusive Use*. In the former, radio spectrum is allocated exclusively to a licensed user/service provider for some period of time (e.g. a few weeks). In the dynamic exclusive use model, a spectrum owner can trade its owned spectrum to a cognitive radio user in a profitable way and on a finer time-scale. In addition, according to how this spectrum trading is carried out in the so-called *Secondary Market*, different sub-models appear for the dynamic exclusive use as described in [48].

In the Open Sharing Model [50], peer unlicensed users share the available spectrum in equal-access conditions. There are three variants of this model, namely, uncontrolled, managed, and private-commons sub-models [18].

Finally, the Hierarchical Access Model adopts a structure with Primary and Secondary (or unlicensed) users (i.e. PUs and SUs) where the licensed spectrum (owned by the primary network) is opened to SUs while limiting the interference perceived by PUs (or licensees). Secondary users operate within the secondary network, sometimes refereed to as the cognitive radio network. Two approaches to spectrum sharing between PUs and SUs are mainly considered [48, 49]: Spectrum underlay and spectrum overlay. For the spectrum underlay case, a SU can transmit concurrently with PUs provided the SU transmit power is limited so that the interference caused to the PUs remains below the interference temperature limit [51, 52]. Spectrum underlay can be used for cognitive radio systems using CDMA[5] or UWB[6] technology, see e.g. [53, 54]. As for spectrum overlay, it does not necessarily impose severe restrictions on the transmission power of SUs, but rather on when and where they may transmit. It directly targets at spatial and temporal spectrum *white spaces* by allowing SUs to identify and exploit local and instantaneous spectrum availability in a non-intrusive manner, see e.g. [55, 56] among others.

1.2.2.1 Dynamic Spectrum Access Architecture

The architecture of a cognitive radio network supporting DSA can be either infrastructure-based or infrastructure-less [14]. In the former case, each unlicensed user

[5] Code Division Multiple Access

[6] Utra Wide Band

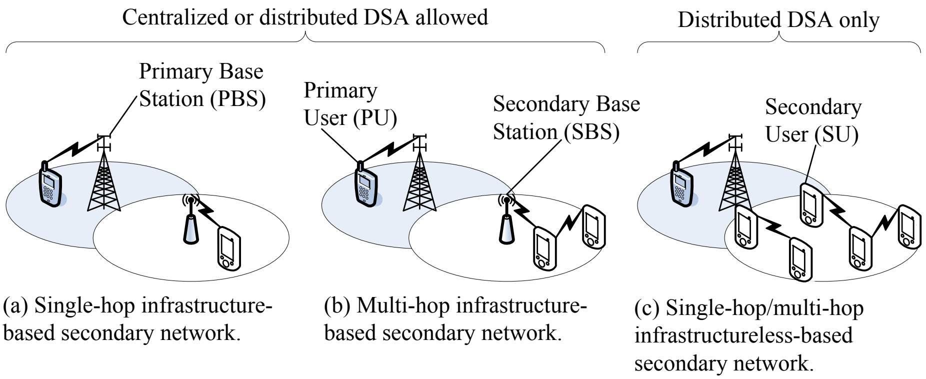

Centralized or distributed DSA allowed
Distributed DSA only
Primary Base Station (PBS)
Primary User (PU)
Secondary Base Station (SBS)
Secondary User (SU)
(a) Single-hop infrastructure-based secondary network.
(b) Multi-hop infrastructure-based secondary network.
(c) Single-hop/multi-hop infrastructureless-based secondary network.

Figure 1.7: Architecture supporting DSA.

transmits to a base-station (BS), if single-hop, or between multiple relaying BSs if multi-hop communication is required [see Fig. 1.7(a) and Fig. 1.7(b)]. In the case of infrastructure-less, or *ad-hoc*, architecture, communication among unlicensed users is on a direct peer-to-peer basis or, alternatively, through other unlicensed users acting as relays, see Fig. 1.7(c).

According to the entity which is responsible to manage spectrum access decisions, cognitive radio networks can be either centralized or distributed [14, 18]. In a centralized architecture, see e.g. [55, 57, 58], the decision of spectrum access is made by a central controller, while in a distributed architecture, [51, 59–61], this decision is made locally by each of the unlicensed users. In general, a centralized DSA requires an infrastructure-based architecture whereas the distributed DSA could be both implemented using infrastructure and ad-hoc architectures [18].

1.2.2.2 Spectrum Awareness

In order to access the radio spectrum in an opportunistic manner, it must be ensured that such spectrum is not being occupied by any licensed user. That is, the secondary system must obtain the current spectrum use pattern showing which frequencies are occupied and which frequencies are available for use in a band of interest at a particular geographic location and a particular time. Several approaches are proposed to identify spectrum opportunities, mainly categorized into *passive awareness* and *active awareness* methods [61].

With passive awareness, primary spectrum use is received outside from one's own secondary communication system. Secondary systems may be informed in real-time, during operation, or in advance. Several passive-awareness alternatives exist.

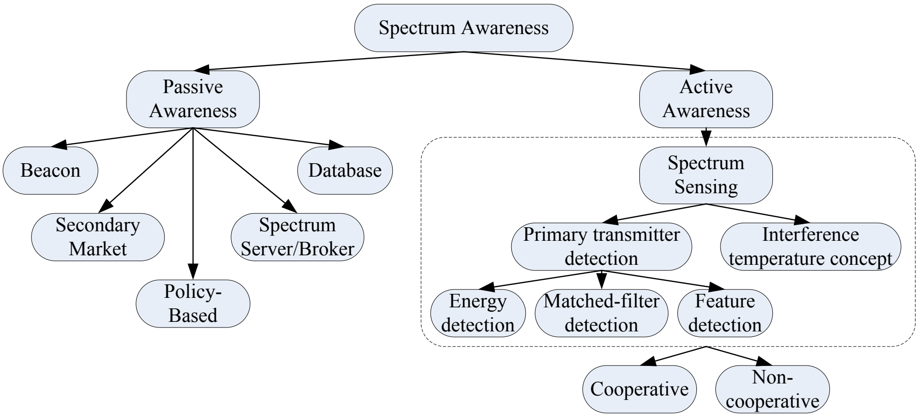


Figure 1.8: Spectrum awareness for DSA.

Secondary users (i.e. unlicensed users) may negotiate spectrum resources with primary users [62] in the secondary market-place [63, 64]. Moreover, the spectrum use pattern can be obtained also from a centralized server (a.k.a. broker server) [57, 65] or, alternatively, by maintaining an up-to-date database [66]. In addition, in a policy-based approach, the primary system use is defined a priori [67].

As for active awareness, secondary users ”actively” sense the surrounding radio environment and adapt their transmission based on the measurements. This can be done in a non-cooperative manner, where nodes decide independently based on their own spectrum observations; or cooperatively, [68], where local measurements are combined before decisions about spectrum use are made. In either way, several spectrum sensing techniques are devised, [69], mainly categorized into two main types, primary transmitter detection and interference temperature concept [14]. Among the primary transmitter detection methods, *energy detection*, *matched filter detection*, and *cyclostationary feature detection* are the ones that have received most attention (see [69] for further details on these methods).

1.3 Problem Formulation, Motivation and Scope

The goal of this dissertation can be summarized as obtaining the utmost and efficient utilization of radio and spectrum resources in a multi-access/multi-service heterogeneous wireless network scenario. By utmost and efficient radio resource and spectrum usage we specifically target at algorithms and mechanisms in order to maximize the overall *user capacity*, defined as the number of users that can be handled in the network while guaranteeing minimum QoS constraints[7]. It is then in the interest of a network operator to maximize the number of users it may handle with satisfactory QoS in order to maximize revenue.

Several methods to improve user capacity exist, ranging from physical to application layer improvements, along with network planning and deployment enhancements. This dissertation focuses on user capacity improvements resulting from the allocation of radio and spectrum resources to different service/user-types. Specifically, each service-type (e.g. voice call, web-browsing session, etc) or user-type (e.g. licensed vs. unlicensed) is assumed to occupy an amount of radio resources. These radio resources depend on the considered underlying access technology, e.g., time-slots in a TDMA[8]-based access, spreading-codes in a CDMA-based access, etc. Additionally, scenarios where spectrum bands (or channels) are shared between licensed and unlicensed users also require efficient spectrum resource management in order to guarantee the licensed user an interference-free operation.

Among the different and various problems arising in the depicted scenario, this thesis is concerned with two problems, namely *P1* and *P2*, along with corresponding sub-problems, as detailed in the following.

With this focus, the first dissertation problem may be formulated as:

P1) *How to select the appropriate RAT for a specific service in a multi-access network such that QoS requirements are met and the overall user capacity is maximized.*

Specifically, this thesis addresses the problem of RAT selection (recall from Section 1.2.1.2) upon call/session establishment, that is, initial RAT selection. In order to assess the performance of this procedure, the scope of this dissertation is limited to this initial RAT selection procedure solely, thus neglecting the operation of an inter-RAT (or vertical) handover. Although considering VHO capabilities would be certainly interesting in order to capture the whole dynamics of the system, it would however obscure the purpose of our study by hiding the effect of initial RAT selection. This interest is motivated by the importance of efficiently designing the initial RAT selection procedure as a first step in the realization of the ABC concept. In addition, this dissertation considers the heterogeneous multi-access network is managed by a single network operator, thus disregarding inter-operator resource

[7] In this thesis, the term *capacity* will be referred in this sense, despite other meanings adopted in the literature.

[8] Time Division Multiple Access

management situations. The heterogeneous network scenarios comprised in this thesis are conformed by wide-spread deployed and ubiquitous GSM/EDGE along with UMTS and WLAN technologies. Nonetheless, proposed methods, algorithms and approaches can be easily extended to include other emerging technologies including, e.g., HSPA, WiMAX, and LTE.

Problem *P1* may in turn be split into several sub-problems of interest which will be also treated in this thesis, as listed hereafter:

P1.1) *How do access selection impairments affect the initial RAT selection procedure.*

P1.2) *How does initial RAT selection influence the radio access congestion of a heterogeneous wireless network.*

Sub-problem *P1.1* relates to accessibility constraints within the initial RAT selection procedure. These constraints may be in the form of coverage availability, mobile terminal capabilities and RAT-service support, among others, thus limiting the options when selecting a RAT. This thesis provides specific rules for access-constrained initial RAT selection. As for sub-problem *P1.2*, it relates to the impact of the RAT selection procedure on potential overload, or congestion, situations that may be reached due to the inherent dynamics of the system. Congestion poses at risk the QoS guarantees in a particular RAT and, more generally, in the whole heterogeneous network. It is therefore important to be able to identify these congestion situations and implement RAT selection procedures such that the occurrence of congestion is minimized.

The second problem addressed in this dissertation can be expressed as:

P2) *How can licensed spectrum utilization be efficiently maximized by means of opportunistic and non-interfering unlicensed use.*

Problem *P2* relates to the Hierarchical Access Model, presented in Section 1.2.2, as a way to accomplish Dynamic Spectrum Access (DSA). This thesis is concerned with the scenario of a primary (i.e. licensed) system which *opens* its spectrum for secondary (i.e. unlicensed) and non-disruptive use. It is further considered that spectrum awareness on the secondary system is implemented via spectrum sensing mechanisms. Spectrum resources are commonly characterized by a partition of bandwidth units dividing the whole spectrum licensed to the primary system.

Problem *P2* may also be divided into several sub-problems of concern in this thesis:

P2.1) *How do spectrum sensing mechanisms affect the operation of primary and secondary systems such that minimum QoS is met.*

P2.2) *How can the secondary user adapt its spectrum requirements according to the demanding service type and spectrum availability.*

Sub-problem *P2.1* is concerned with the way spectrum sensing errors affect the operation of primary users. These spectrum errors, mainly characterized by *false-alarm* and *missed-detection* probabilities, have an opposite behavior. Whereas false-alarm prevents secondary spectrum usage, missed-detection causes spectrum collisions, and thus interference, between primary and secondary users. Unfortunately, both low false-alarm and missed-detection cannot be achieved simultaneously, therefore a trade-off appears in the design of the spectrum sensing mechanism. In this thesis, the *energy detector* (recall from Section 1.2.2.2) will be adopted for spectrum sensing tasks in a non-cooperative fashion.

On the other hand, sub-problem *P2.2* addresses the case where secondary users may flexibly decide the amount of spectrum they wish to access as long as interference with primary users is avoided. The larger the spectrum available for a secondary user the higher the bit-rates this user may achieve. According to the demanding service-type, whether its duration is sensitive or insensitive to the achieved bit-rate (i.e. elastic or inelastic), spectrum flexibility will impact the performance of the system. It is therefore desirable to identify flexible spectrum access mechanisms in order to maximize and efficiently utilize the available spectrum on a non-interfering basis.

The overall motivation of this dissertation through the study of the aforementioned problems resides in the fact that network heterogeneity, provided by multi-access and multi-service characteristics, is a common feature of current wireless networks and will be of future wireless networks too. In this scenario, the need for optimizing the use of radio resources within a pre-defined and fixed spectrum range will still be a concern. Nonetheless, spectrum scarcity and underutilization will have to be addressed by implementing efficient spectrum management techniques achieving a higher and more profitable spectrum use. The increase of user capacity as a final goal to be reached by the proposed methods further motivates the focus of this dissertation.

1.4 Adopted Approach

In this dissertation, the common approach for tackling problems *P1* and *P2* will rely on Markov modeling. Markov chains have been a profusely used tool in modeling problems related with resource allocation in the scope of telecommunications in general, and wireless networks in particular. It presents a mature and well-developed theory, well-recognized and used by the research community. It has also its limitations, as any other mathematical model, which will be discussed in the context of each contribution in this thesis. The reader who is unfamiliar with the theory of Markov chains may find in [70] a good starting point for further understanding on this topic. Additionally, several references addressing the theory of Markov chains in the context of communication networks are provided in [71–73], among others, and references therein.

This thesis will be concerned with so-called *Erlang Loss Systems*, in which a user request not finding a resource will abandon the system without further allocation attempts, that is, a blocked call is cleared. This model is known to fit the nature of voice-call (or circuit-switched) requests and, for the sake of simplicity, it will be further applied to data sessions as well. In addition, the widely-adopted Markovian assumption will hold by considering that arrival process follows a Poisson distribution and that the holding times are exponentially distributed.

Specifically, for the evaluation of initial RAT selection (i.e. problem *P1*), the allocation of resources in the selected RAT is made upon service arrival. Then, we are mostly concerned with the observation of arrival processes. In this case, we adopt the Continuous Time Markov Chain (CTMC) model, which by definition observes the system-state upon arrivals and departures, and evoke the *splitting property* [70] of a Poisson process when selecting the appropriate RAT.

As for the spectrum sharing problem (i.e. problem *P2*), periodical spectrum sensing implementation calls for the adoption of a Discrete Time Markov Chain (DTMC) approach. In this case, the state of the system is observed periodically, which suits the nature of the problem under study.

In either CTMC and/or DTMC approaches, solving Markov chains can be reduced to numerically solve a linear system of equations. In this dissertation, numerical method solutions will be applied to solve these systems. Specifically, iterative methods will be considered. The reader is referred to [74, 75] for a throughout overview and deep explanation of numerical methods for the resolution of Markov chains.

1.5 Dissertation Contributions

With the focus discussed in the previous sections, the aim of this thesis is to provide solutions to the aforementioned problems and sub-problems with the adopted Markovian approach.

With respect to problem *P1*, and related sub-problems, the following contributions are made in this dissertation:

P1.C1) Development of a Markovian framework for the evaluation of RAT selection policies in multi-access/multi-service scenarios.

P1.C2) Characterization of the service-RAT compatibility indicating RATs that do not uphold particular services.

P1.C3) Definition of a probabilistic coverage model enabling the characterization of any overlapping coverage situation among the considered RATs.

P1.C4) Definition of a probabilistic model capturing the availability of multi-mode terminals which support all or a subset of available RATs.

P1.C5) Description of several heuristic RAT selection policies, being categorized into: service-based selection and load balancing selection. Specifically, the service-based policies are defined in a two-service case, with generic voice and data services, in multi-access scenarios comprising 2 and 3 RATs.

P1.C6) Identification the main parameters that influence the performance of the proposed RAT selection policies, mainly: the ratio of offered service traffic (or service-mix), the constrained access for each service request, and the effect of resource-contention between different services in a prticular RAT.

P1.C7) Provisioning of guiding principles and rules for the suitable allocation of voice and data services in constrained multi-access networks.

P1.C8) Definition of a probabilistic model for the characterization of both uplink and downlink radio access congestion situations in heterogeneous networks. The model is evaluated for a particular multi-RAT scenario, considering TDMA and WCDMA[9] access networks, along with generic voice and data services.

[9] Wideband CDMA

P1.C9) Description of a novel initial RAT-selection policy that takes advantage of congestion information provided by the probabilistic model mentioned in *P1.C8*.

Turning to problem *P2*, and related sub-problems, the following contributions are provided in this dissertation:

P2.C1) Development of an analytical framework based on a Discrete Time Markov Chain (DTMC) model for the evaluation of sensing-based secondary opportunistic spectrum access scenarios.

P2.C2) Definition of a probabilistic spectrum sensing model accounting for potential sensing errors in the form of false-alarm and missed-detection. This model allows to use well-known expressions in the literature concerned with such error probabilities under several channel propagation conditions.

P2.C3) Model validation and evaluation studies considering several parameter dependency issues and tradeoffs in order to justify the usefulness of the proposed model for cognitive radio networks system design, realization and operation.

P2.C4) Identification of key parameters influencing the performance of the spectrum sharing model. In particular, sensing periodicity (*how often do we sense?*) and sensing accuracy (*how well do we sense?*). In addition, contribution towards the importance of time-sharing between spectrum sensing (*for how long do we sense?*) and data transmission (*for how long do we transmit?*) is also provided.

P2.C5) Definition of several alternatives for the partition, or channelization, of the available spectrum in order to provide an efficient access and spectrum utilization for both primary and secondary users. Specifically, a *Fixed Channelization Scheme* (FCS) and an *Adaptive Channelization Scheme* (ACS) are proposed as two possible secondary spectrum access mechanisms.

P2.C6) Characterization of secondary service requests. Firstly, *Time-Based Services* (TBSs) aim for the use of a particular amount of bandwidth for a given time. Secondly, *Volume-Based Services* (VBSs) aim at transmitting a given amount of data so that the service duration depends on the achievable bitrate, i.e. on the amount of spectrum bandwidth assigned to each user.

P2.C7) Assessment of the potential gains that can be achieved by correctly selecting the sensing operating point which determines a particular value of the false-alarm and missed-detection probabilities. The suitability of the sensing operation points is determined using the Grade-of-Service (GoS) concept conveniently adopted from ”traditional” telephone networks.

1.5.1 Related Publications

Bearing in mind the above-listed contributions, this section lists the publications by the author which this dissertation is largely based on. Each publication will be linked to one or various contributions listed in the previous section.

This thesis is based on the following journal articles (JAs) and conference articles (CAs), listed hereafter in the order that they will appear in the manuscript.

Regarding problem *P1*, the included publications are:

[JA1] *X. Gelabert, J. Pérez-Romero, O. Sallent, R. Agustí, "A Markovian Approach to Radio Access Technology Selection in Heterogeneous Multiaccess/ Multiservice Wireless Networks," IEEE Transactions on Mobile Computing, Oct. 2008* [76]. Partially includes contribution *P1.C1* by presenting the definition of the Markovian framework for the specific case of two services (voice and data) along with two RATs, GSM/EDGE and UMTS. It also includes contribution *P1.C5* for the specific case of two services and two RATs along with contribution *P1.C6.* Multi-mode terminal availability is also partially covered in this scenario, thus also adhering to contribution *P1.C7.*

[CA1] *X. Gelabert, J. Pérez-Romero, O. Sallent, R. Agustí, "A 4-Dimensional Markov Model for the Evaluation of Radio Access Technology Selection Strategies in Multiservice Scenarios", IEEE Vehicular Technology Conference Fall 2006 (VTC-Fall'06)* [77]. Contains initial results and model formulation for the RAT selection problem in a voice/data GSM-EDGE/UMTS heterogeneous network. Journal [JA1] is an extended version of this article.

[JA2] *X. Gelabert, O. Sallent, J. Pérez-Romero, and R. Agustí, "Performance Evaluation of Radio Access Technology Selection Strategies in Constrained Multi-Access/Multi-Service Wireless Networks", Submitted for possible journal publication, Dec. 2009* [78]. Includes contribution *P1.C1* by presenting a generalized Markov framework accounting for any number of services, RATs and RAT-service support issues (a differentiating aspect from [JA1]). Publication [JA2] also contains contributions ranging from *P1.C2* to *P1.C7.* It assumes three RATs as opposed to only two in [JA1].

[CA2] *X. Gelabert, O. Sallent, J. Pérez-Romero, R. Agustí, "Erlang Capacity Degradation in Multi-Access/Multi-service Wireless Networks due to Terminal/Coverage Restrictions", in Int. Symposium on Wireless Personal Multimedia Communications 2008 (WPMC'08)* [79]. Contains initial contributions towards the work presented in [JA2] related to access limitations in the RAT selection procedure.

[JA3] *X. Gelabert, O. Sallent, J. Pérez-Romero, and R. Agustí, "Radio Access Congestion in Multiaccess/Multiservice Wireless Networks," IEEE Transactions on Vehicular Technology, May 2009* [80]. Contains contribution *P1.C1* as in [JA2]. As in [JA1], it considers two services (voice and data) along with two RATs (generically, TDMA and WCDMA). Contribution *P1.C5* is also addressed. Nonetheless, publication [JA3] mainly focuses on contributions *P1.C8* and *P1.C9* concerning radio access congestion.

[CA3] *X. Gelabert, J. Pérez-Romero, O. Sallent, R. Agustí, "Evaluation of Radio Access Congestion in Heterogeneous Wireless Access Networks", IEEE Global Communications Conference 2008 (GLOBECOM'08)* [81]. Contains initial evaluation of radio access congestion in TDMA/WCDMA scenarios as in [JA3].

Regarding problem *P2*, the related publications by the author are:

[JA4] *X. Gelabert, O. Sallent, J. Pérez-Romero, and R. Agustí, "Spectrum Sharing in Cognitive Radio Networks with Imperfect Sensing: A Discrete-Time Markov Model", Elsevier Computer Networks, Apr. 2010* [82]. Includes contributions *P2.C1* to *P2.C4* with special emphasis on model validation and evaluation (i.e. *P2.C3*). Focus of [JA4] is on analytical proof of the proposed contributions in *P2.C1* and *P2.C2*. Furthermore, relevant performance evaluation as described by contribution *P2.C4* is also provided.

[CA4] *J. Pérez-Romero, X. Gelabert, O. Sallent, R. Agustí, "A Novel Framework for the Characterization of Dynamic Spectrum Access Scenarios (Invited Paper)", in IEEE International Symposium on Personal, Indoor and Mobile Radio Communications 2008 (PIMRC'08)* [83]. Introduces the Markov model further extended in [JA4].

[JA5] *X. Gelabert, O. Sallent, J. Pérez-Romero, and R. Agustí, "Flexible Spectrum Access for Opportunistic Secondary Operation in Cognitive Radio Networks", Submitted for possible journal publication, Dec. 2009* [84]. Relays on contributions *P2.C1* and *P2.C2* from [JA4] in order to provide a flexible spectrum assignment framework as stated by contributions *P2.C5* and *P2.C6*.

[JA6] *X. Gelabert, I. F. Akyildiz, O. Sallent, and R. Agustí, "Operating Point Selection for Primary and Secondary Users in Cognitive Radio Networks," Elsevier Computer Networks, Feb. 2009* [85]. This publication is mainly devoted to contribution *P2.C7* concerned with the suitable selection of the operating point in a sensing-based mechanism for secondary opportunistic access. In addition, it also provides some insights in contributions *P2.C1* and *P2.C2* regarding the model definition.

[CA5] *X. Gelabert, O. Sallent, J. Pérez-Romero, and R. Agustí "Exploiting the Operating Point in Sensing-Based Opportunistic Spectrum Access Scenarios", in IEEE International Conference on Communications 2009* (ICC'09) [86]. Contains derivative work from [JA6] regarding the appropriate selection of the operating point value in sensing-based secondary spectrum access scenarios.

1.5.1.1 Other Supporting Contributions

Besides the aforementioned publications, which constitute the focus of this thesis, several additional contributions by the author have helped in the overall process of developing concepts, ideas and strategies towards the completion of this dissertation. In this sense, the following supporting conference articles (SCAs) have been produced (in chronological order):

[SCA1] *X. Gelabert, J. Pérez-Romero, O. Sallent, R. Agustí, F. Casadevall, "Radio resource management in heterogeneous networks", in Proceedings of the International Working Conference on Performance Modelling and Evaluation of Heterogeneous Networks (HET-NETs'05)* [46]. Presents concepts, functionalities and architectures supporting CRRM strategies in heterogeneous wireless access networks.

[SCA2] *X. Gelabert, J. Pérez-Romero, O. Sallent, R. Agustí, "On the Impact of Multimode Terminals in Heterogeneous Wireless Access Networks", in Second International Symposium of Wireless Communication Systems 2005 (ISWCS'05)* [87]. Contains initial assessment on the performance degradation introduced by single-mode terminals.

[SCA3] *X. Gelabert, J. Pérez-Romero, O. Sallent, R. Agustí, "On the Suitability of Load Balancing Principles in Heterogeneous Wireless Access Networks" in Proc. Int. Symposium on Wireless Personal Multimedia Communications 2005 (WPMC'05)* [88]. Presents the Load Balancing (LB) policy in a GSM/EDGE plus UMTS heterogeneous network scenario, where evaluation is performed using a simulation platform.

[SCA4] *X. Gelabert, J. Pérez-Romero, O. Sallent, R. Agustí, "Congestion Control Strategies in Multi-Access Networks", 3rd International Symposium of Wireless Communication Systems 2006 (ISWCS 06)* [89]. Congestion events in a GSM/EDGE plus UMTS heterogeneous network is addressed using a simulation platform. Several congestion resolution strategies are proposed and evaluated in the case of a congested GSM/EDGE RAT.

[SCA5] *X. Gelabert, J. Pérez-Romero, O. Sallent, R. Agustí, "On Managing Multiple Radio Access Congestion Events in B3G Scenarios", in Proc. 65th Semiannual IEEE Vehicular Technology Conference Spring (VTC-Spring'07)* [90]. Extends the work in [SCA4] by evaluating congestion resolution mechanisms when both GSM/EDGE and UMTS networks undergo radio access congestion.

[SCA6] *O. Sallent, J. Pérez-Romero, X. Gelabert, J. Nasreddine, R. Agustí, F. Casadevall, A. Umbert, J. Olmos, "Gestión Integrada de Redes de Acceso Radio Celulares 2G, 2.5G y 3G", XXII Simpósium Nacional de la Unión Científica Internacional de Radio de 2007 (URSI2007)* [91]. Provides an overview of CRRM strategies, with special emphasis on the evaluation of several RAT selection policies.

In addition, the work in this thesis has been also reflected in the contribution towards the following book chapter (BC):

[BC1] *J. Pérez-Romero, X. Gelabert, O Sallent, "Radio Resource Management for Heterogeneous Wireless Access Networks", Springer, 2008* [92]. Provides a description of interworking and coupling scenarios between several RATs. RRM and CRRM functionalities and models are presented along with the description of several RAT selection alternatives in 3GPP-based technologies scenarios.

Finally, many of concepts related to RRM and CRRM in this dissertation were delivered by the author in the context of a tutorial session at the ISWCS'06[10]. The interested reader may refer to [93] for an electronic version of the presented slides.

[10] International Symposium of Wireless Communication Systems 2006.

1.5.2 Project Involvement

The work carried-out during the doctoral research period has been inevitably linked to research projects being funded by both Spanish (ES) and European Union (EU) entities. This offered the author the opportunity to present the work to project partners, from both industry and academia, obtaining fruitful feedback and comments leading to an overall improvement and assessment on the quality of the work submitted towards this dissertation. In particular, contributions towards EU-funded projects EVEREST [94] and AROMA [95] dealt with RRM and CRRM strategies in heterogeneous wireless networks. In addition, work for EU-project E^3 [96] was devoted to opportunistic spectrum access in spectrum sharing scenarios. Alongside, ES-funded projects COSMOS [97] and, subsequently, COGNOS [98] followed similar paths to those of the abovementioned EU-projects.

While a description of each project is provided in Appendix A, in the following, an overview of the contributions by the author towards each project is summarized:

- EVEREST/COSMOS: After some initial work addressing RRM strategies in GPRS-EDGE, the author contributed with several studies concerning RAT selection and multi-mode terminal availability in the context of a heterogeneous wireless network composed by GSM/EDGE and UMTS RATs. The adopted approach consisted in evaluating such scenarios by utilizing a system-level simulator platform based on commercial software OPNET®.
- AROMA/COSMOS: Contributions towards a Markovian formulation of the RAT selection problem along with the characterization of radio access congestion. In addition, a simulation-based approach to congestion control was also addressed.
- E^3/COGNOS: The scope of this project required to formulate a Markovian framework for the evaluation of opportunistic spectrum access scenarios for licensed/unlicensed spectrum sharing. In addition, spectrum sensing techniques were also modeled using a probabilistic approach.

The outcome of this project involvement can be measured, not only regarding the aforementioned journal and conference publications, but also in the number of contributions towards project deliverables (PD), chronologically listed hereafter:

[PD1] *O. Sallent (Editor) ”EVEREST D11: First report on the evaluation of RRM/CRRM algorithms”, Nov. 2004* [99].

[PD2] *O. Sallent (Editor) "EVEREST D20: Final report on the evaluation of RRM/CRRM algorithms", Oct. 2005* [100]

[PD3] *O. Sallent (Editor) "AROMA D09 - First report on AROMA algorithms and simulation results", Nov. 2006* [101]

[PD4] O. Sallent (Editor) "AROMA D12 - Intermediate report on AROMA algorithms and simulation results", Apr. 2007 [102]

[PD5] *J. Pérez-Romero (Editor) "AROMA D18 - Final report on AROMA algorithms and simulation results", Dec. 2007* [103].

[PD6] *S. Chantaraskul, K. Moessner (Editors) "E^3 D5.3 - First Report on Cognition Enablers Schemes", Apr. 2009* [104]

[PD7] *P. Demestichas, et al. (Editors) "E^3 D5.5 - Final Report on Selected Cognitive Enablers Schemes", Dec. 2009* [105]

In addition, contributions towards the following white paper (WP) in the context of E^3 project was also carried out:

[WP1] *E^3 "White Paper on Spectrum Sensing", Nov. 2009* [106]. The author contributed with a framework for the operating point selection in sensing-based opportunistic spectrum access scenarios.

For the sake of representation, Fig. 1.9 presents the time-line of produced articles in the context of project participation.

1.6 Dissertation Outline

The dissertation outline is intended to match the problem definition discussed in Section 1.5. In particular, we divide the dissertation in two parts, each part addressing problems *P1* and *P2* respectively. Then, Part I, entitled *Radio Resource Management (RRM) in Multi-Access/Multi-Service Wireless Networks*, will be devoted to problem *P1* and related contributions. This part is composed by three chapters where, Chapter 2 comprises the definition of the Markovian framework for the evaluation of RAT selection strategies, Chapter 3 deals with the problem of RAT selection in constrained-access environments and, finally, Chapter 4 analyzes

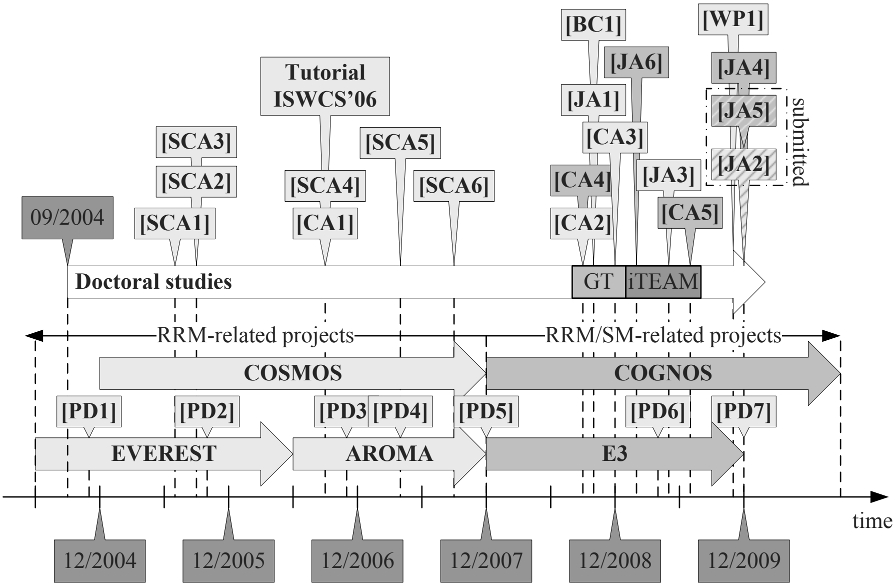


Figure 1.9: Time-line of publications in the context of project participation.

the impact of RAT selection on radio access congestion. In addition, Part II, entitled *Spectrum Management (SM) in Cognitive Radio Networks*, will be devoted to problem *P2* and related contributions. In this part, Chapter 5 will provide a discrete-time Markovian framework supporting spectrum sharing between licensed and unlicensed users with imperfect sensing. Chapter 6 will extend the previous framework to include higher flexibility in the spectrum channelization access along with several alternatives for unlicensed access services. In Chapter 7 the focus will be on how to adjust the sensing mechanism parameters such that some quality of service is ensured for both the licensed and unlicensed users. Finally, Chapter 8 summarizes the dissertation results providing some final remarks and future work.

In this dissertation, and bearing in mind the background provided in this introductory chapter, each following chapter is self-contained in terms of mathematical notation and provided references. In addition, each chapter will provide a concluding summary with the main contributions and remarks.

Bibliography

[1] E. Gustafsson and A. Jonsson, "Always best connected," *Wireless Communications, IEEE [see also IEEE Personal Communications]*, vol. 10, no. 1, pp. 49–55, 2003.

[2] Neeli Prasad and Anand Prasad, Eds., *WLAN Systems and Wireless IP for Next Generation Communications*, Artech House, Inc., Norwood, MA, USA, 2002.

[3] Bo Li, Yang Qin, Chor P. Low, and Choon L. Gwee, "A survey on mobile wimax [wireless broadband access]," *Communications Magazine, IEEE*, vol. 45, no. 12, pp. 70–75, 2007.

[4] Michel Mouly and Marie-Bernadette Pautet, *The GSM System for Mobile Communications*, Telecom Publishing, 1992.

[5] R. Kalden, I. Meirick, and M. Meyer, "Wireless Internet access based on GPRS," *Personal Communications, IEEE*, vol. 7, no. 2, pp. 8–18, August 2002.

[6] Harri Holma and Antti Toskala, *WCDMA for UMTS: Radio Access for Third Generation Mobile Communications*, John Wiley & Sons, Inc., New York, NY, USA, 2000.

[7] A. Furuskar, S. Mazur, F. Muller, and H. Olofsson, "EDGE: enhanced data rates for GSM and TDMA/136 evolution," *Personal Communications, IEEE*, vol. 6, no. 3, pp. 56–66, August 2002.

[8] Harri Holma and Antti Toskala, *HSDPA/HSUPA for UMTS: High Speed Radio Access for Mobile Communications*, John Wiley & Sons, 2006.

[9] Erik Dahlman, Stefan Parkvall, Johan Skold, and Per Beming, *3G Evolution, Second Edition: HSPA and LTE for Mobile Broadband*, Academic Press, 2008.

[10] Pierre Lescuyer and Thierry Lucidarme, *Evolved Packet System (EPS): The LTE and SAE Evolution of 3G UMTS*, Wiley, March 2008.

[11] Harri Holma and Antti Toskala, *LTE for UMTS: OFDMA and SC-FDMA Based Radio Access*, John Wiley & Sons, 2009.

[12] S. Glisic and B. Lorenzo, *Advanced Wireless Networks: Cognitive, Cooperative & Opportunistic 4G Technology, 2nd Edition*, John Wiley & Sons, 2009.

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

[14] Ian F. Akyildiz, Won Y. Lee, Mehmet C. Vuran, and Shantidev Mohanty, "Next generation/dynamic spectrum access/cognitive radio wireless networks: a survey," *Comput. Netw.*, vol. 50, no. 13, pp. 2127–2159, September 2006.

[15] Ramjee Prasad and Albena Mihovska, *New Horizons in Mobile and Wireless Communications: Reconfigurability*, Artech House, Inc., Norwood, MA, USA, 2009.

[16] Friedrich K. Jondral, "Software-defined radio: basics and evolution to cognitive radio," *EURASIP J. Wirel. Commun. Netw.*, vol. 2005, no. 3, pp. 275–283, 2005.

[17] J. Zander and S. Kim, *Radio Resource Management For Wireless Networks*, Artech House Publishers, 2001.

[18] Ekram Hossain, Dusit Niyato, and Zhu Han, *Dynamic Spectrum Access and Management in Cognitive Radio Networks*, Cambridge University Press, 2009.

[19] M. H. Ahmed, "Call admission control in wireless networks: a comprehensive survey," *Communications Surveys & Tutorials, IEEE*, vol. 7, no. 1, pp. 49–68, May 2005.

[20] Majid Ghaderi and Raouf Boutaba, "Call admission control in mobile cellular networks: a comprehensive survey," *Wireless Communications and Mobile Computing*, vol. 6, no. 1, pp. 69–93, 2006.

[21] L. L. H. Andrew, S. V. Hanly, and R. G. Mukhtar, "Active queue management for fair resource allocation in wireless networks," *Mobile Computing, IEEE Transactions on*, vol. 7, no. 2, pp. 231–246, December 2007.

[22] Long Le, Jay Aikat, Kevin Jeffay, and F. Donelson Smith, "The effects of active queue management and explicit congestion notification on web performance," *IEEE/ACM Trans. Netw.*, vol. 15, no. 6, pp. 1217–1230, 2007.

[23] Yaxin Cao and V. O. K. Li, "Scheduling algorithms in broadband wireless networks," *Proceedings of the IEEE*, vol. 89, no. 1, pp. 76–87, August 2002.

[24] Ajay C. Gummalla and John O. Limb, "Wireless medium access control protocols," *Communications Surveys & Tutorials, IEEE*, vol. 3, no. 2, pp. 2–15, November 2009.

[25] S. A. Grandhi, R. D. Yates, and D. J. Goodman, "Resource allocation for cellular radio systems," *Vehicular Technology, IEEE Transactions on*, vol. 46, no. 3, pp. 581–587, August 2002.

[26] Phone Lin and Yi-Bing Lin, "Channel allocation for GPRS," *Vehicular Technology, IEEE Transactions on*, vol. 50, no. 2, pp. 375–387, August 2002.

[27] Seong-Jun Oh, Danlu Zhang, and K. M. Wasserman, "Optimal resource allocation in multiservice CDMA networks," *Wireless Communications, IEEE Transactions on*, vol. 2, no. 4, pp. 811–821, July 2003.

[28] Surachai Chieochan and Ekram Hossain, "Adaptive radio resource allocation in OFDMA systems: a survey of the state-of-the-art approaches," *Wireless Communications and Mobile Computing*, vol. 9, no. 4, pp. 513–527, 2009.

[29] J. Perez-Romero, O. Sallent, and R. Agusti, "A novel approach for multicell load control in W-CDMA," *3G Mobile Communication Technologies, 2004. 3G 2004. Fifth IEE International Conference on*, pp. 688–692, November 2005.

[30] W. Rave, T. Kohler, J. Voigt, and G. Fettweis, "Evaluation of load control strategies in an UTRA/FDD network," *Vehicular Technology Conference, 2001. VTC 2001 Spring. IEEE VTS 53rd*, vol. 4, pp. 2710–2714 vol.4, August 2002.

[31] N. D. Tripathi, J. H. Reed, and H. F. VanLandinoham, "Handoff in cellular systems," *Personal Communications, IEEE*, vol. 5, no. 6, pp. 26–37, August 2002.

[32] J. Pérez-Romero et al., "Common radio resource management: functional models and implementation requirements," in *Personal, Indoor and Mobile Radio Communications. IEEE 16th International Symposium on*, 2005, pp. 2067–71 Vol. 3.

[33] A. Tolli, P. Hakalin, and H. Holma, "Performance evaluation of common radio resource management (CRRM)," in *Communications. IEEE International Conference on*, 2002, pp. 3429–33 vol.5.

[34] D. Niyato and E. Hossain, "Dynamics of network selection in heterogeneous wireless networks: An evolutionary game approach," *Vehicular Technology, IEEE Transactions on*, vol. 58, no. 4, pp. 2008–2017, 2009.

[35] Y. Wei, Y. Hu, and J. Song, "Network selection strategy in heterogeneous multi-access environment," *The Journal of China Universities of Posts and Telecommunications*, vol. 14, pp. 16–49, October 2007.

[36] G. Godor and G. Detari, "Novel network selection algorithm for various wireless network interfaces," *Mobile and Wireless Communications Summit, 2007. 16th IST*, pp. 1–5, July 2007.

[37] Gábor Fodor, Anders Furuskär, and Johan Lundsjö, "On access selection techniques in always best connected networks," in *In Proc. ITC Specialist Seminar on Performance Evaluation of Wireless and Mobile Systems*, 2004.

[38] B. Xing and Nalini Venkatasubramanian, "Multi-constraint dynamic access selection in always best connected networks," in *Mobile and Ubiquitous Systems: Networking and Services, 2005. MobiQuitous 2005. The Second Annual International Conference on*, 2005, pp. 56–64.

[39] I. Koo, A. Furuskar, J. Zander, and Kiseon Kim, "Erlang capacity of multiaccess systems with service-based access selection," vol. 8, no. 11, pp. 662–664, 2004.

[40] A. K. Salkintzis, "Interworking techniques and architectures for WLAN/3G integration toward 4G mobile data networks," *Wireless Communications, IEEE*, vol. 11, no. 3, pp. 50–61, June 2004.

[41] 3GPP, "Feasibility Study on 3GPP System to Wireless Local Area Network (WLAN) Interworking," TR 22.934, 3rd Generation Partnership Project (3GPP), 2007.

[42] ETSI, "Requirements and Architectures for Interworking Between HIPERLAN/2 and 3rd Generation Cellular Systems," ETSI 101 957, European Telecommunications Standards Institute (ETSI), 2001.

[43] 3GPP, "Improvement of RRM across RNS and RNS/BSS," TR 25.881 v5.0.0, 3rd Generation Partnership Project (3GPP), 2001.

[44] 3GPP, "Improvement of RRM across RNS and RNS/BSS (post rel-5) (release 6)," TR 25.891 v0.3.0, 3rd Generation Partnership Project (3GPP), 2003.

[45] O. Sallent, "A perspective on radio resource management in b3g," *Wireless Communication Systems, 2006. ISWCS '06. 3rd International Symposium on*, pp. 30–34, 2006.

[46] X. Gelabert, J. Perez-Romero, O. Sallent, R. Agusti, and F. Casadevall, "Radio resource management in heterogeneous networks," *Proceedings of the International Working Conference on Performance Modelling and Evaluation of Heterogeneous Networks (HET-NETs'05)*, 2005.

[47] J. Mitola, *Cognitive Radio*, Licentiate proposal, KTH, Stockholm, Sweden., 1998.

[48] M. M. Buddhikot, "Understanding dynamic spectrum access: Models,taxonomy and challenges," in *New Frontiers in Dynamic Spectrum Access Networks, 2007. DySPAN 2007. 2nd IEEE International Symposium on*, June 2007, pp. 649–663.

[49] Qing Zhao and B. M. Sadler, "A survey of dynamic spectrum access," *Signal Processing Magazine, IEEE*, vol. 24, no. 3, pp. 79–89, 2007.

[50] W. Lehr and J. Crowcroft, "Managing shared access to a spectrum commons," *New Frontiers in Dynamic Spectrum Access Networks, 2005. DySPAN 2005. 2005 First IEEE International Symposium on*, pp. 420–444, November 2005.

[51] J. Huang, R. A. Berry, and M. L. Honig, "Spectrum sharing with distributed interference compensation," in *New Frontiers in Dynamic Spectrum Access Networks, 2005. DySPAN 2005. 2005 First IEEE International Symposium on*, November 2005, pp. 88–93.

[52] Jianwei Huang, R. A. Berry, and M. L. Honig, "Distributed interference compensation for wireless networks," *Selected Areas in Communications, IEEE Journal on*, vol. 24, no. 5, pp. 1074–1084, May 2006.

[53] Ayman Elezabi, Mohamed Kashef, Mohamed Abdallah, and Mohamed M. Khairy, "Cognitive interference-minimizing code assignment for underlay CDMA networks in asynchronous multipath fading channels," in *IWCMC '09: Proceedings of the 2009 International Conference on Wireless Communications and Mobile Computing*, New York, NY, USA, 2009, pp. 1279–1283, ACM.

[54] H. Arslan and M. E. Sahin, "Cognitive UWB-OFDM: pushing UWB beyond its limit via opportunistic spectrum usage," *Journal of Communications and Networks Special Issue on Spectrum Resource Optimization*, vol. 8, no. 2, pp. 151–157, June 2006.

[55] V. Brik, E. Rozner, S. Banerjee, and P. Bahl, "DSAP: a protocol for coordinated spectrum access," in *New Frontiers in Dynamic Spectrum Access Networks, 2005. DySPAN 2005. 2005 First IEEE International Symposium on*, November 2005, pp. 611–614.

[56] S. Sankaranarayanan, P. Papadimitratos, A. Mishra, and S. Hershey, "A bandwidth sharing approach to improve licensed spectrum utilization," in *New Frontiers in Dynamic Spectrum Access Networks, 2005. DySPAN 2005. 2005 First IEEE International Symposium on*, November 2005, pp. 279–288.

[57] C. Raman, R. D. Yates, and N. B. Mandayam, "Scheduling variable rate links via a spectrum server," in *New Frontiers in Dynamic Spectrum Access Networks, 2005. DySPAN 2005. 2005 First IEEE International Symposium on*, November 2005, pp. 110–118.

[58] S. A. Zekavat and X. Li, "User-central wireless system: ultimate dynamic channel allocation," in *New Frontiers in Dynamic Spectrum Access Networks, 2005. DySPAN 2005. 2005 First IEEE International Symposium on*, November 2005, pp. 82–87.

[59] Lili Cao and Haitao Zheng, "Distributed spectrum allocation via local bargaining," in *Sen-*

sor and Ad Hoc Communications and Networks, 2005. IEEE SECON 2005. 2005 Second Annual IEEE Communications Society Conference on, December 2005, pp. 475–486.

[60] L. Ma, X. Han, and C. C. Shen, "Dynamic open spectrum sharing mac protocol for wireless ad hoc networks," in *New Frontiers in Dynamic Spectrum Access Networks, 2005. DySPAN 2005. 2005 First IEEE International Symposium on*, November 2005, pp. 203–213.

[61] T. Fujii and Y. Suzuki, "Ad-hoc cognitive radio - development to frequency sharing system by using multi-hop network," in *New Frontiers in Dynamic Spectrum Access Networks, 2005. DySPAN 2005. 2005 First IEEE International Symposium on*, November 2005, pp. 589–592.

[62] A. Tonmukayakul and M. B. Weiss, "Secondary use of radio spectrum: A feasibility analysis," in *Telecommunications Policy Research Conference (TPRC'04)*, October 2004.

[63] Federal Communications Commission FCC, "Principles for promoting the efficient use of spectrum by encouraging the development of secondary markets," ET Docket no. 00-401, December 2000.

[64] Jon M. Peha and Sooksan Panichpapiboon, "Real-time secondary markets for spectrum," *Telecommunications Policy*, vol. 28, no. 7-8, pp. 603–618, September 2004.

[65] Milind M. Buddhikot, Paul Kolodzy, Scott Miller, Kevin Ryan, and Jason Evans, "DIMSUMNet: New Directions in Wireless Networking Using Coordinated Dynamic Spectrum Access," in *World of Wireless Mobile and Multimedia Networks, 2005. WoWMoM 2005. Sixth IEEE International Symposium on a*, Los Alamitos, CA, USA, 2005, vol. 1, pp. 78–85, IEEE Computer Society.

[66] T. X. Brown, "An analysis of unlicensed device operation in licensed broadcast service bands," in *New Frontiers in Dynamic Spectrum Access Networks, 2005. DySPAN 2005. 2005 First IEEE International Symposium on*, November 2005, pp. 11–29.

[67] S. Mangold, Zhun Zhong, K. Challapali, and Chun-Ting Chou, "Spectrum agile radio: radio resource measurements for opportunistic spectrum usage," in *Global Telecommunications Conference, 2004. GLOBECOM '04. IEEE*, December 2004, vol. 6, pp. 3467–3471 Vol.6.

[68] A. Ghasemi and E. S. Sousa, "Collaborative spectrum sensing for opportunistic access in fading environments," in *New Frontiers in Dynamic Spectrum Access Networks, 2005. DySPAN 2005. 2005 First IEEE International Symposium on*, December 2005, pp. 131–136.

[69] T. Yucek and H. Arslan, "A survey of spectrum sensing algorithms for cognitive radio applications," *Communications Surveys & Tutorials, IEEE*, vol. 11, no. 1, pp. 116–130, March 2009.

[70] Leonard Kleinrock, *Queueing Systems. Volume 1: Theory*, Wiley-Interscience, 1 edition, January 1975.

[71] V. B. Iversen, "Teletraffic engineering and network planning (2009 version)," http://www.com.dtu.dk/teletraffic/, Technical University of Denmark, May 2009.

[72] Piet V. Mieghem, *Performance Analysis of Communications Networks and Systems*, Cambridge University Press, New York, NY, USA, 2005.

[73] G. R. Dattatreya, *Performance Analysis of Queuing and Computer Networks*, Chapman & Hall/CRC, 2008.

[74] William J. Stewart, *Introduction to the numerical solution of Markov chains*, Princeton University Press, 1994.

[75] Gunter Bolch, Stefan Greiner, Hermann de Meer, and Kishor S. Trivedi, *Queueing Networks and Markov Chains: Modeling and Performance Evaluation with Computer Science Applications*, WileyBlackwell, 2nd edition edition, May 2006.

[76] X. Gelabert, J. Pérez-Romero, O. Sallent, and R. Agustí, "A Markovian approach to radio access technology selection in heterogeneous multiaccess/multiservice wireless networks," *Mobile Computing, IEEE Transactions on*, vol. 7, no. 10, 2008.

[77] X. Gelabert, J. Perez-Romero, O. Sallent, and R. Agusti, "A 4-Dimensional Markov Model for the Evaluation of Radio Access Technology Selection Strategies in Multiservice Scenarios," in *Proc. 64th Semi-annual IEEE Vehicular Technology Conference Fall (VTC-Fall'06)*, Montreal, Canada, September 25-28, 2006.

[78] X. Gelabert, O. Sallent, J. Pérez-Romero, and R. Agustí, "Performance evaluation of radio access technology selection strategies in constrained multi-access/multi-service wireless networks," *Submitted for journal publication*, December 2009.

[79] X. Gelabert, O. Sallent, J. Perez-Romero, and R. Agusti, "Erlang Capacity Degradation in Multi-Access/Multi-service Wireless Networks due to Terminal/Coverage Restrictions," in *11th Int. Symposium on Wireless Personal Multimedia Communications (WPMC'08)*, Lapland, Finland, September 8-11, 2008.

[80] X. Gelabert, O. Sallent, J. Perez-Romero, and R. Agusti, "Radio access congestion in multiaccess/multiservice wireless networks," *Vehicular Technology, IEEE Transactions on*, vol. 58, no. 8, pp. 4462–4475, May 2009.

[81] X. Gelabert, J. Perez-Romero, O. Sallent, and R. Agusti, "Evaluation of radio access congestion in heterogeneous wireless access networks," in *Global Telecommunications Conference, 2008. IEEE GLOBECOM 2008. IEEE*, 2008, pp. 1–6.

[82] X. Gelabert, O. Sallent, J. Pérez-Romero, and R. Agustí, "Spectrum sharing in cognitive radio networks with imperfect sensing: A discrete-time markov model," *Elsevier Computer Networks*, April 2010.

[83] J. Perez-Romero, X. Gelabert, O. Sallent, and R. Agusti, "A novel framework for the characterization of dynamic spectrum access scenarios," in *Personal, Indoor and Mobile Radio Communications, 2008. PIMRC 2008. IEEE 19th International Symposium on*, 2008, pp. 1–6.

[84] X. Gelabert, O. Sallent, J. Pérez-Romero, and R. Agustí, "Flexible spectrum access for opportunistic secondary operation in cognitive radio networks," *Submitted for journal publication*, December 2009.

[85] X. Gelabert, I. F. Akyildiz, O. Sallent, and R. Agustí, "Operating point selection for primary and secondary users in cognitive radio networks," *Computer Networks*, February 2009.

[86] X. Gelabert, O. Sallent, J. Perez-Romero, and R. Agusti, "Exploiting the operating point in sensing-based opportunistic spectrum access scenarios," in *Communications, 2009. ICC '09. IEEE International Conference on*, August 2009, pp. 1–6.

[87] X. Gelabert, J. Pérez-Romero, O. Sallent, and R. Agustí, "On the impact of multimode terminals in heterogeneous wireless access networks," in *Second International Symposium of Wireless Communication Systems 2005 (ISWCS'05)*, Siena, Italy, September 5-7, 2005.

[88] X. Gelabert, J. Perez-Romero, O. Sallent, and R. Agusti, "On the suitability of load balancing principles in heterogeneous wireless access networks," in *Int. Symposium on Wireless Personal Multimedia Communications (WPMC'05)*, Aalborg, Denmark, September 18-22., 2005.

[89] X. Gelabert, J. Pérez-Romero, O. Sallent, and R. Agustí, "Congestion control strategies in multi-access networks," in *Wireless Communication Systems, 2006. ISWCS '06. 3rd International Symposium on*, 2006, pp. 579–583.

[90] X. Gelabert, J. Perez-Romero, O. Sallent, and R. Agusti, "On managing multiple radio access congestion events in b3g scenarios," in *Proc. 65th Semi-annual IEEE Vehicular Technology Conference Spring (VTC-Spring'07)*, Dublin, Ireland, April 22-25, 2007.

[91] O. Sallent, J. Pérez-Romero, X. Gelabert, J. Nasreddine, R. Agustí, F. Casadevall, A. Umbert, and J. Olmos, "Gestión integrada de redes de acceso radio celulares 2g, 2.5g y 3g [in spanish]," in *XXII Simpósium Nacional de la Unión Científica Internacional de Radio de 2007, URSI2007*, La Laguna, Tenerife, Spain, September, 2007.

[92] J. Pérez-Romero, X. Gelabert, and O. Sallent, "Radio resource management for heterogeneous wireless access networks," in *Heterogeneous Wireless Access Networks - Architectures and Protocols*, E. Hossain, Ed., chapter 5, pp. 133–166. Springer, 2008.

[93] X. Gelabert, "Topics on Common Radio Resource Management (CRRM) Strategies for QoS provisioning over Heterogeneous (Beyond 3G) Wireless Radio Access Networks," `http://www.tsc.upc.edu/grcm/images/stories/tutorial_iswcs06_rrm_in_b3g.pdf`, September 2006, Tutorial at ISWCS'06 [URL accessed March 1, 2010].

[94] "EVEREST: Evolutionary Strategies for Radio Resource Management in Cellular Heterogeneous Networks," `http://www.everest-ist.upc.es/`, [URL accessed March 1, 2010].

[95] "AROMA: Advanced Resource Management Solutions for future all IP Heterogeneous Mobile Radio Environments," `http://www.aroma-ist.upc.edu/`, [URL accessed March 1, 2010].

[96] "E^3: End-to-End Efficiency," `https://ict-e3.eu/`, [URL accessed March 1, 2010].

[97] "COSMOS: Calidad de servicio extremo a extremo y flexibilidad espectral en redes móviles heterogéneas," `http://www.cosmos.upc.edu/`, [URL accessed March 1, 2010].

[98] "COGNOS: Gestión cognitiva de recursos radio y espectro radioeléctrico en redes móviles heterogéneas con provisión de calidad de servicio extremo a extremo," `http://www.cognos.upc.edu/`, [URL accessed March 1, 2010].

[99] O. Sallent (Editor), "EVEREST D11 - First report on the evaluation of RRM/CRRM algorithms," `http://www.everest-ist.upc.es/`, November 2004, [Available upon request. Accessed March 1, 2010].

[100] O. Sallent (Editor), "EVEREST D20 - Final report on the evaluation of RRM/CRRM algorithms," `http://www.everest-ist.upc.es/`, October 2005, [Available upon request. Accessed March 1, 2010].

[101] O. Sallent (Editor), "AROMA D09 - First report on AROMA algorithms and simulation results," `http://www.aroma-ist.upc.edu/`, November 2006, [Available upon request. Accessed March 1, 2010].

[102] O. Sallent (Editor), "AROMA D12 - Intermediate report on AROMA algorithms and simulation results," `http://www.aroma-ist.upc.edu/`, April 2007, [Available upon request. Accessed March 1, 2010].

[103] J. Pérez-Romero (Editor), "AROMA D18 - Final report on AROMA algorithms and simulation results," `http://www.aroma-ist.upc.edu/`, December 2007, [Available upon request. Accessed March 1, 2010].

[104] S. Chantaraskul and K. Moessner (Editors), "E^3 D5.3 - First Report on Cognition Enablers Schemes," `https://ict-e3.eu/`, April 2009, [Available upon request. Accessed March 1, 2010].

[105] P. Demestichas et al. (Editors), "E^3 D5.5 - Final Report on Selected Cognitive Enablers Schemes," `https://ict-e3.eu/`, December 2009, [Available upon request. Accessed March 1, 2010].

[106] E^3, "E^3 White Paper on Spectrum Sensing," `https://ict-e3.eu/project/white_papers/E3_White_Paper_Sensing.pdf`, November 2009, [Available online. Accessed March 1, 2010].

Part I

Radio Resource Management (RRM) in Multi-Access/ Multi-Service Wireless Networks

Chapter 2

A Markovian Approach to Radio Access Technology Selection in Heterogeneous Multi-Access/ Multi-Service Wireless Networks

This chapter addresses the problem of Radio Access Technology (RAT) selection in heterogeneous multi-access/multi-service scenarios. For such purpose, a Markov model is proposed to compare the performance of various RAT selection policies within these scenarios. The novelty of the approach resides in the embedded definition of the aforementioned RAT selection policies within the Markov chain. In addition, the model also considers the constraints imposed by those users with terminals that only support a subset of all the available RATs (i.e. multi-mode terminal capabilities). Furthermore, several performance metrics may be measured to evaluate the behaviour of the proposed RAT selection policies under varying offered traffic conditions. In order to illustrate the validation and suitability of the proposed model, some examples of operative radio access networks are provided, including the GSM/EDGE Radio Access Network (GERAN) and the UMTS Radio Access Network (UTRAN), as well as several service-based, load-balancing and terminal-driven RAT selection strategies. The flexibility exhibited by the presented model enables to extend these RAT selection policies to others responding to diverse criteria. The model is successfully validated by means of comparing the Markov model results with those of system-level simulations.

2.1 Initial RAT Selection

Inherent to heterogeneous networks, to select an appropriate RAT for an incoming user requesting a given service becomes a key CRRM issue. This RAT selection can be carried out considering different criteria (such as, e.g., service type, load conditions, etc.) with the final purpose of enhancing overall capacity, resource utilisation and service quality.

Although the RAT selection problem has been covered in a number of papers, see for example [1–3], the proposed methodology usually relies on system level simulations in order to extract some relevant performance metrics to compare different strategies. The analytical approach to the RAT selection problem, however, has been less addressed in the literature. To the authors knowledge, only a few analytical proposals have been developed, e.g. [4], [5] and [6]. In [4], Lincke et al. propose an analytical approach to the problem of traffic overflowing between several RATs using a multi-dimensional Markov model. However, in order to derive a closed form solution by means of applying independence between service types, Markov states in this model indicate the number of sessions of each service that are being allocated in whole composite network, but not on which RAT each session is being served. In [5], a near-optimum service allocation is proposed in order to maximize the combined multi-service capacity. The authors assumed an a priori knowledge of the services that need to be allocated, rather than modelling user arrival process. In [6], Koo et al. assess the separate and common Erlang capacity of a multi-access/multi-service system. For this purpose, an Erlang Loss queuing approach is assumed and a closed product form expression for the equilibrium probability is provided. Nevertheless, this assumption implies that the fractional traffic loads of each service offered to each system are known, so the approach is only valid to evaluate some basic RAT selection policies.

In this chapter, the proposed analytical model entails a more flexible framework by assuming that only the total offered traffic to the multi-RAT system for each service is known. Thus, fractional traffic arriving to each RAT will be dependant on the chosen RAT selection scheme which is fully embedded in the model. This feature constitutes the main innovative contribution of this work and differentiates it from previous approaches to the problem. In particular, the model describes the allocation of two service types onto two RATs, which allows the definition of a wide range of RAT selection policies taking into account several criteria, such as service type, network conditions, terminal types, etc. Finally, the proposed model also captures the availability of multi-mode terminals, i.e. those that can operate on both RATs, so as to reflect a more realistic medium term scenario considering the flexibility constraints of those terminals supporting one single RAT (i.e. single-mode terminals).

Multi-dimensional Markov models have been widely used in the field of networking to model the behaviour of communication networks under variable traffic load conditions [7]. In the analysis of this chapter, the focus is on two RATs with different underlying access methodologies: the Time Division Multiple Access (TDMA) and the Wideband Code Division Multiple Access (WCDMA). These two access schemes may, e.g., represent 3rd Generation Partnership Project (3GPP) standardised technologies GSM/EDGE and UMTS respectively [8, 9], although the model could be adapted to other standards by a proper change in the parameter values.

The chapter is organized as follows: Section 2.2 deals with the problem statement and the considered approach to solve it. Section 2.3 presents the analytical model and the notation that will be used throughout the chapter. In section 2.4, various RAT selection policies are described by means of the proposed model. Section 2.5 presents the performance metrics that will be used to evaluate the behaviour of initial RAT selection policies in section 2.6. Finally, section 2.7 deals with the chapter conclusions.

2.2 Problem Statement

For the evaluation of the forthcoming RAT selection strategies a scenario is assumed where a TDMA-based and a WCDMA-based technology coexist and provide coverage over a same area. Generically, one can characterise this scenario by means of a Markov chain represented by a (N+M)-dimensional state, denoted as $S_{(t_1,t_2,\ldots,t_N,w_1,w_2,\ldots,w_M)}$, where t_n $(1 \leq n \leq N)$ and w_m $(1 \leq m \leq M)$ relate to the N and M dimensions corresponding to TDMA and WCDMA RATs respectively. Each of these dimensions can represent a single or a combination of communication characteristics, such as service type (e.g. voice or data), user communication status (e.g. active or queued users), transmission rate, amount of allocated resources, etc.

In this chapter a 4D Markov chain is considered accounting for two service types, generically voice and data, being served over the aforementioned RATs, TDMA and WCDMA, in order to model the system behaviour. Therefore, let $S_{(i,j,k,l)}$ represent the state in which i voice users and j data users are being served through TDMA; and k voice users and l data users are being served through WCDMA. These indices represent the number of active simultaneous voice calls and data sessions being carried out at a given time.

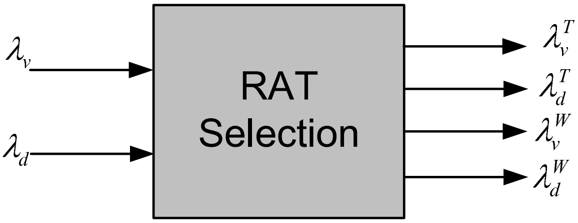


Figure 2.1: Mapping of total-to-fractional arrival rates given by initial RAT selection.

Transitions between states within the Markov chain will occur due to call/session arrivals or due to call/session departures. Regarding traffic patterns, it is supposed that voice calls and data sessions are generated according to Poisson processes with rates λ_v and λ_d respectively. As for voice-call holding time and data session time, they follow exponential distributions with means $1/\mu_v$ and $1/\mu_d$ correspondingly. It is assumed that only transitions between neighbouring states (those that only differ in a single increment/decrement in a sole state dimension) are allowed. This prevents situations where more than one call/session arrives or departs from a given state at the same time.

Within the set of CRRM functions devoted to efficiently manage the available resources in a heterogeneous network, the RAT selection plays a key role in deciding the most appropriate RAT for a given service at a given time. In that sense, the algorithm operation might then respond to specific policies taking into account both technical and/or economical aspects (e.g. operator or user preferences). In the context of the proposed Markov framework it is important to notice that, given the total voice call and data session arrival rates, λ_v and λ_d respectively, the adopted RAT selection policy will determine the arrival rates into each RAT (see Fig. 2.1). Consequently, the RAT selection policy will modulate the transition rates between the states $S_{(i,j,k,l)}$ in the Markov chain according to a predefined RAT selection policy. Mathematically, given a generic RAT selection policy denoted as $\pi_{(i,j,k,l)}$, we may introduce the following function:

$$\begin{array}{rccc} \pi_{(i,j,k,l)}: & \mathbb{R}^2 & \longrightarrow & \mathbb{R}^4 \\ & (\lambda_v, \lambda_d) & \longrightarrow & \left(\lambda_v^T, \lambda_d^T, \lambda_v^W, \lambda_d^W\right) \end{array} \tag{2.1}$$

where λ_v^T, λ_d^T, λ_v^W, λ_d^W represent the fractional arrival rates of each service into to each of the available RATs given by policy $\pi_{(i,j,k,l)}$. In this way, by an appropriate definition of the RAT selection policies, it is possible to embed those into the Markov chain and evaluate the performance of the system by considering that only the total voice and data offered traffic, i.e. λ_v and λ_d, are known parameters. This approach differentiates the presented work from previous mentioned studies, [4] [5] [6], and constitutes the main innovative contribution of this work which will be fully developed in the next sections.

2.3 The 4D Markov Model State Space

In the following, the Markov model state space containing the total set of feasible states is presented. Clearly, if the capacity in terms of number of supported users in each RAT is assumed to be fixed, a finite number of states $S_{(i,j,k,l)}$ (called feasible states) limited by the number of allowable users of each service in each RAT must exist.

This limit is usually set by the RAT-specific Call Admission Control (CAC) procedures, that determine if a given user should be admitted or not, so as to guarantee some minimum QoS requirements to users already admitted in the system. Because CAC is dependant on the underlying technology, the set of feasible states in TDMA, S^T, and WCDMA, S^W, can be individually defined as:

$$\begin{aligned} S^T &= \left\{ S_{(i,j,k,l)} | 0 \leq f^T_{(i,j)} \leq 1, \forall k,l \right\} \\ S^W &= \left\{ S_{(i,j,k,l)} | 0 \leq f^W_{(k,l)} \leq 1, \forall i,j \right\} \end{aligned} \tag{2.2}$$

where $f^T_{(i,j)}$ and $f^W_{(k,l)}$ are defined as the feasibility conditions which account for the CAC procedures in TDMA and WCDMA correspondingly.

Consequently, we can define the set of feasible states, S, which include all states $S_{(i,j,k,l)}$ that satisfy the CAC procedures in each of the systems. Then, a given state $S_{(i,j,k,l)}$ is said to be feasible, if it satisfies that $S_{(i,j,k,l)} \in S$ with $S = S^T \cap S^W$, i.e. a state is only feasible if it is feasible in both TDMA and WCDMA systems.

In the following subsections, the state feasibilities for TDMA and WCDMA are presented. In this chapter, CAC procedures are based on the reverse link (uplink) in order to determine the number of allowable users in each RAT, which it is assumed to be the most restricting case. Nonetheless, Chapter 4 will relax this assumption by considering both uplink and downlink effects in the CAC procedure.

2.3.1 TDMA State Feasibility

The resource allocation for voice and data services in a TDMA-based technology, such as, e.g. GSM/EDGE, relies on the *capacity on demand* principle [10]. Briefly, a data user can transmit data over a number of simultaneous channels or timeslots (TSLs). Moreover, several data users can be multiplexed over a same TSL for coordinated data transmission by means of an efficient scheduling mechanism.

Given that voice and data users can demand different amounts of resources and that these resources are shared between them, mechanisms to referee the sharing among voice and data traffic are needed [11]. In this chapter, and for the sake of simplicity, it is assumed that the total available capacity is shared between voice and data traffic on a first-come-first-served basis with no service priority.

If C is the total number of available channels (TSLs) available for voice and data services in the cell, the maximum number of voice users being served through TDMA, i, is upper-bounded by $i \leq C$. Considering the uplink (UL) direction, assuming voice and data users are granted with a single channel for each connection[1], and that a maximum number of n_C data users are allowed to share the same TSL, the maximum number of simultaneous data users being served through TDMA must satisfy $j \leq n_C C$. Since voice and data services share the total amount of resources, the previous conditions may be expressed jointly as:

$$0 \leq i/C + j/n_C C \leq 1 \tag{2.3}$$

which implicitly defines the state feasibility condition for the TDMA system, i.e. $f^T_{(i,j)} = i/C + j/n_C C$.

2.3.2 WCDMA State Feasibility

In WCDMA-based systems, the UL load factor ($L^W_{(k,l)}$) condition must hold in order to ensure that users are granted the desired capacity for their demanding services. Considering k voice users and l data users being served in WCDMA, the UL load factor condition for a single-cell may be expressed as [9]:

$$0 \leq L^W_{(k,l)} \leq \eta_{max} \tag{2.4}$$

where

$$L^W_{(k,l)} = k\left[\frac{W/R_{b,v}}{\left(E_b/N_0\right)_v} + 1\right]^{-1} + l\left[\frac{W/R_{b,d}}{\left(E_b/N_0\right)_d} + 1\right]^{-1} \tag{2.5}$$

with W the chip rate; $R_{b,v}$ and $R_{b,d}$ the bit rate granted to voice and data services; $\left(E_b/N_0\right)_v$ along with $\left(E_b/N_0\right)_d$ the target bit-energy-to-noise-density ratio after despreading and decoding for voice and data users; and η_{max} the admission threshold. By choosing an appropriate value for η_{max}, quality requirements of admitted users (e.g. in terms of bit error rate) depending on the coverage conditions can be ensured [9]. From (2.4), the state feasibility condition of WCDMA system, $f^W_{(k,l)}$, is easily identified as $f^W_{(k,l)} = L^W_{(k,l)}/\eta_{max}$.

[1] Although the consideration of multi-slot capabilities in the model would be feasible, at this point, this would complicate the algebra and the model while not bringing substantial added value on the methodology and approach in this chapter. Thus a single TSL is allocated to data users. Nonetheless, multi-slot capabilities will be addressed when dealing with congestion control in Chapter 4.

2.3.3 Call Admission Control and Blocking States

Once the state space has been defined by means of the feasibility conditions in each RAT, let us define, for the sake of convenience, the set of states in which the acceptance of a new user would force a transition to an unfeasible state $S_{(i,j,k,l)} \notin S$. Under these circumstances, the RAT in question is said to be in a blocking state. Let $S^{\rho}_{b,\sigma}$ denote the set of feasible states where the acceptance of a service type σ user in RAT ρ forces the state to move to an unfeasible state. Then, the fractional per-service/per-RAT blocking states for voice and data services, i.e. $\sigma = \{v, d\}$, in TDMA and WCDMA RATs, $\rho = \{T, W\}$, are defined as:

$$\begin{array}{l} S^{T}_{b,v} = \left\{ S_{(i,j,k,l)} \in S | S_{(i+1,j,k,l)} \notin S \right\} \\ S^{T}_{b,d} = \left\{ S_{(i,j,k,l)} \in S | S_{(i,j+1,k,l)} \notin S \right\} \\ S^{W}_{b,v} = \left\{ S_{(i,j,k,l)} \in S | S_{(i,j,k+1,l)} \notin S \right\} \\ S^{W}_{b,d} = \left\{ S_{(i,j,k,l)} \in S | S_{(i,j,k,l+1)} \notin S \right\} \end{array} \tag{2.6}$$

If S^{ρ}_{b} denotes the set of feasible states where the acceptance of any service type user in RAT ρ forces the state to move to an unfeasible state, we have:

$$S^{\rho}_{b} = S^{\rho}_{b,v} \cap S^{\rho}_{b,d} \tag{2.7}$$

Assuming that a given service type user can be allocated in either of the existing RATs provided the one selected by the RAT selection policy is blocked, we can define service blocking states where the acceptance of a given service type user $\sigma = \{v, d\}$ forces the current state to move to an unfeasible state in each of the considered RATs $\rho = \{T, W\}$. Bearing in mind (2.6), the per-service blocking set $S_{b,\sigma}$ can be defined as:

$$S_{b,\sigma} = S^{T}_{b,\sigma} \cap S^{W}_{b,\sigma} \tag{2.8}$$

Finally, if S_b defines the set of states where the acceptance of any service type user in any of the available RATs forces the state to move to an unfeasible state, then the total blocking states are defined as:

$$S_b = S^{T}_{b} \cap S^{W}_{b} = S^{T}_{b,v} \cap S^{T}_{b,d} \cap S^{W}_{b,v} \cap S^{W}_{b,d} \tag{2.9}$$

2.4 Radio Access Technology Selection Policies and State Transitions

Based on the relation provided by (2.1), we can conveniently define the RAT selection policies as functions that map the total arrival rates λ_v and λ_d into fractional arrival rates of each service into each system (i.e. λ_v^T, λ_d^T, λ_v^W and λ_d^W) depending on the current state information. Then, in a given state $S_{(i,j,k,l)}$, relation (2.1) can be rewritten as:

$$\begin{array}{cccc} \pi_{(i,j,k,l)}: & \mathbb{R}^2 & \longrightarrow & \mathbb{R}^4 \\ & \begin{pmatrix} \lambda_v \\ \lambda_d \end{pmatrix}^T & \longrightarrow & \begin{pmatrix} \alpha_{(i,j,k,l)}\lambda_v\delta_{(i+1,j,k,l)} \\ \beta_{(i,j,k,l)}\lambda_d\delta_{(i,j+1,k,l)} \\ \bar{\alpha}_{(i,j,k,l)}\lambda_v\delta_{(i,j,k+1,l)} \\ \bar{\beta}_{(i,j,k,l)}\lambda_d\delta_{(i,j,k,l+1)} \end{pmatrix}^T \end{array} \tag{2.10}$$

where, given RAT selection policy $\pi_{(i,j,k,l)}$, $\alpha_{(i,j,k,l)}$ and $\bar{\alpha}_{(i,j,k,l)} = (1 - \alpha_{(i,j,k,l)})$ are the functions determining the fraction of voice users into TDMA and WCDMA respectively, and $\beta_{(i,j,k,l)}$ along with $\bar{\beta}_{(i,j,k,l)} = (1 - \beta_{(i,j,k,l)})$ are the functions governing the fractional data arrival rates into TDMA and WCDMA respectively. Furthermore, function $\delta_{(i,j,k,l)}$ is an indicator function which will guarantee that non-feasible states, i.e. $S_{(i,j,k,l)} \notin S$, are not taken into account in the transitions, thus:

$$\delta_{(i,j,k,l)} = \begin{cases} 1 & \text{if } S_{(i,j,k,l)} \in S \\ 0 & \text{otherwise} \end{cases} \tag{2.11}$$

The proposed model also allows us to take into consideration scenarios in which not all terminals have multi-mode capabilities. In that respect, assume that the availability of terminals that support both RATs (multi-mode) is given by p which is defined as the fraction of terminals with multi-mode capabilities. Accordingly, the ratio of terminals that only support TDMA (single-mode) is given by $\bar{p} = 1-p$. The rationale behind this assignment resides in the fact that terminals supporting more recent technologies, such as WCDMA, will most probably support preceding technologies like TDMA. The converse, however, will be less usual. Then, the fractional traffic derived into each RAT stated in (2.10) may be rewritten as:

$$\begin{array}{cccc} \pi_{(i,j,k,l)}: & \mathbb{R}^2 & \longrightarrow & \mathbb{R}^4 \\ & \begin{pmatrix} \lambda_v \\ \lambda_d \end{pmatrix}^T & \longrightarrow & \begin{pmatrix} \alpha^p_{(i,j,k,l)}\lambda_v\delta_{(i+1,j,k,l)} \\ \beta^p_{(i,j,k,l)}\lambda_d\delta_{(i,j+1,k,l)} \\ \bar{\alpha}^p_{(i,j,k,l)}\lambda_v\delta_{(i,j,k+1,l)} \\ \bar{\beta}^p_{(i,j,k,l)}\lambda_d\delta_{(i,j,k,l+1)} \end{pmatrix}^T \end{array} \tag{2.12}$$

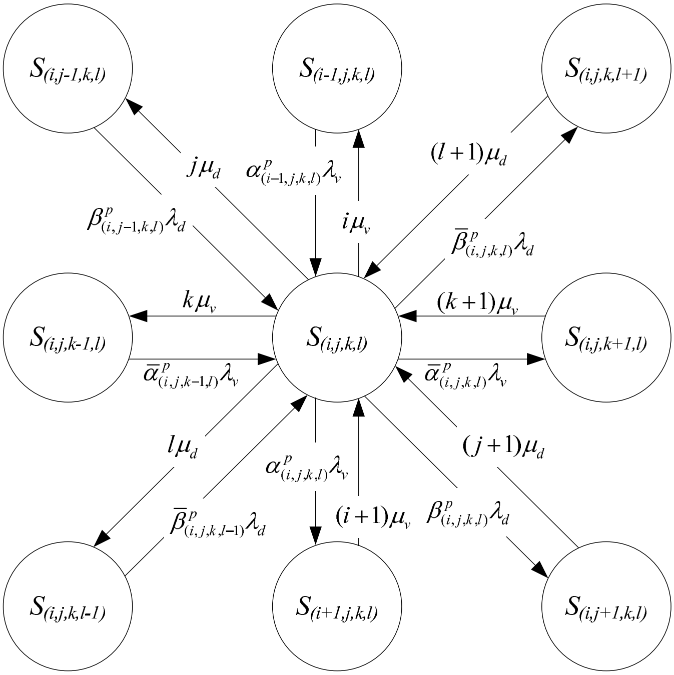

$S_{(i,j-1,k,l)}$
$S_{(i-1,j,k,l)}$
$S_{(i,j,k,l+1)}$
$j\mu_d$
$\alpha^p_{(i-1,j,k,l)}\lambda_v$
$(l+1)\mu_d$
$\beta^p_{(i,j-1,k,l)}\lambda_d$
$i\mu_v$
$\bar{\beta}^p_{(i,j,k,l)}\lambda_d$
$k\mu_v$
$(k+1)\mu_v$
$S_{(i,j,k-1,l)}$
$S_{(i,j,k,l)}$
$S_{(i,j,k+1,l)}$
$\bar{\alpha}^p_{(i,j,k-1,l)}\lambda_v$
$\bar{\alpha}^p_{(i,j,k,l)}\lambda_v$
$l\mu_d$
$(j+1)\mu_d$
$\alpha^p_{(i,j,k,l)}\lambda_v$
$\bar{\beta}^p_{(i,j,k,l-1)}\lambda_d$
$(i+1)\mu_v$
$\beta^p_{(i,j,k,l)}\lambda_d$
$S_{(i,j,k,l-1)}$
$S_{(i+1,j,k,l)}$
$S_{(i,j+1,k,l)}$

Figure 2.2: State transition diagram for a general state.

where $\alpha^p_{(i,j,k,l)}$ and $\beta^p_{(i,j,k,l)}$ relate to the RAT selection policy assignment considering the presence of both multi-mode and single-mode terminals. In particular, voice and data traffic offered to TDMA will consist of not only the traffic allocated by means of the applied RAT selection policy, but also by the traffic that does not support WCDMA. This can be expressed as:

$$\begin{aligned}\alpha^p_{(i,j,k,l)}\lambda_v &= \left[\alpha_{(i,j,k,l)} + (1-\alpha_{(i,j,k,l)})(1-p)\right]\lambda_v \\ &= \left(\bar{p} + \alpha_{(i,j,k,l)}p\right)\lambda_v \\ \beta^p_{(i,j,k,l)}\lambda_d &= \left[\beta_{(i,j,k,l)} + (1-\beta_{(i,j,k,l)})(1-p)\right]\lambda_d \\ &= \left(\bar{p} + \beta_{(i,j,k,l)}p\right)\lambda_d \end{aligned} \tag{2.13}$$

where $\alpha_{(i,j,k,l)}$ and $\beta_{(i,j,k,l)}$ relate to the policy decision (Note that if $p = 1$, i.e. all terminals are multi-mode, thus expression (2.12) becomes expression (2.10)).

Given the fractional arrival rates provided in (2.12), the state transition diagram at a particular non-boundary state $S_{(i,j,k,l)}$ may be built (Fig. 2.2). By inspection of Fig. 2.2 we may deduce the Steady-State Balance Equation (SSBE) for a given state $S_{(i,j,k,l)}$ as:

$$\begin{aligned}
&P_{(i,j,k,l)}[\alpha^p_{(i,j,k,l)}\lambda_v\delta_{(i+1,j,k,l)} + i\mu_v\delta_{(i-1,j,k,l)} \\
&+\bar{\alpha}^p_{(i,j,k,l)}\lambda_v\delta_{(i,j,k+1,l)} + k\mu_v\delta_{(i,j,k-1,l)} \\
&+\beta^p_{(i,j,k,l)}\lambda_d\delta_{(i,j+1,k,l)} + j\mu_d\delta_{(i,j-1,k,l)} \\
&+\bar{\beta}^p_{(i,j,k,l)}\lambda_d\delta_{(i,j,k,l+1)} + l\mu_d\delta_{(i,j,k,l-1)}] \\
&= \alpha^p_{(i-1,j,k,l)}\lambda_v P_{(i-1,j,k,l)}\delta_{(i-1,j,k,l)} \\
&+(i+1)\mu_v P_{(i+1,j,k,l)}\delta_{(i+1,j,k,l)} \\
&+\bar{\alpha}^p_{(i,j,k-1,l)}\lambda_v P_{(i,j,k-1,l)}\delta_{(i,j,k-1,l)} \\
&+(k+1)\mu_v P_{(i,j,k+1,l)}\delta_{(i,j,k+1,l)} \\
&+\beta^p_{(i,j-1,k,l)}\lambda_d P_{(i,j-1,k,l)}\delta_{(i,j-1,k,l)} \\
&+(j+1)\mu_d P_{(i,j+1,k,l)}\delta_{(i,j+1,k,l)} \\
&+\bar{\beta}^p_{(i,j,k,l-1)}\lambda_d P_{(i,j,k,l-1)}\delta_{(i,j,k,l-1)} \\
&+(l+1)\mu_d P_{(i,j,k,l+1)}\delta_{(i,j,k,l+1)}
\end{aligned} \tag{2.14}$$

where $P_{(i,j,k,l)}$ is the steady state probability of being in state $S_{(i,j,k,l)}$.

Once the SSBEs are obtained for all states $S_{(i,j,k,l)} \in S$, numerical methods may be used to solve the system of equations given by the SSBEs plus the normalisation constraint:

$$\sum_{S_{(i,j,k,l)} \in S} P_{(i,j,k,l)} = 1 \tag{2.15}$$

The proposed analytical approach allows us to define a wide range of RAT selection policies taking into account several allocation criteria, such as service type, load, network conditions, etc. In particular, some of the policies presented in [1] and [12] will be adapted to our Markov model in the following subsections.

2.4.1 Random (RND) RAT Selection Policy

For illustrative purposes, this policy randomly selects the RAT on which the call/session will be carried out. Assume TDMA is selected randomly for voice and data users with a probability of α and β respectively. In the same way, WCDMA is selected with a probability $(1-\alpha)$ and $(1-\beta)$ for voice and data users correspondingly.

If the system is in a voice blocking state, i.e. $S_{(i,j,k,l)} \in S^T_{b,v}$ or $S_{(i,j,k,l)} \in S^W_{b,v}$, or in a data blocking state, i.e. $S_{(i,j,k,l)} \in S^T_{b,d}$ or $S_{(i,j,k,l)} \in S^W_{b,d}$, then the arrival rates of voice and data users to non-blocked states happen with probability equal to the unity. This ensures that a call/session will not be dropped due to the random allocation policy if resources exist in the opposite RAT to the one chosen by the policy. Then, the values of $\alpha_{(i,j,k,l)}$ and $\beta_{(i,j,k,l)}$ in (2.10) may be written as:

$$\begin{aligned}
\alpha_{(i,j,k,l)} &= \begin{cases} \alpha & \text{if } S_{(i,j,k,l)} \notin (S^T_{b,v} \cup S^W_{b,v}) \\ 1 & \text{if } S_{(i,j,k,l)} \in S^W_{b,v} \\ 0 & \text{if } S_{(i,j,k,l)} \in S^T_{b,v} \end{cases} \\
\beta_{(i,j,k,l)} &= \begin{cases} \beta & \text{if } S_{(i,j,k,l)} \notin (S^T_{b,d} \cup S^W_{b,d}) \\ 1 & \text{if } S_{(i,j,k,l)} \in S^W_{b,d} \\ 0 & \text{if } S_{(i,j,k,l)} \in S^T_{b,d} \end{cases}
\end{aligned} \tag{2.16}$$

2.4.2 Service-Based #1 (SB#1) RAT Selection Policy

This policy intends to allocate voice users to TDMA and data users to WCDMA. If the assignment is not possible, i.e. the chosen RATs are at full capacity, the voice users are directed to WCDMA and data users to TDMA.

Bearing this in mind, a voice arrival is not allowed in WCDMA, i.e. the transition $S_{(i,j,k,l)} \rightarrow S_{(i,j,k+1,l)}$ is not allowed, unless we are in a TDMA voice blocking state ($S_{(i,j,k,l)} \in S^T_{b,v}$). Moreover, a data session arrival will not be accommodated in TDMA, i.e. the transition $S_{(i,j,k,l)} \rightarrow S_{(i,j+1,k,l)}$ is not allowed, unless we are in a WCDMA data blocking state, that is $S_{(i,j,k,l)} \in S^W_{b,d}$. In order to take these restrictions into account in the global balance equations, the functions $\alpha_{(i,j,k,l)}$ and $\beta_{(i,j,k,l)}$ in (2.10) which define the feasibility of a voice arrival in WCDMA and the feasibility of a data arrival in TDMA can be defined as:

$$\begin{aligned}
\alpha_{(i,j,k,l)} &= \begin{cases} 1 & \text{if } S_{(i,j,k,l)} \notin S^T_{b,v} \\ 0 & \text{if } S_{(i,j,k,l)} \in S^T_{b,v} \end{cases} \\
\beta_{(i,j,k,l)} &= \begin{cases} 1 & \text{if } S_{(i,j,k,l)} \in S^W_{b,d} \\ 0 & \text{if } S_{(i,j,k,l)} \notin S^W_{b,d} \end{cases}
\end{aligned} \tag{2.17}$$

2.4.3 Service-Based #2 (SB#2) RAT Selection Policy

This policy, acting as opposite to the SB#1 policy, intends to allocate voice users to WCDMA and data users to TDMA. If the assignment is not possible, i.e. the chosen RATs are at full capacity, the voice users are directed to TDMA and data users to WCDMA.

Keep in mind that a voice call will be not admitted in TDMA, i.e. the transition $S_{(i,j,k,l)} \rightarrow S_{(i+1,j,k,l)}$ will be not allowed, unless no capacity is left for voice users in WCDMA, that is $S_{(i,j,k,l)} \in S_{b,v}^{W}$. Similarly, data users will be admitted in WCDMA, i.e. the transition $S_{(i,j,k,l)} \rightarrow S_{(i,j,k,l+1)}$, only if no capacity is left in TDMA to accommodate the data session, i.e. $S_{(i,j,k,l)} \in S_{b,d}^{T}$. In order to account for these limitations in the arrival rates, functions $\alpha_{(i,j,k,l)}$ and $\beta_{(i,j,k,l)}$ in (2.10) denoting the feasibility of data arrival rates in TDMA and of voice arrival rates in WCDMA can be expressed as:

$$\begin{aligned} \alpha_{(i,j,k,l)} &= \begin{cases} 1 & \text{if } S_{(i,j,k,l)} \in S_{b,v}^{W} \\ 0 & \text{if } S_{(i,j,k,l)} \notin S_{b,v}^{W} \end{cases} \\ \beta_{(i,j,k,l)} &= \begin{cases} 1 & \text{if } S_{(i,j,k,l)} \notin S_{b,d}^{T} \\ 0 & \text{if } S_{(i,j,k,l)} \in S_{b,d}^{T} \end{cases} \end{aligned} \tag{2.18}$$

2.4.4 Load Balancing (LB) RAT Selection Policy

The load balancing (LB) policy intends to allocate users to the RAT that undergoes a lower load situation at a given time. In particular, transitions between a source state and possible destination states will depend on the measured load at each destination state.

Before expressing this notion in terms of transition rates in our Markov model, it is convenient to define the load metrics in both RATs.

In TDMA-based GSM/EDGE, the TSL utilization factor, initially defined in [8], may be used to measure the load in a given state $S_{(i,j,k,l)} \in S$ as:

$$L_{(i,j)}^{T} = n_{(i,j)}/C \tag{2.19}$$

where C is the total number of available channels (TSLs) in the cell devoted to voice and data traffic services and $n_{(i,j)}$ is the number of occupied channels (TSLs) when i voice users and j data users are currently being served in TDMA. For the case of data users requiring a single slot for their uplink connection, $n_{(i,j)} = \min(C, i+j)$. Note that this definition of load will not account for multiple users sharing a same TSL nor users using multiple TSLs.

On the other hand, the load in a WCDMA-based system may be calculated by means of the uplink load factor $L_{(k,l)}^{W}$, defined in (2.5), scaled by $\eta_{\max}$.

In order to determine whether the incoming user demanding a given service should be allocated to TDMA or to WCDMA, functions $\alpha_{(i,j,k,l)}$ and $\beta_{(i,j,k,l)}$ in (2.10) will take the following values:

$$\begin{aligned}
\alpha_{(i,j,k,l)} &= \begin{cases} 1 & \text{if } (L^T_{(i+1,j)} < L^W_{(k+1,l)}/\eta_{\max}) \text{ or} \\ & \text{if } (S_{(i,j,k,l)} \notin S^T_{b,v} \wedge S_{(i,j,k,l)} \in S^W_{b,v}). \\ 0 & \text{if } (L^T_{(i+1,j)} > L^W_{(k+1,l)}/\eta_{\max}) \text{ or} \\ & \text{if } (S_{(i,j,k,l)} \in S^T_{b,v} \wedge S_{(i,j,k,l)} \notin S^W_{b,v}). \\ 0.5 & \text{otherwise} \end{cases} \\
\beta_{(i,j,k,l)} &= \begin{cases} 1 & \text{if } (L^T_{(i,j+1)} < L^W_{(k,l+1)}/\eta_{\max}) \text{ or} \\ & \text{if } (S_{(i,j,k,l)} \notin S^T_{b,d} \wedge S_{(i,j,k,l)} \in S^W_{b,d}). \\ 0 & \text{if } (L^T_{(i,j+1)} > L^W_{(k,l+1)}/\eta_{\max}) \text{ or} \\ & \text{if } (S_{(i,j,k,l)} \in S^T_{b,d} \wedge S_{(i,j,k,l)} \notin S^W_{b,d}). \\ 0.5 & \text{otherwise} \end{cases}
\end{aligned} \tag{2.20}$$

which account for the load levels in each of the corresponding RATs given voice call and data session arrivals.

2.4.5 Multi-Mode Terminal Driven (MMTD) RAT Selection Policy

With the purpose of taking advantage of terminal availability characteristics, we may use this information to decide the most appropriate RAT for an incoming call/session. In this sense, we may attempt to allocate single-mode users to TDMA and multi-mode users to WCDMA. Multi-mode users would eventually be allocated to TDMA if no capacity was left in WCDMA. With this policy we try to minimise the impact of single-mode terminals being served in TDMA given the higher allocation flexibility of multi-mode terminals. To account for the situations where no voice or data capacity is available in WCDMA and consequently multi-mode users are allocated, if possible, in TDMA, we define the following indicator functions, $\alpha_{(i,j,k,l)}$ and $\beta_{(i,j,k,l)}$ in (2.13), for each feasible state $S_{(i,j,k,l)}$:

$$\begin{aligned}
\alpha_{(i,j,k,l)} &= \begin{cases} 1 & \text{if } S_{(i,j,k,l)} \notin S^W_{b,v} \\ 0 & \text{if } S_{(i,j,k,l)} \in S^W_{b,v} \end{cases} \\
\beta_{(i,j,k,l)} &= \begin{cases} 1 & \text{if } S_{(i,j,k,l)} \notin S^W_{b,d} \\ 0 & \text{if } S_{(i,j,k,l)} \in S^W_{b,d} \end{cases}
\end{aligned} \tag{2.21}$$

2.5 Performance Metrics

In order to compute the steady state probabilities $P_{(i,j,k,l)}$ we must solve the global steady-state balance equations given by the application of the aforementioned RAT selection policies for all feasible states $S_{(i,j,k,l)} \in S$. This may be carried out using numerical methods; in particular, an iterative power procedure will be utilized for such task [13]. In general, the dimensionality of the Markov chain, M_d, can be computed as the product of K available RATs and J supported services (assuming all services are supported on all RATs). Obviously, the higher the number of services and/or RANs, the higher the dimensionality of our model. As a result, the computational complexity to numerically solve these systems is a well-known fact and increases with the state dimension. Nevertheless, typically up to 3 or 4 RANs are available and, although services are high in number, not all RATs support all services, which may lower the impact on the Markov dimensionality. In addition, the higher the number of states in the Markov model, N_s, the more computation resources are needed as explained in the following. For the particular case of two services, voice and data, along with two RATS, i.e. TDMA and WCDMA, the resulting number of states can be computed as:

$$N_s = N_s^T \cdot N_s^W \tag{2.22}$$

with the number of states in TDMA, N_s^T, being

$$N_s^T = \frac{(C+1)(n_C C+2)}{2} \tag{2.23}$$

and the number of states in WCDMA, N_s^W, yielding

$$N_s^W \approx \left\lfloor \frac{(\eta_{\max} \cdot \left[\frac{W/R_{b,v}}{(E_b/N_0)_v} + 1 \right] + 1)(\eta_{\max} \cdot \left[\frac{W/R_{b,d}}{(E_b/N_0)_d} + 1 \right] + 1)}{2} \right\rfloor \tag{2.24}$$

where $\lfloor x \rfloor$ denotes the integer value of x.

In our case, the iterative power method is used to solve the system of equations provided in (2.14). Its operation is based on iteratively performing the product of a probability vector $\mathbf{p}$ (of dimension $N_s \times 1$) with the $N_s \times N_s$ transition probability matrix ($\boldsymbol{P}$). If k iterations are needed for convergence then a total number of $k \times N_s^2$ multiplications are needed. The number of k iterations needed to satisfy convergence is based on the following relative measure [13]:

$$\max_i \left(\frac{\left| p_i^{(k)} - p_i^{(k-1)} \right|}{p_i^{(k)}} \right) < \varepsilon \tag{2.25}$$

where $p_i^{(k)}$ are elements of vector $\mathbf{p^{(k)}}$, which denotes the probability distribution after the *k-th* iteration, and ε is the required solution accuracy which is in our case set to 10^{-6}.

Fortunately, matrix $\boldsymbol{P}$ is usually sparse, i.e. it contains a large amount of zero entries. Then, if N_z is the total number of non-zero entries in matrix $\boldsymbol{P}$, a total of $k \times N_z$ multiplications are now required. In this sense, the limiting factor would be in terms of memory storage requirements rather than in terms of computational complexity of operation and solution convergence time. Nevertheless, state-of-the-art computers are able to support these high memory storage requirements.

Then, performance metrics may be directly derived from the steady state probabilities, $P_{(i,j,k,l)}$, as described in the following.

2.5.1 Blocking Probabilities

Making use of the blocking state sets defined formerly in section 2.3.3, the generalized form of the blocking probability of a service type σ in a given RAT ρ may be expressed as:

$$P_{b,\sigma}^{\rho} = \sum_{S_{(i,j,k,l)} \in S_{b,\sigma}^{\rho}} P_{(i,j,k,l)} \tag{2.26}$$

with $\sigma = \{v, d\}$ and $\rho = \{T, W\}$.

If we are interested in the blocking probability of a particular service type σ over all the possible RATs, this can be computed as:

$$P_{b,\sigma} = \sum_{S_{(i,j,k,l)} \in S_{b,\sigma}} P_{(i,j,k,l)} \tag{2.27}$$

Finally, the total blocking probability may be computed as:

$$P_b = \sum_{S_{(i,j,k,l)} \in S_b} P_{(i,j,k,l)} \tag{2.28}$$

2.5.2 Carried Traffic

The average carried traffic, or average number of users, may also be computed from the steady state probabilities $P_{(i,j,k,l)}$. The fractional average number of users demanding a given service σ in a given RAT ρ can be derived numerically from:

$$N_{\sigma}^{\rho} = E[x] \quad \text{with } x = \begin{cases} i & \text{if } \rho = T,\, \sigma = v \\ j & \text{if } \rho = T,\, \sigma = d \\ k & \text{if } \rho = W,\, \sigma = v \\ l & \text{if } \rho = W,\, \sigma = d \end{cases} \tag{2.29}$$

and $E[x]$ the expectation of x defined as:

$$E[x] = \sum_{S_{(i,j,k,l)} \in S} x \cdot P_{(i,j,k,l)} \tag{2.30}$$

Similarly, the average number of users in each RAT ρ is computed as:

$$N^{\rho} = N_{v}^{\rho} + N_{d}^{\rho} \tag{2.31}$$

The per-service average number of users in the system is defined by:

$$N_{\sigma} = N_{\sigma}^{T} + N_{\sigma}^{W} \tag{2.32}$$

Finally, the total average number of users in the system yields:

$$N = N_{v}^{T} + N_{d}^{T} + N_{v}^{W} + N_{d}^{W} \tag{2.33}$$

2.5.3 System Load

Load metrics are also key performance indicators which can be obtained from the steady state probabilities. Bearing in mind the load definitions given in (2.19) and (2.5), the average TSL utilization factor in TDMA yields:

$$L^{T} = E[L_{(i,j)}^{T}] \tag{2.34}$$

and the average uplink load factor in WCDMA may be computed as:

$$L^{W} = E[L_{(k,l)}^{W}] \tag{2.35}$$

2.5.4 Peak Throughput

Throughput definitions are also intrinsic to the underlying access scheme and will be, consequently, defined individually for TDMA and WCDMA systems.

2.5.4.1 TDMA Throughput

The throughput in TDMA at a given state $S_{(i,j,k,l)}$ can be expressed as the sum of voice and data throughput contributions as:

$$\Gamma^T_{(i,j)} = i \cdot \kappa_v + \min(C - i, j) \cdot \kappa_d \tag{2.36}$$

where κ_v and κ_d are the voice and data TSL bit rates respectively and the term $\min(C - i, j)$ accounts for the number of data users transmitting at κ_d bits per second. If i voice users are being served in TDMA, and they require a whole TSL, then at most $(C - i)$ data users will be able to transmit at κ_d bits per second. If $j < (C - i)$, then j data users transmit at κ_d bits per second.

It is important to note that although the throughput per voice user will be κ_v, for data users, the effect of TSL sharing will contribute to a decrease in throughput per data user as the number of multiplexed data TSLs increases. Actually, the throughput per data user will be equal to $\kappa_d \cdot \min(C - i, j)/j$.

Then, the total average throughput in TDMA becomes:

$$\Gamma^T = E[\Gamma^T_{(i,j)}] \tag{2.37}$$

2.5.4.2 WCDMA Throughput

Throughput delivered in WCDMA-based systems at a given state $S_{(i,j,k,l)}$ can be calculated as:

$$\Gamma^W_{(k,l)} = k \cdot R_{b,v} + l \cdot R_{b,d} \tag{2.38}$$

where $R_{b,\sigma}$ is the granted bit rate of a σ service type user.

Then, the average throughput in WCDMA is obtained as:

$$\Gamma^W = E[\Gamma^W_{(k,l)}] \tag{2.39}$$

2.5.4.3 Total Aggregate Throughput

Considering the combined throughput carried by both RATs, TDMA and WCDMA, the total aggregate throughput, Γ_A, becomes:

$$\Gamma_A = \Gamma^T + \Gamma^W \tag{2.40}$$

2.6 Results

In order to illustrate the performance of the presented RAT selection policies, the GSM/EDGE Radio Access Network (GERAN) and the UMTS Radio Access Network (UTRAN) will be used as representatives of TDMA and WCDMA technologies respectively.

The performance of the system is evaluated under different offered voice and data traffic loads, T_v and T_d, where $T_v = \lambda_v/\mu_v$ and $T_d = \lambda_d/\mu_d$. The considered system parameters for numerical evaluation are represented in Table 2.1.

Under these assumptions, and as will be shown in the following numerical results, UTRAN exhibits a higher capacity in terms of maximum number of allowable voice and data users as compared to GERAN. Indeed, the $C = 8$ channels in GERAN correspond to a 200 kHz bandwidth single-carrier configuration, while for UTRAN a total bandwidth of 5 MHz is available [9].

2.6.1 Markov Model Validation

In order to validate the results provided by the Markov model, a system-level simulator has been developed. This simulator assumes a more realistic behavior than the model by considering that data users intend to transmit a particular amount of data (bits), which follows a Pareto distribution [14]. In this case, the data holding time will depend on the bit rate allocated to the user in the selected RAT rather than being modeled by an exponential distribution. RAT selection is performed and CAC procedures follow the same principles as the state feasibility conditions imposed for the Markov model. Once users are allocated in the appropriate RAT, statistics are measured on a discrete-time basis. In addition, another simulator considering the same assumptions as in the Markov model has been used to validate the correctness of the algebra, but not shown here due to lack of space.

Table 2.1: System Parameters For Numerical Evaluation

Parameter	Symbol	Value
Number of channels in GERAN	C	8
Maximum number of simultaneous users sharing a same TSL in GERAN	n_C	3
Bit rate for voice users in GERAN	κ_v	12.2 kbps
Bit rate for data users in GERAN	κ_d	44.8 kbps
Chip-rate in UTRAN	W	3.84 Mcps
Required bit-energy-to-noise-density ratio for voice traffic in UTRAN	$(E_b/N_0)_v$	6 dB
Required bit energy-to-noise-density ratio for data traffic in UTRAN	$(E_b/N_0)_d$	5 dB
Bit rate for voice users in UTRAN	$R_{b,v}$	12.2 kbps
Bit rate for data users in UTRAN	$R_{b,d}$	44.8 kbps
Maximum UL load factor	η_{max}	1
Multi-mode terminal availability (unless otherwise stated)	p	1

In the following, we compare the results obtained via the Markov model with the results obtained through simulation using the Load Balancing criterion for the RAT selection policy. Fig. 2.3 shows the loads in GERAN and UTRAN considering an offered traffic load of 16 and 8.33 Erlangs and a range of voice traffic for the LB case. Loads in both RATs tend to follow each other as the total offered traffic increases. Note how the simulated values (represented as bullets) follow the same trend to those obtained via the Markov model and a good matching exists. Fig. 2.4 shows the number of served users of each service in UTRAN and GERAN. Bearing in mind that, with the current parameter setting (see Table 2.1), GERAN offers a much lower capacity than UTRAN (i.e. one single carrier of 200 kHz for GERAN as opposed to 5 MHz for UTRAN); many more users are needed in UTRAN in order to balance the loads. The number of data users in each RAT is kept constant while offered voice traffic varies between 1.2 and 36 Erlangs. For 84 Erlangs, data users are forced to share TSLs in GERAN, explaining the increase/decrease of data users in GERAN and UTRAN respectively. Fig. 2.5 shows the total voice blocking probability in the combined GERAN/UTRAN system as defined in (2.27). Clearly, the higher the offered traffic the higher the blocking probability gets. Again, the

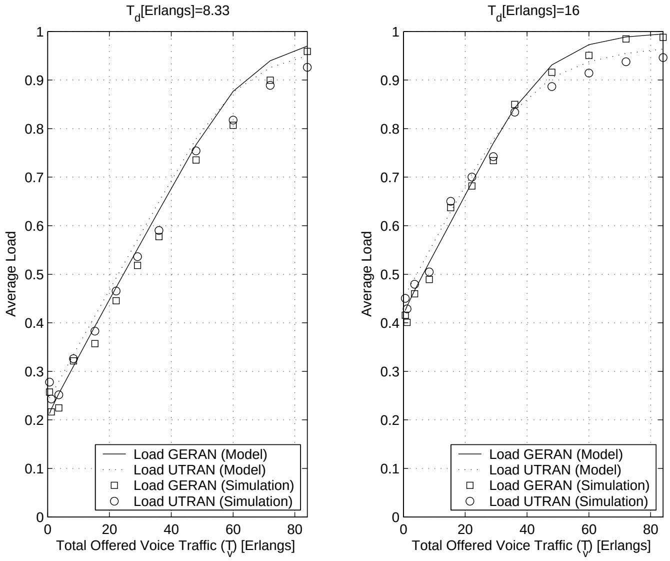


Figure 2.3: Average load in each RAT under varying traffic.

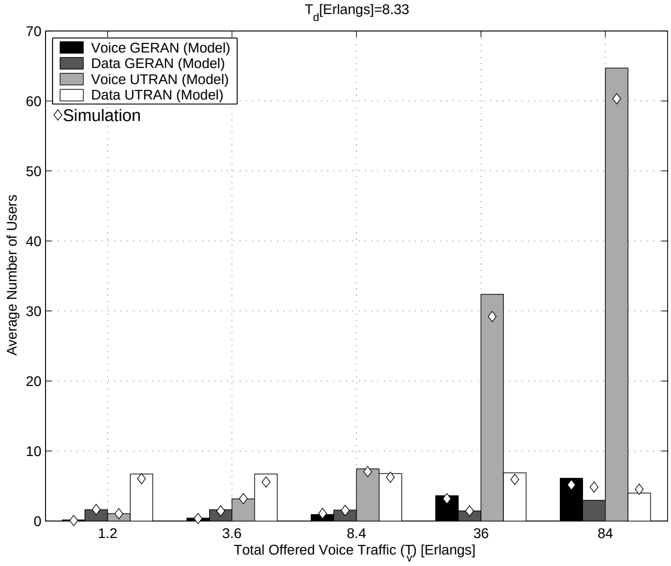


Figure 2.4: Average number of served users in each RAT.

simulated results (marked with bullets) match the Markov model behavior. Finally, Fig. 2.6 shows the throughput performance in each of the RATs for both the model and the simulated approaches under varying traffic conditions. At this point, the proposed Markov model has been validated, as shown in Figs. 2.3-2.6, and its

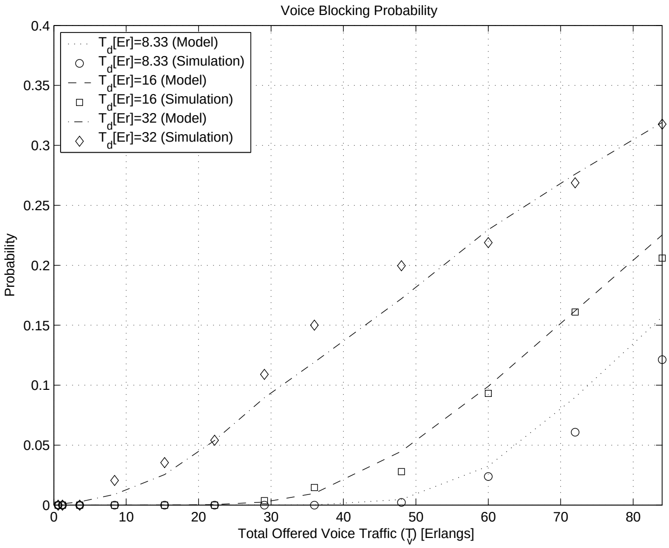


Figure 2.5: Total voice blocking probabilities under varying traffic.

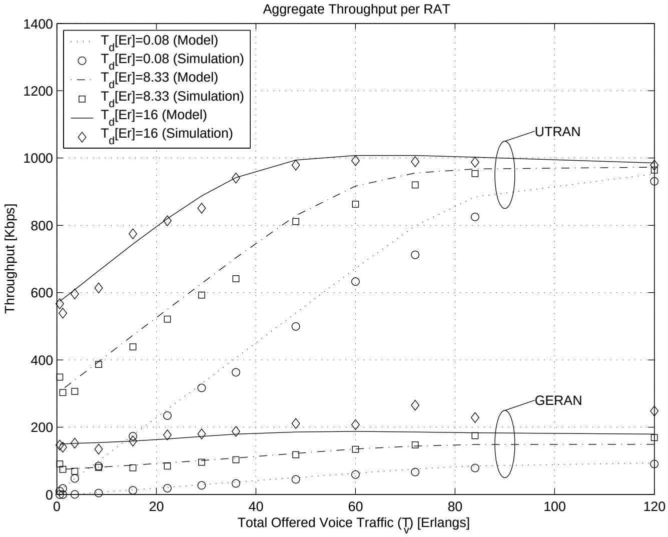


Figure 2.6: Total throughput per RAT under varying traffic.

suitability for testing different RAT selection policies confirmed. The comparison of such RAT selection policies is provided in the forthcoming subsections.

2.6.2 RAT Selection Policies Comparison

This section provides illustrative results depicting the behavior of the presented RAT selection policies (except for policy MMTD, given that $p = 1$ and thus does not apply in this case). The focus is set on how the different policies allocate different services over the existing RATs by means of representing the probability, $P_{(i,j,k,l)}$, of having a given number of voice and data users in each RAT for several traffic mix conditions. In the following, statistical user distribution is represented with 2D discrete graphs with axis indicating the number of voice and data users, and probabilities depicted by grey-scaled shaded regions, where dark regions indicate high probability values and light regions low probability values. Step-wise admission limits are plotted for both systems (denoted as *feasibility region*) and, for the case of GERAN, the limit upper-bounding the region where data users are not sharing resources is also plotted (which is denoted as *data non-reuse region*).

In this sense, Fig. 2.7 shows the user distribution provided by policy SB#1. For a traffic mix of $T_v = 3.6$ and $T_d = 16$ Erlangs, GERAN is able to handle its share of voice users while UTRAN manages the offered data traffic. This causes users in GERAN being exclusively distributed over the voice user axis with no data users at all (Fig. 2.7 upper-left). Accordingly, in UTRAN, users are spanned over the data user axis with no voice user component (Fig. 2.7 upper-right). If voice traffic is increased so that GERAN is not able to handle all the requests, SB#1 policy will re-direct voice traffic to UTRAN. Consequently, user distribution in GERAN is concentrated on the maximum number of allowed voice users (Fig. 2.7 lower-left). In UTRAN, users are now distributed over both axes due to overflowed voice traffic from GERAN (Fig. 2.7 lower-right). For the case of SB#2 policy, see Fig. 2.8, an analogous study may be made on the statistical user distributions. When both GERAN and UTRAN are able to manage their shares of data and voice users respectively (Fig. 2.8 upper left and right), data and voice user distributions lay on the corresponding axis in GERAN and UTRAN respectively. If data traffic is increased so that GERAN is unable to handle all data traffic, UTRAN will have to manage with its share of voice users plus data users that could not be allocated in GERAN. This behavior can be observed in Fig. 2.8 lower left and right figures. It is worth noting how SB#2 provides high reuse of data resources in GERAN. The rationale behind LB policy is to maintain both loads in GERAN and UTRAN at the same level. By doing so, and according to the load definitions presented earlier on, data users in GERAN will not be forced to share resources until UTRAN is fully loaded. This may be observed in Fig. 2.9. For offered traffic values of $T_v = 36$ and $T_d = 8.33$ Erlangs, UTRAN remains in a half-loaded situation given that user distribution remains far from the admission limit (see Fig. 2.9 upper-right).

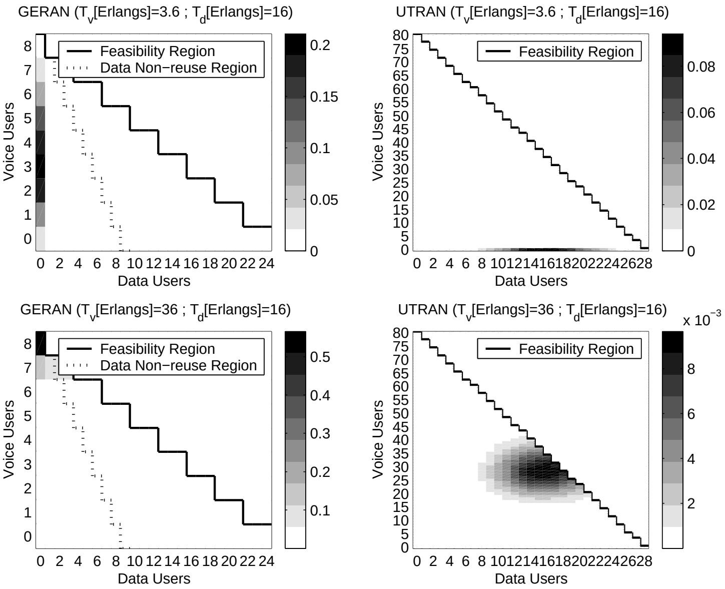


Figure 2.7: Statistical user distributions in GERAN and UTRAN with policy SB#1 for two different service mixes.

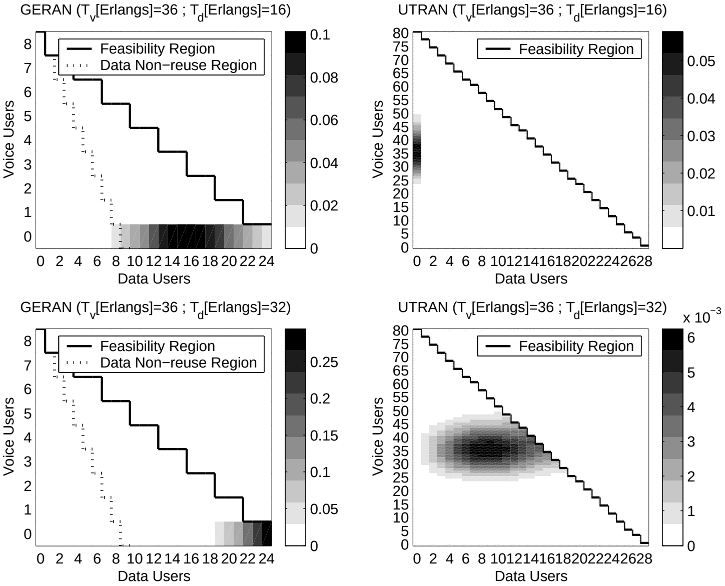


Figure 2.8: Statistical user distributions in GERAN and UTRAN with policy SB#2 for two different service mixes.

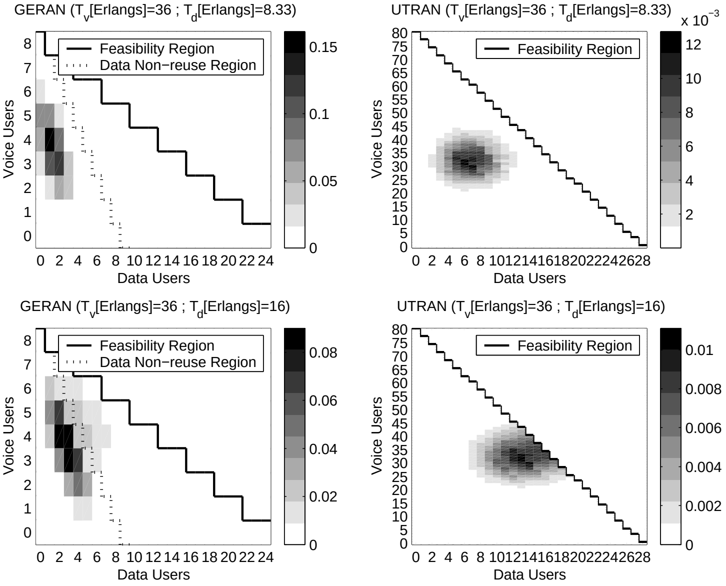


Figure 2.9: Statistical user distributions in GERAN and UTRAN with policy LB for two different service mixes.

Consequently, in GERAN (Fig. 2.9 upper-left) data users do not share resources since user distribution lies below the data non-reuse region indicated by the dotted line. If traffic is increased such that UTRAN achieves fully loaded situations (Fig. 2.9 lower-right), data users in GERAN will then start to share resources which is indicated by user distribution falling above the data non-reuse region as observed in Fig. 2.9 lower-left. Finally, Fig. 2.10 shows the statistical user distribution when applying policy RND. As compared to LB policy, RND policy forces data users to share resources in GERAN even if they could be more efficiently managed by UTRAN system. Moreover, higher blocking situations in GERAN are achieved indicated by the proximity of user distribution to the admission limit.

2.6.3 Throughput Comparison

In the following, a comparison between the different RAT selection policies by evaluating a number of different performance measures is provided. Fig. 2.11 shows the performance in terms of aggregate throughput for the different presented RAT selection policies, i.e. SB#1, SB#2 and LB (where RND policy is not shown given its poor performance and the lack of space). For medium data traffic loads (i.e. $T_v = 16$ and $T_d = 32$ Erlangs), we can clearly see how LB outperforms all other policies, with the worst overall behavior corresponding to SB#2. As it will be shown

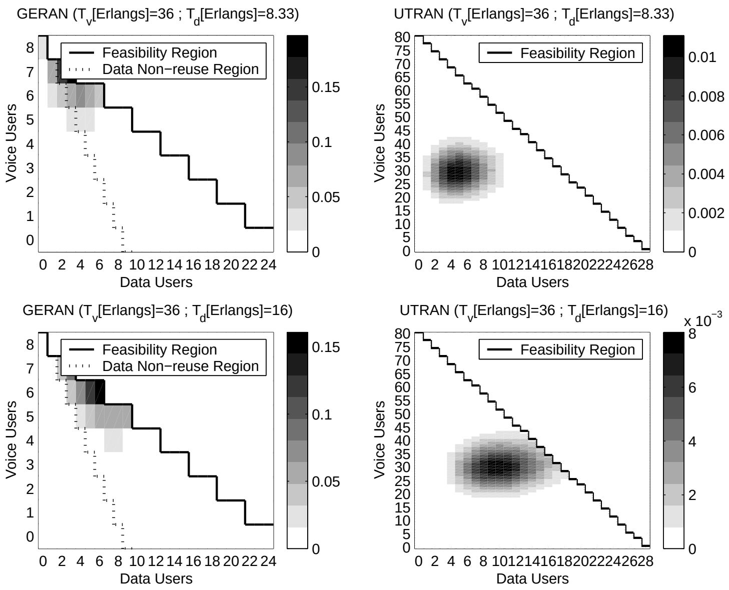


Figure 2.10: Statistical user distributions in GERAN and UTRAN with policy RND for two different service mixes.

in the following, a major cause of throughput degradation is due to excessive TSL sharing in GERAN. In this sense, SB#1 allocates data users to UTRAN, so this problem is partially avoided. However, when voice load is increased, data users may be also allocated in GERAN causing throughput to exhibit similar performances than for other policies. In contrast, SB#2 allocates data users to GERAN causing high TSL reuse even for low voice loads thus negatively impacting on the total aggregate throughput. Notice that for the case of LB policy, load definitions provided in (2.19) and (2.5) prevent data users in GERAN to share TSL with other users unless UTRAN is fully loaded. Finally, an overall throughput degradation is noted when the offered voice traffic increases, in particular when the offered data traffic is $T_d = 32$ Erlangs (Fig. 2.11 left). This is caused by a major number of admitted voice users which contribute with lower throughput values (12.2 kbps for a single voice user) than those offered by data users (a maximum of 44.8 kbps for a single data user). As previously mentioned, a cause for aggregate throughput degradation may be found in the excessive reuse of data TSLs in GERAN. Fig. 2.12 shows, for several offered traffic configurations, the total throughput per data user when each of the considered RAT selection policies is applied. As it can be observed, SB#2 policy provides an excessive reuse of TSLs and thus throughput per data user is lower than that of SB#1 and LB policies. On the other hand, SB#1 policy will direct data users to UTRAN which exhibits a better throughput performance and therefore data users are less penalized by TSL reuse. Finally,

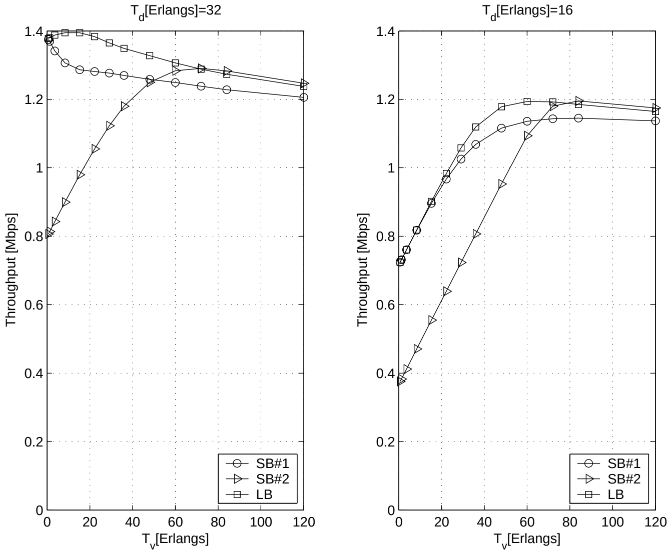


Figure 2.11: Aggregate throughput values for policies SB#1, SB#2 and LB under varying traffic conditions.

LB policy will prevent data sharing in GERAN as long as UTRAN may handle offered traffic. In these cases, throughput per data user provided by LB policy is maximum, i.e. 44.8 kbps (see Fig. 2.12 upper left and right). However, if traffic increases, UTRAN will no longer be able to manage all traffic and thus LB will force data users in GERAN to share TSL which will consequently cause throughput per data user to decrease. Another important issue to take into account when designing RAT selection policies, is to provide a high ratio of admitted users in the system or, equivalently, to provide low blocking probabilities. Fig. 2.13 illustrates the blocking probability obtained by the presented policies when offered data traffic is $T_v = 16$ and $T_d = 32$ Erlangs. Clearly, SB#2 provides a lower blocking probability as compared to SB#1 and LB policies. In GERAN, the allocation of a voice user implies a timeslot consumption of $1/C$ while as for a data user this consumption is $1/n_C C$. Therefore, it is more resource-consuming (n_C times more) to allocate voice users in GERAN than data users. On the other hand, the resource consumption in UTRAN may be quantified by means of the load factor definition, given in (2.5), with $[W/R_{b,v}(E_b/N_0)_v + 1]^{-1}$ and $[W/R_{b,d}(E_b/N_0)_d + 1]^{-1}$ the fractions of loads consumed by voice and data users respectively. Bearing in mind the simulation parameters given in Table 2.1, it can be shown that a data user in UTRAN demands more resources than a voice user. In this sense, it is much more suitable in the considered scenario to allocate voice users in UTRAN and data users in GERAN, thus explaining the better behavior of SB#2 as opposed to SB#1. LB on the other side will provide a performance in-between SB#2 and SB#1.

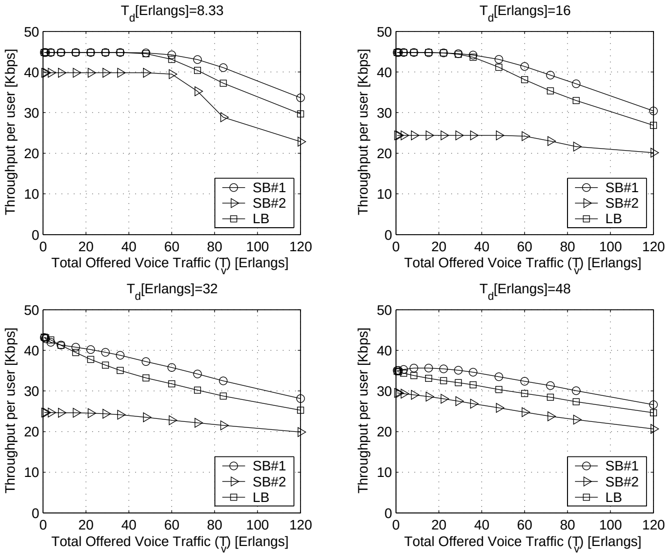


Figure 2.12: Throughput per data user for different RAT selection policies and traffic mixes.

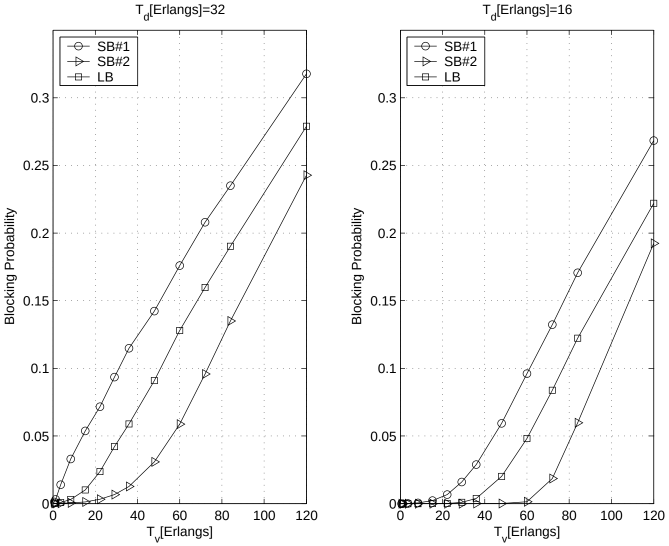


Figure 2.13: Total blocking probabilities for policies SB#1, SB#2 and LB.

As previously stated, to choose the most suitable RAT selection policy may depend on a number of quality requirements. In particular, it has been studied that both the data throughput per user and the blocking probability can decide the appropriateness of one RAT selection policy with respect to another. Therefore, in the following, the achievable capacity in terms of the maximum number of allowable

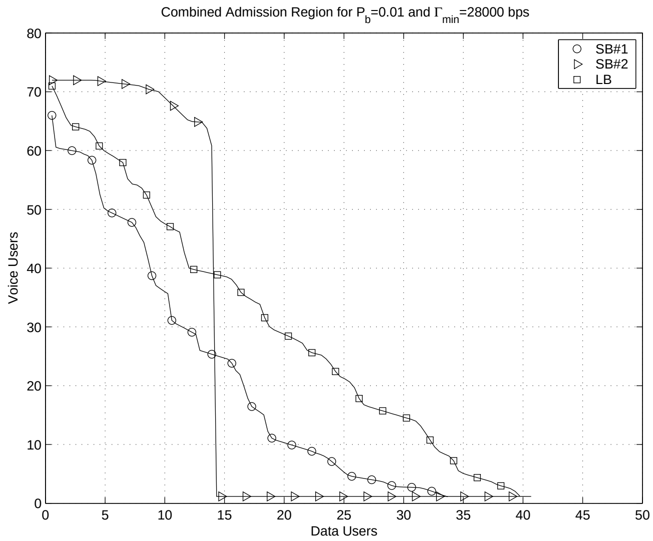


Figure 2.14: Combined admission regions for policies SB#1, SB#2 and LB in a blocking probability limited scenario.

voice and data users (admission regions) satisfying some blocking probability and data throughput per user requirements will be studied. Based on these measurements, the most appropriate RAT selection policy may be decided.

Fig. 2.14 shows the admission regions achieved by the different policies under equal traffic conditions. In this first case, a blocking probability limited scenario is considered where it is required a maximum blocking probability of $P_b = 0.01$ and a minimum data throughput per user of $\Gamma_{min} = 28$kbps (recall that the maximum achievable throughput per data user is 44.8kbps). Results in Fig. 2.14 indicate that although SB#2 policy provides a larger admission region than SB#1, given by a better response in terms of blocking probability, the SB#2 policy is severely limited by the data throughput per user requirements for high data loads. This causes policy LB to better trade-off both blocking probability and throughput requirements.

Fig. 2.15, considers a data throughput per user limited scenario, where the target blocking probability is set to 5% and the minimum required data throughput per user is $\Gamma_{min} = 44$kbps. In this case, SB#2 no longer provides a larger admission region than SB#1 due to the throughput requirements. As expected, SB#1 provides lower data resource utilization in GERAN thus being capable of allocating somewhat more users than policy LB.

Finally, if the system is limited by both blocking and throughput requirements, via setting $P_b = 0.01$ and $\Gamma_{min} = 44$kbps, the performance shown in Fig. 2.16 is

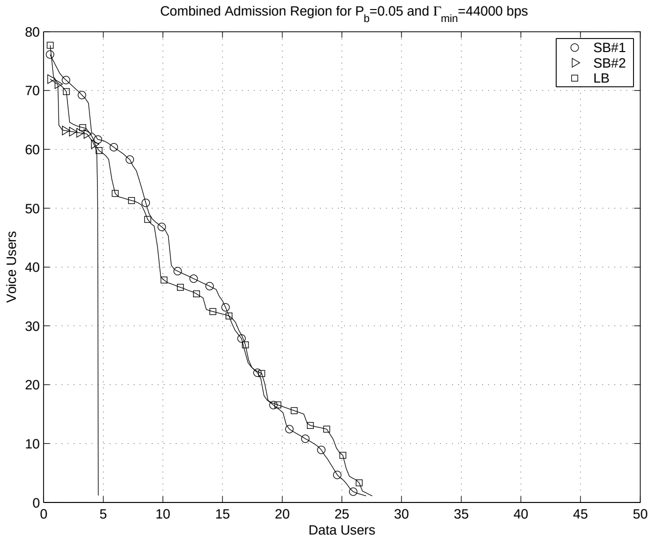


Figure 2.15: Combined admission regions for policies SB#1, SB#2 and LB in a throughput limited scenario.

obtained. Once again, throughput requirements are too severe for SB#2 and thus provides a smaller admission region than SB#1 and LB. LB on the other hand exhibits a better performance in terms of throughput and blocking probability than SB#1, therefore providing the largest admission region.

2.6.4 Multi-mode Terminal Availability Impact on Initial RAT Selection

Fig. 2.17 reflects the impact of multi-mode terminal availability over the performance in terms of total aggregate throughput when using policies SB#1 (left) and LB (right). Specifically, we plot the normalized throughput degradation measured as the relative difference in aggregate throughput when single-mode terminals are present with respect to the case of all terminals being multi-mode, mathematically expressed as:

$$D^p = (\Gamma_T^1 - \Gamma_T^p)/\Gamma_T^1 \tag{2.41}$$

where Γ_T^p is the total aggregate throughput for a multi-mode terminal probability equal to p. The overall behavior of having low number of multi-mode terminals is translated into a higher throughput degradation which is represented in Fig. 2.17.

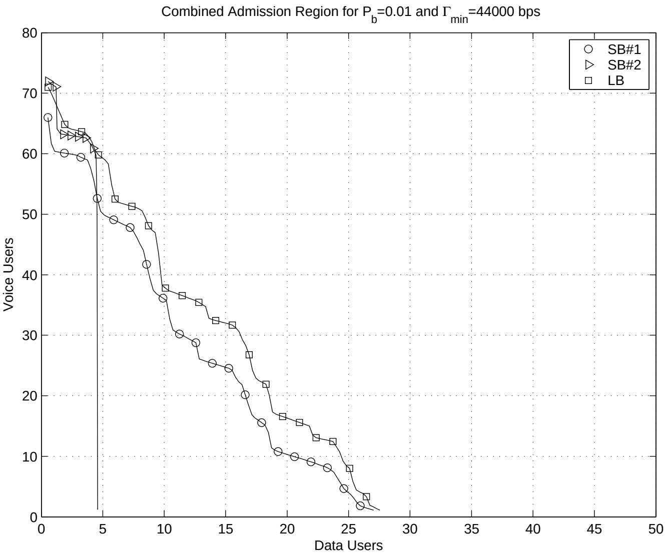


Figure 2.16: Combined admission regions for policies SB#1, SB#2 and LB in a blocking probability and throughput limited scenario.

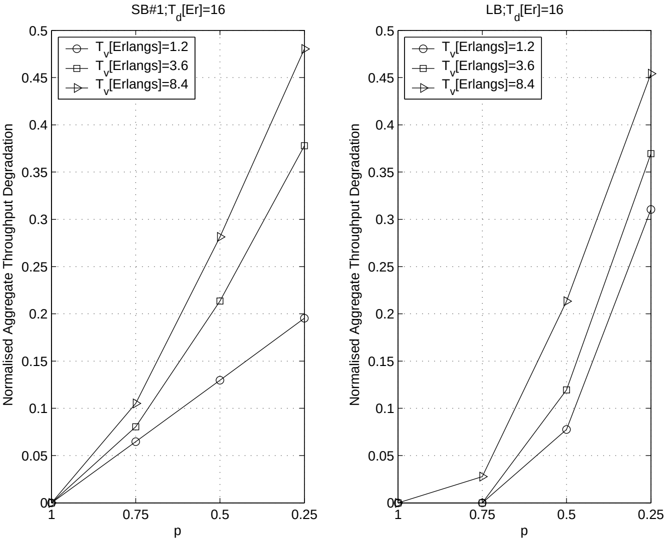


Figure 2.17: Impact of multi-mode terminal probability (p) when applying SB#1 and LB policies.

Fig. 2.18 compares, in terms of aggregate throughput, policies MMTD, SB#1 and LB in a scenario with 50% of terminals with multi-mode capabilities. With SB#1, GERAN handles voice users plus single-mode users thus showing a poorer performance than MMTD which allocates voice multi-mode users to UTRAN. On

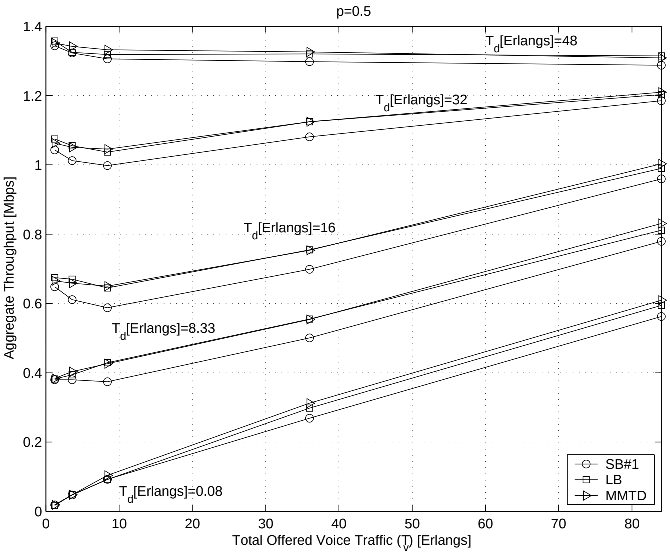


Figure 2.18: Total aggregate throughput provided by SB#1, LB and MMTD policies.

the other hand, LB policy shows higher flexibility in allocating multi-mode users as compared to SB#1. As a result, similar performance is observed between LB and MMTD.

2.7 Chapter Summary

It is widely established that RAT selection procedures in multi-service/multi-access scenarios play a key role in the provision of CRRM functionalities. In this chapter, a Markovian framework for the allocation of multiple services in multiple RATs is presented. It allows the evaluation of several RAT selection policies considering different allocation criteria which are fully embedded in the model. In addition, the model captures the availability of multi-mode terminals so as to consider the flexibility constraints of single-mode terminals. In particular, two different underlying radio access schemes are studied: Time Division Multiple Access (TDMA) and Wideband Code Division Multiple Access (WCDMA). In this context, generic voice and data sessions are to be allocated to the aforementioned RATs given particular RAT selection policies which comprise: two service-based schemes, namely SB#1 and SB#2, along with a load-balancing (LB) and a terminal-driven (MMTD) schemes, and finally, a random policy (RND). Results have confirmed the validity and suitability of the model which has been evaluated for the aforementioned

RAT selection policies. Results indicate that a trade-off between the average data throughput per user and the total blocking probability arises when comparing SB#1 and SB#2. As revealed, this trade-off may be suitably managed by the appliance of the LB policy. Finally, RAT selection is also performed taking into account the multi-mode terminal availability information, indicating that this input must not be avoided to achieve a higher utilization of the offered resources.

Bibliography

[1] J. Pérez-Romero, O. Sallent, and R. Agustí, "Policy-based Initial RAT Selection algorithms in Heterogeneous Networks," in *7th IFIP International Conference on Mobile and Wireless Communication Networks, 2005 (MWCN'05)*, September 2005.

[2] Gábor Fodor, Anders Furuskär, and Johan Lundsjö, "On access selection techniques in always best connected networks," in *In Proc. ITC Specialist Seminar on Performance Evaluation of Wireless and Mobile Systems*, 2004.

[3] O. Yilmaz, A. Furuskar, J. Pettersson, and A. Simonsson, "Access Selection in WCDMA and WLAN Multi-Access Networks," in *2005 IEEE 61st Vehicular Technology Conference*. 2005, pp. 2220–2224, IEEE.

[4] Susan Lincke-Salecker and Cynthia S. Hood, "Integrated networks that overflow speech and data between component networks," *International Journal of Network Management*, vol. 12, no. 4, pp. 235–257, 2002.

[5] A. Furuskar and J. Zander, "Multiservice allocation for multiaccess wireless systems," vol. 4, no. 1, pp. 174–184, 2005.

[6] I. Koo, A. Furuskar, J. Zander, and Kiseon Kim, "Erlang capacity of multiaccess systems with service-based access selection," vol. 8, no. 11, pp. 662–664, 2004.

[7] Gunter Bolch, Stefan Greiner, Hermann de Meer, and Kishor S. Trivedi, *Queueing Networks and Markov Chains: Modeling and Performance Evaluation with Computer Science Applications*, WileyBlackwell, 2nd edition edition, May 2006.

[8] T. Halonen, J. Romero, and J. Melero, *GSM, GPRS and EDGE Performance: Evolution Towards 3G/UMTS*, John Wiley & Sons, 2003.

[9] Jordi Pérez-Romero, Oriol Sallent, Ramon Agustí, and Miguel Ángel Díaz-Guerra, *Radio Resource Management Strategies in UMTS*, John Wiley & Sons, 2005.

[10] Christian Bettstetter, Hans-Jorg Vogel, and Jorg Eberspacher, "GSM phase 2+ general packet radio service GPRS: Architecture, protocols, and air interface," *IEEE Communications Surveys & Tutorials*, vol. 2, no. 3, pp. 2–14, 1999.

[11] M. Ermel, K. Begain, T. Müller, J. Schüler, and M. Schweigel, "Analytical comparison of different GPRS introduction strategies," in *MSWIM '00: Proceedings of the 3rd ACM international workshop on Modeling, analysis and simulation of wireless and mobile systems*, New York, NY, USA, 2000, pp. 3–10, ACM.

[12] X. Gelabert, J. Perez-Romero, O. Sallent, and R. Agusti, "On the suitability of load balancing principles in heterogeneous wireless access networks," in *Int. Symposium on Wireless Personal Multimedia Communications (WPMC'05)*, Aalborg, Denmark, September 18-22., 2005.

[13] William J. Stewart, *Introduction to the numerical solution of Markov chains*, Princeton University Press, 1994.

[14] ETSI, "Selection procedures for the choice of radio transmission technologies of the UMTS," TR 101 112 V3.2.0, European Telecommunications Standards Institute (ETSI), 1998-04.

Chapter 3

Radio Access Selection Strategies in Constrained Multi-Access/Multi-Service Wireless Networks

The benefits of jointly managing the combined set of radio resources offered by heterogeneous networks consisting of several Radio Access Technologies (RATs) have been profusely studied and assessed in recent years. This notion has been coined as Common Radio Resource Management (CRRM)[1]. Nevertheless, most of the existing work assumes scenarios where all RATs are accessible (provided the RAT is not at full capacity) to those users demanding a particular service. If this is so, the obtained benefits become rather optimistic given that we neglect the fact that the deployed RATs may have different coverage overlapping conditions among them and that users may not have terminals that support all RATs (i.e. multimode terminals). In this chapter we extend the Markov framework developed in Chapter 2 in order to capture the effect of having different coverage overlapping conditions along with the capability of certain terminals to support all or a subset of available RATs. Probabilistic models are obtained for the characterization of a wide range of coverage and terminal heterogeneity scenarios. Extensive performance evaluation is carried out in order to identify those parameters influencing the suitability of a particular initial RAT selection strategy, that is, to choose the most suitable RAT at call/session initiation among those accessible. Results indicate that suitable RAT selection is tightly dependent on: the ratio between the different offered traffic loads (or service-mix), the contention (if any) for radio resources by different services in particular RATs, and the identification of access-constrained RATs due

[1] Sometimes being referred to as Joint RRM (JRRM) or Multiple RRM (MRRM) in the literature.

to lack of coverage and/or terminal capability. In this chapter we provide specific guidelines and rules concerning resource allocation for the utmost utilization of radio resources, in terms of Erlang capacity, and enhanced perceived quality by the users, in terms of achievable throughput.

3.1 Constrained RAT Selection: Motivation and Problem Statement

This chapter addresses the problem of initial RAT selection for the case of having several RATs with different coverage overlapping conditions over a given area. In addition, the capability of certain terminals to support all or a subset of available RATs is also captured. Moreover, the case where specific services are not supported by particular RATs is also considered.

Initial RAT selection is the CRRM functionality that has attracted a major attention from the research community in recent years (see e.g. survey [1] and references therein). Despite this effort, and to the author's best knowledge, the majority of existing works assume that, upon a service request at session initiation, all available RATs are eligible candidates for the RAT selection procedure [2–9]. Hence, disregarding the fact that different coverage overlapping conditions may exist among the RATs forming the heterogeneous network. In general, the spatial deployment of Base Station (BS) sites and/or Access Points (APs) corresponding to different RATs usually result in inhomogeneous and different coverage areas with diverse overlapping coverage settings among the RATs. Even in the case of co-located deployments, i.e. with BSs and APs sharing the same spatial location, coverage ranges of different RATs usually differ from each other thus leading to coverage unavailability regions. The term *coverage* as used in this chapter will define the ability of a particular RAT to provide access to its resources which can indistinctly refer to the spatial coverage range and/or the BS/AP deployment ability to "cover" a set of users. Then, in the case of homogeneous spatial user distribution the probability that a given user is "covered" by a specific RAT is merely dependent on the coverage area sizes. Conversely, non-homogeneous spatial user distributions in the form of *hot-spots*, or high-density populated areas may cause particular coverage regions to be more populated depending on the specific BS/AP deployments.

On the other hand, it is commonly assumed in related literature, see e.g. [2–8], that all users are provided with multimode terminals, as opposed to singlemode terminals, thus being able to establish communication through any available RAT, provided suitable coverage conditions apply. Along with the deployment and roll-out of new RATs, terminal capabilities are accordingly increased providing enhanced connectivity, capabilities and compliancy with these new emerging RATs. In addition, users owning multimode terminals may have the opportunity to choose their preferred RAT for particular services and even prevent from using particular RATs. For example, terminals which allow energy-saving third party applications to be used such that voice calls are preferably directed to 2G instead of 3G according to the power consumption. In addition, in WLAN-capable devices, users may prefer this RAT as opposed to 3G since the cost-per-Mbyte is usually lower (or even free of cost), hence disabling 3G access for data services. Whether hardware-constrained or user-constrained, coexistence of limited-access terminals with multi-mode terminals entails new challenges so as to efficiently manage the coordinated set of resources offered by all RATs and, therefore, must not be disregarded.

Finally, the capability of existing RATs in providing specific services in a given area is also a major concern for the RAT selection procedure. In this case, we deal with the problem that some RATs may not be able to handle particular service types, such as the case of video-streaming in a GSM network. Hence, the compliant service types handled by each RAT in the heterogeneous network may influence the suitability of a particular RAT selection procedure according to the offered traffic load of each service as we will see later on. In addition, cases where RATs handle one single service type whereas other RATs share their resources among various services will be also addressed in this chapter.

Bearing in mind the above, ubiquitous coverage and multimode terminal availability along with full service-RAT compatibility may cause optimistic considerations on the obtained gains through CRRM given they disregard limitations in access. These limitations will pose constraints to the problem of RAT selection in heterogeneous networks and will lead to a degraded network operation.

3.1.1 Related Work and Main Contributions of the Chapter

The interest of the Third Generation Partnership Project (3GPP) standardization body in heterogeneous networks was initiated by the identification of some CRRM architectures and procedures for integrated GSM/EDGE and UMTS networks in [10, 11]. Few works address the problem of terminal heterogeneity along with radio coverage conditions, and even fewer in an analytical form. Among the related work, in [9], Lincke identifies the degradation introduced by the limited operation of single-mode terminals on several traffic overflowing rules among RATs. In particular, this Chapter considers three services, namely 2G-speech, 3G-speech and

3G-data and a single allocation principle: direct 3G-speech to 2G first in order to free resources in 3G for 3G-data services. If 2G is at full capacity, overflowing of 3G-speech users with dual-mode terminal capabilities towards 3G is carried out so as to free resources for 2G capable single-mode terminals. Several overflow-return principles are examined. Similarly, work by Falowo in [12] also addresses the identical problem using the same Erlangian approach. In this case, as in [9], offered traffic loads into each RAT is known a-priori, therefore disregarding the impact of RAT selection policies and allocation principles. In [13], multi-mode terminal heterogeneity in heterogeneous networks is evaluated (under simulation analysis) proposing a traffic-aware adaptive resource reservation scheme based to account for both single-mode and dual-mode terminal users. As in [9, 12], the study is limited to a particular scenario and no generalization is provided. Moreover, [9, 12, 13] assume perfect coverage of all RATs over the area of interest.

This chapter departs from a developed Markov model in Chapter 2 in order to evaluate the resource allocation of different services into a number of RATs. The *main contributions* of this chapter, which differentiates it from previous studies [9, 12, 13], are listed hereafter:

- Extend the work in Chapter 2 to generalize the number of supported services and RATs.
- Service-RAT compatibility characterizing RATs that do not uphold particular services.
- Definition of a probabilistic coverage model enabling the characterization of any overlapping coverage situation among the considered RATs.
- Definition of a probabilistic model capturing the availability of multi-mode terminals which support all or a subset of available RATs.
- Describe a number of heuristic RAT selection policies which fall into two categories: service-based selection and load balancing selection. Hence, we account for several allocation principles as opposed to only one in [9, 12].
- Identify the main parameters that influence the performance of the proposed RAT selection policies, mainly: the ratio between offered voice and data traffic (or service-mix), the constrained access for each service request, and the effect of resource-contention between different services.
- Provide guiding principles and rules for the suitable allocation of voice and data services in the considered multi-access network. In this sense, we aim at maximizing Erlang capacity and throughput.

Summarizing, *this chapter addresses the effect of access impairments in the form of coverage, multi-mode terminal availability and RAT-service compliance in a heterogeneous network scenario along with an efficient implementation of RAT selection procedures taking into account the aforementioned limitations.*

The rest of the chapter is outlined as follows. In Section 3.2, the Markov framework modeling the allocation of multiple services on multiple RATs is defined. Subsequently, Section 3.3 addresses the probabilistic coverage and terminal heterogeneity models used in this chapter. Section 3.4 deals with the definition of RAT selection policies within the Markov model subject to coverage and/or terminal constraints. The model is particularized for the allocation of voice and data services in a GSM/UMTS/WLAN heterogeneous network in Section 3.5. Next, in Section 3.6, scenario and input parameters are defined along with the definition of relevant performance metrics. Numerical evaluation of the proposed framework is carried out in Section 3.7. Finally, Section 3.8 concludes the chapter with some final remarks and future paths.

3.2 A General Framework for Multi-Service Allocation in Multi-Access Networks

In the following subsections, the definition of a Markov framework that models the allocation of multiple services in multi-access networks is presented. The allocation model departs from the one in Chapter 2 where, in addition, we have included a higher degree of generality (in terms of number of supported RATs and services) and also service-RAT compatibility issues, thus offering a wider range of applicability. Essentially, any Markov model formulation problem involves the identification of the state space followed by the definition of the state transition rates, which specify the allocation of each service into each RAT, and, finally, the steady state balance equations. Before proceeding with the aforesaid steps, some preliminary definitions and notations are required along with the considered model assumptions.

3.2.1 Notation, preliminary definitions and model assumptions

Consider a heterogeneous network consisting of K RATs given by the set $\mathcal{K} = \{r_1, r_2, ..., r_k, ..., r_K\}$. A number of J different service classes $\mathcal{J} = \{s_1, s_2, ..., s_j, ..., s_J\}$

are available to users[2]. Each RAT supports either all or a subset of the J traffic classes. So as to account for RATs that do not uphold particular traffic classes, a service-RAT compatibility matrix, $\mathbf{B} = [b_{kj}] \in \mathbb{Z}^{K \times J}$, can be defined with elements $b_{kj} = 1$ if RAT $k \in \{1, 2, \ldots, K\}$ supports traffic type $j \in \{1, 2, \ldots, J\}$ and $b_{kj} = 0$ otherwise. According to $\mathbf{B}$, the number of supported services by a given RAT k, J_k, is $J_k = \sum_{j=1}^{J} b_{kj}$. Moreover, the number of RATs that support a specific service j, K_j, is given by $K_j = \sum_{k=1}^{K} b_{kj}$. Then, the Markov state dimension that accounts for the allocation of each supported service into each RAT is given by $M = \sum_{k=1}^{K} J_k = \sum_{j=1}^{J} K_j$. For convenience, the set of supported RAT-service indices $\mathcal{I} = \{(k, j) : b_{k,j} = 1\}$ is defined, where the cardinality of $\mathcal{I}$ is $|\mathcal{I}| = M$.

We may now define the row vector

$$\mathbf{N}_k = [N_{k,j}]_{(k,j) \in \mathcal{I}} \in \mathbb{Z}^{1 \times J_k}, \tag{3.1}$$

indicating the number of allocated service type j users in RAT k. Notice that we consider only those supported services in RAT k by imposing $(k, j) \in \mathcal{I}$.

Taking into account the number of available RATs, the number of allocated users of each supported service in each RAT may be written as a row vector

$$\mathbf{N} = [\mathbf{N}_1, \mathbf{N}_2, \ldots, \mathbf{N}_k, \ldots, \mathbf{N}_K] \in \mathbb{Z}^{M}, \tag{3.2}$$

with elements $\mathbf{N}_k$ given in (3.1), and where $\mathbf{N}$ will be the M-dimensional index that uniquely identifies each state, hereon denoted as $S_{\mathbf{N}}$, in the Markov chain model. Note that for the specific case of two RATs and two services with full support depicted in Chapter 2, the state index was simply (i, j, k, l).

As for traffic models, similar to Chapter 2, there is the need to define the characteristics of both arrival and departure processes. For the call/session arrival process of a particular service type j, it can be safely assumed, and widely used in the literature, that this process follows a Poisson distribution. As for the departure process of a particular service type j, which relates to the call/session holding time, some studies show that exponential call holding time distribution assumption is reasonable [14, 16], while others advocate for gamma and lognormal distributions [17]. Since the focus of this chapter is the evaluation of RAT selection schemes and the analytical model of coverage and terminal limitations, rather than the traffic model itself, the exponential distribution assumption is considered since it makes the Markov model more tractable. Identical approach has been extensively used by [18–21]. The reader is also referred to Chapter 2 (in particular to Figs. 2.3, 2.4

[2] In the following, and unless otherwise stated, numerical indices k and j will refer to particular RAT and service respectively.

and 2.5 in Section 2.6.1) where the Markov model was compared against a system level simulator considering both exponential and lognormal distributions revealing that, for the purpose of RAT selection evaluation and comparison, the exponential assumption suitably fulfills our needs.

We are interested in determining the impact of a particular initial RAT selection scheme when non-ideal coverage and terminal availability conditions apply along with RAT-service compliance issues. Then, as motivated in Section 1.3, it will be assumed that a user initially assigned to a particular RAT will remain in the same RAT throughout the duration of the call/session. Hence, no inter-RAT handover (i.e. VHO) will be considered in our model. Although considering VHO capabilities would be certainly interesting in order to capture the whole dynamics of the system it would however obscure the purpose of our study by hiding the effect of initial RAT selection. This interest is motivated by the importance of efficiently designing the initial RAT selection procedure as a first step in the realization of the ABC concept [22]. In addition, perfect link conditions will be assumed, thus no disruptions due to link quality failures will be considered.

3.2.2 Defining the Markov state space

Assuming that the capacity of a particular RAT k, defined as the maximum allowable number of users of each service type it may handle, is upper-bounded; a finite number of states $S_{\mathbf{N}}$ arises. The limit on the number of states is set by RAT-specific Call Admission Control (CAC) procedures that determine the maximum number of users this RAT may admit in order to guarantee some minimum QoS requirements. In terms of the number of states, we define the set of feasible states in RAT k, $\mathcal{S}^k$, as

$$\mathcal{S}^k = \{S_{\mathbf{N}} : 0 \leq f^k_{\mathbf{N}_k} \leq 1\}, \tag{3.3}$$

where $f^k_{\mathbf{N}_k}$ is the *feasibility condition* which accounts for the CAC procedures in RAT k by determining if a given state $S_{\mathbf{N}}$ is feasible in RAT k provided $f^k_{\mathbf{N}_k}$ lays between 0 and 1. In addition, a state $S_{\mathbf{N}}$ is said to be feasible if it satisfies $S_{\mathbf{N}} \in \mathcal{S}$ with $\mathcal{S} = \bigcap_{k=1}^{K} \mathcal{S}^k$, i.e. if it is feasible in all RATs. Note that (3.3) is the generalization of (2.2) in Chapter 2.

3.2.3 State transition rates and RAT selection policies

Transitions between states $S_{\mathbf{N}} \in \mathcal{S}$ in the resulting M-dimensional Markov chain happen due to service arrival rates, i.e. $\boldsymbol{\lambda} = [\lambda_1, \lambda_2, \ldots, \lambda_j, \ldots \lambda_J]$, or due to service departure rates $\boldsymbol{\mu} = [\mu_1, \mu_2, \dots, \mu_j, \dots, \mu_J]$, where λ_j indicates the arrival rate of the *j-th* service type and μ_j indicates the departure rate of the *j-th* service. As mentioned in section 3.2.1, it is assumed that rates λ_j and μ_j are Poisson and exponentially distributed respectively. Moreover, service rates μ_j will not depend on the allocated bit-rate in a given specific RAT, thus assuming that a user type will remain, on average, $1/\mu_j$ seconds in the selected RAT. Provided not all services may be supported by all RATs, recall from section 3.2.1, the supported arrival rates into RAT k, $\boldsymbol{\lambda}_k$, can be defined as the row vector

$$\boldsymbol{\lambda}_k = [\lambda_{k,j}]_{(k,j)\in\mathcal{I}} \in \mathbb{R}^{1\times J_k}, \tag{3.4}$$

with $\lambda_{k,j}$ the arrival rate of the *j-th* supported service type in RAT k. Note that $\boldsymbol{\lambda}_k$ is a subset of elements in $\boldsymbol{\lambda}$ determined by compatibility matrix $\mathbf{B}$.

As stated in Section 3.1, the initial RAT selection procedure is responsible of allocating users, upon call/session establishment, to the most appropriate RAT according to some predetermined criteria. In the Markov model, this notion translates into the definition of specific traffic allocation policies which determine the fraction of traffic arrival rates into each available RAT. This means that a particular traffic allocation policy, denoted generically as π, will govern the transition arrival rates into each state in the state space. Then, for each state $S_{\mathbf{N}} \in \mathcal{S}$, a traffic allocation policy π determines the specific transition arrival rates $\lambda^{\pi}_{(\mathbf{N},k,j)}$, indicating the transition arrival rate of service j into RAT k, with $(k,j) \in \mathcal{I}$.

Specifically, a particular policy π may be implemented by means of *policy actions* $\Theta^{\pi}_{(\mathbf{N},k,j)} \in [0,1]$ determining the fraction of supported traffic j into RAT k in state $S_{\mathbf{N}}$:

$$\lambda^{\pi}_{(\mathbf{N},k,j)} = \Theta^{\pi}_{(\mathbf{N},k,j)} \lambda_{k,j}. \tag{3.5}$$

Note that (3.5) evokes the splitting property of the Poisson process $\lambda_{k,j}$, which, by considering $\Theta^{\pi}_{(\mathbf{N},k,j)} \in [0,1]$, ensures that $\lambda^{\pi}_{(\mathbf{N},k,j)}$ also follows a Poisson distribution (see e.g. [23]).

3.2.4 Steady State Balance Equations (SSBEs)

In equilibrium, the SSBE in $S_{\mathbf{N}} \in \mathcal{S}$ results from equaling the inflow rate to the outflow rate [24], i.e.

$$\begin{gathered} P_{\mathbf{N}}\Big[\sum_{(k,j)\in\mathcal{I}} \lambda_{k,j}\Theta^{\pi}_{(\mathbf{N},k,j)}\delta_{(\mathbf{N}+\mathbf{a}_{k,j})} + N_{k,j}\mu_j\delta_{(\mathbf{N}-\mathbf{a}_{k,j})}\Big] = \\ \sum_{(k,j)\in\mathcal{I}} \lambda_{k,j}\Theta^{\pi}_{(\mathbf{N}-\mathbf{a}_{k,j},k,j)}P_{(\mathbf{N}-\mathbf{a}_{k,j})}\delta_{(\mathbf{N}-\mathbf{a}_{k,j})} + (N_{k,j}+1)\mu_j P_{(\mathbf{N}+\mathbf{a}_{k,j})}\delta_{(\mathbf{N}+\mathbf{a}_{k,j})} \end{gathered}, \tag{3.6}$$

where $P_{\mathbf{N}}$ is the state $S_{\mathbf{N}}$ probability, and $\mathbf{a}_{k,j} \in \mathbb{Z}_{+}^{M}$ is a row vector containing all zeroes except for the *j-th* supported service in RAT k element which equals to 1. In addition, $\delta_{\mathbf{N}}$ is an indicator function ensuring that non-feasible states are not considered, i.e. $\delta_{\mathbf{N}} = 1$ if $S_{\mathbf{N}} \in \mathcal{S}$ and $\delta_{\mathbf{N}} = 0$ otherwise.

Once the SSBEs are determined for all states $S_{\mathbf{N}} \in \mathcal{S}$, the resulting system of equations given by the SSBEs plus the normalization constraint $\sum_{S_{\mathbf{N}} \in S} P_{\mathbf{N}} = 1$ should be solved in order to obtain the steady state probabilities $P_{\mathbf{N}}$. In this chapter, the steady state probabilities ($P_{\mathbf{N}}$) are obtained using the iterative *power method* (refer to [23] for further details).

3.3 Coverage and Terminal Models

This section presents the corresponding probabilistic models accounting for coverage and terminal heterogeneity scenarios. Later, both models are merged into the RAT-service eligibility concept which captures joint effects and constitutes a key element in the RAT selection procedure defined in Section 3.4 consequently affecting the resource allocation model given in Section 3.2. Note that the RAT-service compatibility issue has already been captured by compatibility matrix $\mathbf{B}$ defined in Section 3.2.1. We therefore implicitly consider in the remainder that the RAT-service pair (k, j) is compatible, i.e. $(k, j) \in \mathcal{I}$.

3.3.1 Probabilistic multimode terminal availability model

Consider the total set of existing terminal types associated to a particular service j, denoted as $\mathcal{T}$, can be partitioned into $N = 2^K$ mutually exclusive terminal types, denoted by $\mathcal{T}_{i,j}$ with $i \in \{0, 1, ..., N-1\}$. Each terminal type $\mathcal{T}_{i,j}$ supports a subset of K available RATs. For convenience, we define a $2^K \times K$ matrix $\mathbf{T}$ containing all possible terminal support combinations where elements $t_{(i,k)} = 1$ in $\mathbf{T}$ indicate that terminal type $\mathcal{T}_{i,j}$ supports RAT k and $t_{(i,k)} = 0$ otherwise. Each terminal type $\mathcal{T}_{i,j}$ is associated with a known probability $\Upsilon_{i,j}$=Pr$\{\mathcal{T}_{i,j}\}$ which can be obtained, e.g., through detailed market studies on the availability of each of the terminal types in a specified location. In addition, Table 3.1 shows the multimode terminal availability scenario formulation for the case of $K = 3$ RATs (i.e r_1, r_2 and r_3) and where the notation $\Upsilon_{i,j} = P_{t,j}\{\bigcap_{k=0}^{K} x_k\}$, with $x_k \in \{r_k, \bar{r}_k\}$, indicates the probability that RATs r_k are supported and RATs $\bar{r}_k$ are not supported by the terminal with

Table 3.1: Multimode terminal availability scenario example with $K = 3$ RATs, thus $N = 2^K = 8$.

Type	Matrix $\mathbf{T}$ $(t_{(i,k)})$	Description	Probability
$(\mathcal{T}_{i,j})$	r_1 r_2 r_3		$\Upsilon_{i,j} = \Pr\{\mathcal{T}_i\}$
$\mathcal{T}_{0,j}$	0 0 0	No RAT supported	$\Upsilon_{0,j} = P_{t,j}\{\bar{r}_1 \cap \bar{r}_2 \cap \bar{r}_3\}$
$\mathcal{T}_{1,j}$	1 0 0	r_1 terminal only	$\Upsilon_{1,j} = P_{t,j}\{r_1 \cap \bar{r}_2 \cap \bar{r}_3\}$
$\mathcal{T}_{2,j}$	0 1 0	r_2 terminal only	$\Upsilon_{2,j} = P_{t,j}\{\bar{r}_1 \cap r_2 \cap \bar{r}_3\}$
$\mathcal{T}_{3,j}$	1 1 0	r_1, r_2 terminal only	$\Upsilon_{3,j} = P_{t,j}\{r_1 \cap r_2 \cap \bar{r}_3\}$
$\mathcal{T}_{4,j}$	0 0 1	r_3 terminal only	$\Upsilon_{4,j} = P_{t,j}\{\bar{r}_1 \cap \bar{r}_2 \cap r_3\}$
$\mathcal{T}_{5,j}$	1 0 1	r_1, r_3 terminal only	$\Upsilon_{5,j} = P_{t,j}\{r_1 \cap \bar{r}_2 \cap r_3\}$
$\mathcal{T}_{6,j}$	0 1 1	r_2, r_3 terminal only	$\Upsilon_{6,j} = P_{t,j}\{\bar{r}_1 \cap r_2 \cap r_3\}$
$\mathcal{T}_{7,j}$	1 1 1	r_1, r_2, r_3 terminal	$\Upsilon_{7,j} = P_{t,j}\{r_1 \cap r_2 \cap r_3\}$

requesting service j. For example, $\Upsilon_{3,j} = P_{t,j}\{r_1 \cap r_2 \cap \bar{r}_3\}$ is the probability that user u_j's terminal supports RATs r_1 and r_2 but not r_3.

In addition, the probability associated to particular terminal characteristics can be computed regarding the probabilities $\Upsilon_{i,j}$, such as, e.g., the probability that a user (u_j) terminal supports RAT r_k, $P_{t,j}\{r_k\}$, or the probability that a particular terminal supports RAT r_k but not RAT $r_{k'}$, $P_{t,j}\{r_k \cap \bar{r}_{k'}\}$, etc.

It is worthwhile noting that the probability of a particular user's terminal of supporting a specific RAT r_k, $P_{t,j}\{r_k\}$, depends on the demanding service type j. Indeed, a service type user u_j must have a terminal that supports, at least, one RAT capable of offering such service j (e.g. a user with a GSM-only terminal will not attempt to establish a video-call).

3.3.2 Probabilistic coverage availability model

For the coverage model, a similar approach to the multimode availability case is adopted. Assume we can partition the whole area of interest $\mathcal{A}$ into $N = 2^K$ mutually exclusive areas denoted by $\mathcal{A}_i$ with $i \in \{0, 1, ..., N-1\}$ (i.e. a complete space tessellation). Similar to the terminal availability case in Section 3.3.1, we define a $2^K \times K$ matrix $\mathbf{A}$ containing all possible coverage overlapping combinations where elements $a_{(i,k)} = 1$ in $\mathbf{A}$ indicate that a user in area $\mathcal{A}_i$ is covered by RAT k and $a_{(i,k)} = 0$ otherwise. Each area $\mathcal{A}_i$ is associated to a known probability $\Psi_i = \Pr\{\mathcal{A}_i\}$, which captures not only the probability that a given user is in a particular area (spatial location) but also the ability of a given RAT to cover a user (i.e. efficient coverage planning) and that such user is covered during the

Table 3.2: Coverage availability scenario example with $K = 3$ RATs ($N = 2^K = 8$).

Coverage event ($\mathcal{A}_i$)	Matrix $\mathbf{A}$ ($a_{(i,k)}$) r_1 r_2 r_3	Description	Probability $\Psi_i = \Pr\{\mathcal{A}_i\}$
$\mathcal{A}_0$	0 0 0	No coverage at all	$\Psi_0 = P_c\{\bar{r}_1 \cap \bar{r}_2 \cap \bar{r}_3\}$
$\mathcal{A}_1$	1 0 0	r_1 coverage only	$\Psi_1 = P_c\{r_1 \cap \bar{r}_2 \cap \bar{r}_3\}$
$\mathcal{A}_2$	0 1 0	r_2 coverage only	$\Psi_2 = P_c\{\bar{r}_1 \cap r_2 \cap \bar{r}_3\}$
$\mathcal{A}_3$	1 1 0	r_1, r_2 coverage only	$\Psi_3 = P_c\{r_1 \cap r_2 \cap \bar{r}_3\}$
$\mathcal{A}_4$	0 0 1	r_3 coverage only	$\Psi_4 = P_c\{\bar{r}_1 \cap \bar{r}_2 \cap r_3\}$
$\mathcal{A}_5$	1 0 1	r_1, r_3 coverage only	$\Psi_5 = P_c\{r_1 \cap \bar{r}_2 \cap r_3\}$
$\mathcal{A}_6$	0 1 1	r_2, r_3 coverage only	$\Psi_6 = P_c\{\bar{r}_1 \cap r_2 \cap r_3\}$
$\mathcal{A}_7$	1 1 1	r_1, r_2, r_3 coverage	$\Psi_7 = P_c\{r_1 \cap r_2 \cap r_3\}$

call/session lifetime. As in the previous section, this probability Ψ_i constitutes an input to our model and could be retrieved in practice through detailed coverage and user spatial distribution studies. Moreover, Table 3.2 shows the coverage scenario formulation for $K = 3$ RATs (r_1, r_2 and r_3) and where the notation $\Psi_i = P_c\{\bigcap_{i=0}^{N-1} y\}$, with $y \in \{r_k, \bar{r}_k\}$, indicates the probability that RATs r_k are covered and RATs $\bar{r}_k$ are not covered. For example, $\Psi_5 = P_c\{r_1 \cap \bar{r}_2 \cap r_3\}$ is the probability that a user is covered by RATs r_1 and r_3 but not by r_2.

Accordingly, the probability associated to specific coverage situations can be computed considering those probabilities given by Ψ_i, such as, e.g., the probability that RAT r_k covers a user, $P_c\{r_k\}$, or the probability that a particular user is covered by RAT r_k but not by RAT $r_{k'}$, $P_c\{r_k \cap \bar{r}_{k'}\}$, etc.

Note that it has been considered that coverage conditions are independent of the requesting service-type, thus the coverage probability of a specific RAT, $P_c\{r_k\}$, does not depend on j. However, considering different spatial distributions for each service could be easily adopted in the model and would follow a similar approach as for the terminal probability case described in Section 3.3.1.

3.3.3 RAT-service eligibility

For a particular RAT r_k to be eligible by a given user u_j (i.e. requesting service j) during a RAT selection process, coverage conditions and terminal characteristics must allow such selection. Then, the probability that a specific RAT r_k is eligible for a given user u_j is the probability, $P_{e,j}\{r_k\}$ (henceforth the *eligibility probability*), that: 1) RAT r_k supports service j, i.e. $(k, j) \in \mathcal{I}$; 2) user u_j has terminal capabilities to uphold RAT r_k and; 3) RAT r_k provides coverage to user u_j during its call/session, i.e.

$$P_{e,j}\{r_k\} = P_{t,j}\{r_k\} \cdot P_c\{r_k\}, \tag{3.7}$$

where $P_{t,j}\{r_k\}$ is the probability that user u_j's terminal supports RAT r_k (which implicitly means that RAT r_k supports service j, see Section 3.3.1); and $P_c\{r_k\}$ the probability that this user is covered by RAT r_k (see Section 3.3.2). Note in (3.7) that coverage and terminal processes follow independent distributions.

Similarly we may be interested, as reflected in Section 3.4, in computing the joint eligibility probability of events such as $P_{e,j}\{r_m \cap \bar{r}_n\}$, i.e. the probability that RAT r_m is eligible but not r_n. For example, to compute the probability $P_{e,j}\{r_1 \cap \bar{r}_2\}$, we explore all possible combinations of coverage and terminal events causing RAT r_1 to be eligible but not RAT r_2, yielding

$$P_{e,j}\{r_1 \cap \bar{r}_2\} = (\Upsilon_{1,j} + \Upsilon_{3,j} + \Upsilon_{5,j} + \Upsilon_{7,j}) \cdot (\Psi_1 + \Psi_5) + (\Upsilon_{1,j} + \Upsilon_{5,j}) \cdot (\Psi_3 + \Psi_7).$$

Summarizing, (3.7) allows capturing in a single parameter, the eligibility probability $P_{e,j}\{r_k\}$, the effect of terminal availability, coverage-related issues and the capability of certain RATs to support specific services, which allows us to easily represent and evaluate a wide variety of scenarios.

3.4 RAT Selection under Coverage and Multimode Terminal Availability Constraints

Considering the Markov framework in Section 3.2, and for the broad case of having K RATs, $\mathcal{K} = \{r_1, r_2, ..., r_k, ..., r_K\}$, it is assumed that the RAT selection procedure over a given user, u_j, demanding service type j in state $S_{\mathbf{N}} \in \mathcal{S}$ prioritizes the candidate RATs according to the ordered set

$$\mathcal{R}^{\pi}_{\mathbf{N},j} = \{r^*_1, r^*_2, \ldots r^*_i \ldots r^*_{K_j}\}, \tag{3.8}$$

where r^*_1 is the RAT with highest priority and $r^*_{K_j}$ is the RAT with lowest priority, with $r^*_i \in \mathcal{K}$ for $i = \{1, 2, 3 \ldots K_j\}$. Note that the set of candidate RATs given by (3.8) are those RATs that support service j, thus containing K_j elements, with K_j the number of RATs that support a specific service j (see Section 3.2.1). In the eventual, but not unusual, case of two or more RATs with the same priority, random selection will be performed among those RATs so as to enforce strict priority order in set $\mathcal{R}^{\pi}_{\mathbf{N},j}$.

Assuming that a RAT selection policy π in state $S_{\mathbf{N}} \in \mathcal{S}$ prioritizes the RATs according to the set $\mathcal{R}^{\pi}_{\mathbf{N},j}$ in (3.8), the goal is to determine the fraction of supported traffic j into each RAT k, $\lambda^{\pi}_{(\mathbf{N},k,j)}$, as defined in (3.5). These expressions will be affected by the eligibility of particular RATs according to coverage and terminal availability models given in Section 3.3. For representation purposes, let the Boolean $\mathcal{C}_{j,k}$ indicate that free resources for service j in RAT k are available, and $\bar{\mathcal{C}}_{j,k}$ otherwise. Moreover, consider the function $k = h_i$ which returns the index k of the *i-th* prioritized RAT (r_i^*) in set $\mathcal{R}^{\pi}_{\mathbf{N},j}$.

Then, for the case of the RAT with highest priority, $k = h_1$, we have expression

$$\lambda^{\pi}_{(\mathbf{N},h_1,j)} = \Theta^{\pi}_{(\mathbf{N},h_1,j)} \cdot \lambda_{h_1,j} = \begin{cases} \lambda_{h_1,j} \cdot P_{e,j}\{r_{h_1}\}, & \text{if } \mathcal{C}_{j,h_1} \\ 0, & \text{otherwise} \end{cases}, \tag{3.9}$$

where it states that service j will be allocated to r_1^* (i.e the preferred option) provided r_1^* has enough capacity to admit this user and that r_1^* is *eligible* (i.e. terminal supports RAT r_1^* and it is covered by RAT r_1^*. Otherwise, the arrival rate of service j into RAT r_1^* is zero.

For the case of the second preferred RAT, r_2^*, the fraction of traffic j into $k = h_2$ yields

$$\lambda^{\pi}_{(\mathbf{N},h_2,j)} = \Theta^{\pi}_{(\mathbf{N},h_2,j)} \cdot \lambda_{h_2,j} = \begin{cases} \lambda_{h_2,j} \cdot P_{e,j}\{\bar{r}_{h_1} \cap r_{h_1}\}, & \text{if } \mathcal{C}_{j,h_1} \wedge \mathcal{C}_{j,h_2} \\ \lambda_{h_2,j} \cdot P_{e,j}\{r_{h_2}\}, & \text{if } \bar{\mathcal{C}}_{j,h_1} \wedge \mathcal{C}_{j,h_2} \\ 0, & \text{otherwise} \end{cases}. \tag{3.10}$$

The first case in (3.10) states that, provided enough capacity is left in r_1^* and r_2^* for service j, traffic will be allocated to r_2^* (i.e. the second preferred option) if RAT r_1^* is not eligible (i.e. preferred option not eligible) whereas r_2^* is. The second case in (3.10) states that, if no capacity is left for allocating service j in r_1^*, this service will be allocated in r_2^* provided it has enough capacity and is eligible.

A similar reasoning to that of (3.9,3.10) would follow for the definition of $\lambda^{\pi}_{(\mathbf{N},h_i,j)}$ with $i = \{3, 4, ..., K_j\}$.

3.5 Case Study: Voice and Data Services in a GSM/UMTS/WLAN Deployment Scenario

In order to assess and evaluate the proposed model we consider a practical case study in which J=2 services, generic voice and data, request to be allocated in a heterogeneous network scenario comprising $K{=}3$ deployed RATs, namely: GSM, UMTS and WLAN. Although a higher number of services and RATs could be considered and supported by the proposed model, in practice, this would difficult the interpretation of the results while not adding substantial value to the model evaluation itself. The adopted RATs for this case study have been chosen based on their widespread use and deployment ubiquity; however, other RAT representatives or enhancements to those considered here (e.g. GPRS/EDGE, WIMAX, HSPA, LTE, etc.) could be easily adopted instead. For example, to account for HSDPA, expressions in [25] could be adopted into our state feasibility model. Similarly, [26] presents analogue expressions for OFDMA schemes such as those employed by WIMAX and LTE. In addition, GPRS/EDGE characterization into a similar model was already captured by the authors in [14].

Coverage and multi-mode terminal availability will be modeled according to what described in Sections 3.3.1 and 3.3.2. In the following subsections, the resulting state space that arises from considering the abovementioned RATs are defined along with the formulation of the RAT selection procedures which affect the state transition rates as specified in Section 3.4. In the analysis below, a simplified terminology is used to ease tractability and understanding: let v and d represent voice and data indexes, along with g, u and w representing GSM, UMTS and WLAN indexes respectively.

3.5.1 State space Definition

It is considered that GSM has the capability of supporting only voice services, while UMTS may uphold both voice and data traffic; and finally, WLAN is capable of supporting only data traffic. In this specific case, the compatibility matrix $\mathbf{B}$ defined in Section 3.2.1 is given by:

$$\mathbf{B} = \begin{bmatrix} b_{gv} & b_{gd} \\ b_{uv} & b_{ud} \\ b_{wv} & b_{wd} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 0 & 1 \end{bmatrix}, \tag{3.11}$$

from which we deduce that the Markov state dimension in this particular case study is $M = 4$, and that the vector that uniquely defines a state can be expressed as

$$\mathbf{N} = [\mathbf{N}_g, \mathbf{N}_u, \mathbf{N}_w] = [N_{g,v}, N_{u,v}, N_{u,d}, N_{w,d}], \tag{3.12}$$

indicating the number of users of each supported service type into each RAT. Additionally, indicator vector $\mathbf{a}_{k,j} \in \mathbb{Z}^M$ generically defined in Section 3.2.4, yields in this case

$$\mathbf{a}_{g,v} = [1, 0, 0, 0], \mathbf{a}_{u,v} = [0, 1, 0, 0], \mathbf{a}_{u,d} = [0, 0, 1, 0], \mathbf{a}_{w,d} = [0, 0, 0, 1]. \tag{3.13}$$

Then, the state space results from applying specific CAC procedures in each RAT in order to determine the maximum number of allowable users of each type in each RAT, i.e. the user capacity, as described in the following subsections. Given the focus is on the evaluation of RAT selection policies in access-constrained scenarios, and not on providing accurate capacity models for GSM, UMTS and WLAN, simple state feasibility models are presented for the considered RATs. It is also assumed the uplink direction regarding user capacity, which was also considered in 2. Nonetheless, more elaborate and accurate models could be considered, if desired, and easily adopted into our framework. In this sense, Chapter 4 assumes both uplink and downlink directions in a similar framework.

It is assumed, as e.g. in [3, 5, 14], that the per-RAT feasibility condition, $f^k_{\mathbf{N}_k}$, follows a common structure given, in general terms, by

$$f^k_{\mathbf{N}_k} = \sum\nolimits_{(k,j)\in\mathcal{I}} N_{k,j} \cdot \Delta\gamma_{k,j}, \tag{3.14}$$

i.e the feasibility condition is a summative contribution of the number of admitted users of each supported service ($N_{k,j}$) times the radio resource consumption of such service ($\Delta\gamma_{k,j}$) in RAT k. Note also that the quantity $1/\Delta\gamma_{k,j}$ defines the maximum number of users of a particular service type j allowed in RAT k. The radio resource consumption (henceforth, RC), as shown later on, will be of particular interest when analyzing the performance of different RAT selection policies in several access-limited scenarios.

Next, feasibility conditions in GSM, UMTS and WLAN are derived in order to obtain, first, the set of feasible states in each RAT and, second, the total set of feasible states in the heterogeneous network.

3.5.1.1 GSM state feasibility

In Time Division Multiple Access (TDMA) systems, such as GSM, a total amount of C channels (or timeslots, TSLs) are shared among users within a time frame. A voice user is assumed to occupy a whole TSL throughout the call duration. Hence, the RC of a voice user in GSM is

$$\Delta\gamma_{g,v} = 1/C, \tag{3.15}$$

and the feasible state space in GSM yields

$$\mathcal{S}^g = \{S_{\mathbf{N}} :\ 0 \le N_{g,v} \cdot \Delta\gamma_{g,v} \le 1\}. \tag{3.16}$$

3.5.1.2 UMTS state feasibility

In systems based on Wideband Code Division Multiple Access (WCDMA), such as UMTS, the maximum number of users it may admit, given by vector $\mathbf{N_u}$, can be computed imposing that the uplink load factor, $\eta(\mathbf{N_u})$, should be kept below a given threshold η_{max} [27]:

$$\eta(\mathbf{N}_u) = (1+f)\cdot\Big(\sum_{j\in\{v,d\}} \alpha_j N_{u,j} \left[W/(R_j\theta_j)+1\right]^{-1}\Big) \leq \eta_{max}, \tag{3.17}$$

where α_v (α_d) is the voice (data) activity factor, R_v (R_d) is the allocated bit-rate for voice (data) users, θ_v (θ_d) is the target bit-energy-to-noise ratio for voice (data) users, W is the chip-rate, and f is the average uplink inter-cell-to-intra-cell interference. Moreover, in (3.17) the common assumption, see Chapter 2, that all users of a particular service demand the same bit-rate and target bit-energy-to-noise ratios has been considered. In addition, users transmit on a dedicated channel (DCH) with fixed allocated bit-rate. Note in (3.17) that, as opposed to what considered in Chapter 2, we take into account the inter-cell interference (f) and the activity factor (α_j).

From (3.17) we identify the RC in UMTS as

$$\Delta\gamma_{u,j} = (1+f)\cdot\alpha_j \left[\eta_{max}\left(W/(\theta_j\cdot R_j)\right)\right]^{-1}, \tag{3.18}$$

where index $j \in \{v, d\}$ accounts for voice and data services respectively, and the state space for UMTS can be written as

$$\mathcal{S}^u = \{S_{\mathbf{N}} :\ 0 \leq N_{u,v}\cdot\Delta\gamma_{u,v} + N_{u,d}\cdot\Delta\gamma_{u,d} \leq 1\}. \tag{3.19}$$

3.5.1.3 WLAN state feasibility

In order to account for the maximum allowable number of data users in WLAN, $N_{w,d}^{max}$, and hence for the set of feasible states, we impose some minimum throughput-per-user requirements to be met for admitted users. In this sense, we make use of some expressions provided in [28] [see equation (8) therein and related] to compute the throughput-per-user for a particular traffic class. From [28], the maximum number of simultaneously transmitting users, $n_{w,d}^{\max}$, can be computed such that a minimum throughput-per-data-user, denoted by $R_{w,d}^{\min}$, is met. In addition, assuming admitted data users exhibit an activity factor given by α_d, and are therefore not constantly transmitting, the maximum number of admitted users in WLAN, $N_{w,d}^{\max}$, can be computed as

$$N_{w,d}^{max} = \left\lfloor n_{w,d}^{max} / \alpha_d \right\rfloor, \tag{3.20}$$

where $\lfloor x \rfloor$ indicates the lowest closest integer to x.

Consequently, the RC of data service in WLAN is

$$\Delta\gamma_{w,d} = 1/N_{w,d}^{max} \tag{3.21}$$

and the feasible state space in WLAN is given by

$$\mathcal{S}^w = \{S_{\mathbf{N}} : \ 0 \leq N_{w,d} \cdot \Delta\gamma_{w,d} \leq 1\}. \tag{3.22}$$

3.5.1.4 Global state feasibility

Finally, once the feasibility conditions in each RAT are defined, the global Markov state feasibility space, $\mathcal{S}$, yields

$$\mathcal{S} = \{\mathbf{N} : \mathbf{N}_g \in \mathcal{S}^g, \mathbf{N}_u \in \mathcal{S}^u, \mathbf{N}_w \in \mathcal{S}^w\}, \tag{3.23}$$

whith sets $\mathcal{S}_g$, $\mathcal{S}_u$ and $\mathcal{S}_w$ given by (3.16), (3.19) and (3.22).

3.5.2 RAT Selection Formulation

RAT selection policies may consider a number of different criteria in order to determine the most appropriate RAT for a specific user at a given time: requesting service, path-loss conditions, resource-consumption, etc., see e.g. [3, 7–9]. Without loss of generality, we present here some representative policies in order to evaluate the RAT selection process when non-ideal coverage and multi-mode terminal availability conditions are assumed. These policies are mainly based and inspired on those previously introduced in Chapter 2. In the following, a brief description is provided.

3.5.2.1 Service-Based (SB) RAT Selection Policy

Based on the requesting service type, in our case study voice or data, several RAT selection policies can be defined. Following the notation given by (3.8), Table 3.3 shows the possible allocation principles for voice and data services in the GSM/UMTS/WLAN combined network. Since voice service is only supported by GSM and UMTS, and data service is supported solely by UMTS and WLAN, the priority set given by (3.8) for each service will have a cardinality of 2, i.e. $\mathcal{R}^{\pi}_{\mathbf{N},j} = \{r^*_1, r^*_2\}$. Hence, if the highest priority RAT (r^*_1) cannot allocate the requesting service due to non available capacity, the lowest priority RAT (r^*_2) is selected to attempt admission.

Table 3.3: Service-based RAT selection policies.

Policy	Acronym	RAT Priority		Description
		$\mathcal{R}^{\pi}_{\mathbf{N},v} = \{r^*_1, r^*_2\}$	$\mathcal{R}^{\pi}_{\mathbf{N},d} = \{r^*_1, r^*_2\}$	
Service-Based #1	SB#1	$\mathcal{R}^{SB\#1}_{\mathbf{N},v} = \{g, u\}$	$\mathcal{R}^{SB\#1}_{\mathbf{N},d} = \{w, u\}$	Voice priority to GSM, data priority to WLAN.
Service-Based #2	SB#2	$\mathcal{R}^{SB\#2}_{\mathbf{N},v} = \{u, g\}$	$\mathcal{R}^{SB\#2}_{\mathbf{N},d} = \{u, w\}$	Voice and data priority to UMTS.
Service-Based #3	SB#3	$\mathcal{R}^{SB\#3}_{\mathbf{N},v} = \{g, u\}$	$\mathcal{R}^{SB\#3}_{\mathbf{N},d} = \{u, w\}$	Voice priority to GSM, data priority to UMTS.
Service-Based #4	SB#4	$\mathcal{R}^{SB\#4}_{\mathbf{N},v} = \{u, g\}$	$\mathcal{R}^{SB\#4}_{\mathbf{N},d} = \{w, u\}$	Voice priority to UMTS, data priority to WLAN.

Noteworthy is that resource contention will solely happen in UMTS since this RAT handles both voice and data services, whereas GSM only handles voice users and WLAN only handles data users. From this perspective, SB#1 can be regarded as a *late-contention* policy since it first directs traffic to the RATs that uniquely handles that particular service, i.e. voice to GSM and data to WLAN. Conversely, SB#2 is regarded as an *early-contention* policy in which both services first compete for resources in UMTS. In the same way, SB#3 can be characterized as a *data-greedy* policy since it avoids contention by directing data users to UMTS and voice users to GSM. In this case, data users make the most of UMTS resources. Conversely, SB#4 can be regarded as a *voice-greedy* policy by actuating in the opposite way as SB#3.

3.5.2.2 Load-Based (LB) RAT Selection Policy

This policy selects the RAT according to the current load levels in each of the accessible RATs by selecting the RAT that undergoes a minimum load status. Accordingly, we must define load metrics for each RAT at a given state $S_{\mathbf{N}}$, referred to as $L_{\mathbf{N}}^{k}$. Noting that the feasibility condition in (3.14) represents the fraction of consumed resources in a particular RAT, we may simply define the load as

$$L_{\mathbf{N}}^{k} = f_{\mathbf{N}_k}^{k} = \sum\nolimits_{(k,j)\in\mathcal{I}} N_{k,j} \cdot \Delta\gamma_{k,j}, \tag{3.24}$$

where the RC of each RAT ($\Delta\gamma_{k,j}$) should be conveniently particularized for GSM, UMTS and WLAN by respectively considering (3.15), (3.18) and (3.21).

Once the load in each RAT has been defined, the LB RAT selection policy (LB for short) can be defined in terms of the prioritized set given by (3.8) as:

$$\mathcal{R}_{\mathbf{N},v}^{LB} = \left\{ \underset{k\in\{g,u\}}{\arg\min} \left\{ L_{\mathbf{N}+\mathbf{a}_{k,v}}^{k} \right\}, \underset{k\in\{g,u\}}{\arg\max} \left\{ L_{\mathbf{N}+\mathbf{a}_{k,v}}^{k} \right\} \right\}, \tag{3.25}$$

$$\mathcal{R}_{\mathbf{N},d}^{LB} = \left\{ \underset{k\in\{u,w\}}{\arg\min} \left\{ L_{\mathbf{N}+\mathbf{a}_{k,d}}^{k} \right\}, \underset{k\in\{u,w\}}{\arg\max} \left\{ L_{\mathbf{N}+\mathbf{a}_{k,d}}^{k} \right\} \right\}, \tag{3.26}$$

for voice and data services respectively, where vectors $\mathbf{a}_{k,v}$ and $\mathbf{a}_{k,d}$ are given in (3.13) for $k \in \{g, u, w\}$. According to (3.25) and (3.26), the selected RAT for a specific service, voice or data, will be the one that provided the allocation of the requesting service results in a lower load status. If the RAT exhibiting the lowest load is at full capacity, and thus the service cannot be accommodated, then the RAT that undergoes a higher load condition is chosen. Note that the formulation in (3.25) and (3.26) neglects the case when both target RATs exhibit equal load in the event of allocating the requesting service. In this eventual case, the selected RAT would be chosen randomly with equal probability among both candidates.

3.6 Parameter Settings, Scenario Definitions and Metrics of Interest

In this section the presented model will be evaluated considering different coverage and multimode terminal availability scenarios. The impact of presented RAT selection policies will be of special interest in order to determine the most suitable allocation principle in access-constrained scenarios. Preceding this evaluation, the considered scenarios must be identified and justified.

3.6.1 Access-Impaired Scenario Definitions

According to the eligibility probability of a specific RAT k for a given service j, $P_{e,j}\{r_k\}$, presented in (3.7), several scenarios can be characterized according to its component coverage and multimode terminal availability probabilities, $P_c\{r_k\}$ and $P_{t,j}\{r_k\}$ respectively, as defined in the following.

3.6.1.1 Coverage Availability Scenario Definitions

In this case, assuming all users with multimode terminals (i.e. $P_{t,j}\{r_k\}=1\forall k, j$), coverage conditions are indicated by probability $P_c\{r_k\}$ defined in Section 3.3.2 through the values of Ψ_i for $k \in \{g, u, w\}$. Accordingly, Table 3.4 contains the definition of three representative Coverage Scenarios (CS), namely CS1, CS2 and CS3. These scenarios mainly differ in the ability of WLAN to provide coverage to users. The adopted scenario definitions for CS1, CS2 and CS3 are motivated considering the adoption of the specified technologies along time, where CS1, CS2 and CS3 can be regarded as the coverage conditions in three different and consecutive time-epochs. Therefore, we have assumed the case where GSM is a widely adopted technology throughout CS1-CS3, whereas recently adopted technologies such as WLAN have experienced an increase in coverage. In the case of UMTS, some slight increase in coverage has also been assumed, although in less proportion to WLAN.

Although the model supports the definition of more diverse scenarios, the adopted definition will still allow assessing the model while favoring result interpretation and discussion.

Table 3.4: Coverage availability scenarios for numerical evaluation (Zero-probability events not shown).

Coverage event	Matrix $\mathbf{A}$ ($a_{(i,k)}$)	Description	Probability $\Psi_i = \Pr\{\mathcal{A}_i\}$		
($\mathcal{A}_i$)	$g\ u\ v$		CS1	CS2	CS3
$\mathcal{A}_1$	1 0 0	GSM coverage only	0.2	0.2	0.0
$\mathcal{A}_3$	1 1 0	GSM/UMTS coverage	0.6	0.2	0.0
$\mathcal{A}_7$	1 1 1	GSM/UMTS/WLAN coverage	0.2	0.6	1.0
Probability that a user is covered by each RAT ($P_c\{r_k\}$)					
		$P_c\{g\}$	1.0	1.0	1.0
		$P_c\{u\}$	0.8	0.8	1.0
		$P_c\{w\}$	0.2	0.6	1.0

Table 3.5: Terminal availability scenarios for numerical evaluation (zero-probability types not shown).

Type	Matrix $\mathbf{T}$ ($t_{(i,k)}$)	Description	Probability $\Upsilon_{i,j}$				
($\mathcal{T}_{i,j}$)			TS1		TS2		TS3
	$g\ u\ v$		v	d	v	d	v/d
$\mathcal{T}_{1,j}$	1 0 0	GSM terminal only	0.3	0.0	0.1	0.0	0.0
$\mathcal{T}_{3,j}$	1 1 0	GSM/UMTS terminal	0.6	0.9	0.5	0.6	0.0
$\mathcal{T}_{7,j}$	1 1 1	GSM/UMTS/WLAN terminal	0.1	0.1	0.4	0.4	1.0
Probability that a terminal supports each RAT ($P_{t,j}\{k\}$)							
		$P_{t,j}\{g\}$	1.0	-	1.0	-	1.0
		$P_{t,j}\{u\}$	0.7	1.0	0.9	1.0	1.0
		$P_{t,j}\{w\}$	-	0.1	-	0.4	1.0

3.6.1.2 Multi-mode Terminal Availability Scenarios

As for the scenarios concerning the availability of multi-mode terminals, the model presented in Section 3.3.1 enables the definition of a wide range of cases. Following the same principles as in the coverage scenarios in Section 3.6.1.2, it is assumed a number of terminal scenarios, namely TS1 to TS3, which refer to successive time-epochs, as reflected in Table 3.5. Specifically, three terminal types are considered in our case study: terminals only supporting GSM (type $\mathcal{T}_{1,j}$), terminals supporting both GSM and UMTS (type $\mathcal{T}_{3,j}$), and, finally, terminals supporting all three RATs (type $\mathcal{T}_{7,j}$). This choice is motivated by the fact that terminal evolution along time will adopt the capabilities of newly deployed RATs along with forerunner technologies (i.e. backward compatibility is assumed). As reflected in Table 3.5, it is considered the case where type $\mathcal{T}_{1,j}$ and $\mathcal{T}_{3,j}$ terminal penetration (given by probability $\Upsilon_{i,j}$) will decrease along time favoring the introduction of multi-mode terminal type $\mathcal{T}_{7,j}$ which experiences an increased penetration. Recall from Section 3.3.1, that terminal probability distribution is given for both voice and data users separately. Consequently, and bearing in mind that GSM is not able to handle data users, it is assumed that all voice users are equipped with GSM-capable terminals (i.e. $P_{t,v}\{g\} = 1$) while all data users are equipped with UMTS-capable terminals ($P_{t,d}\{u\} = 1$).

3.6.2 Performance Metrics

A wide range of performance metrics can be computed once the steady-state probabilities, $P_{\mathbf{N}}$, have been determined by solving the steady-state balance equations given in Section 3.2.4. In the following, a set of relevant metrics are presented and defined which will later be used to evaluate the performance of the proposed model. For convenience, we define the expectation of x, over the set space $\mathcal{S}$, as:

$$\mathbb{E}\left[x\right] = \sum\nolimits_{S_{\mathbf{N}} \in \mathcal{S}} x \cdot P_{\mathbf{N}}. \qquad (3.27)$$

3.6.2.1 Average Number of Served Users (Served Traffic)

The service/user distribution amongst the available RATs will be indicative of the RAT selection policy operation and, therefore, results of special interest. The average number of served service-type j users in RAT k ($N_{(j,k)}$) can be computed from the steady state probabilities, $P_{\mathbf{N}}$, as [similar to (2.29)]

$$N_{(j,k)} = \mathbb{E}\left[\mathbf{a}_{k,j} \cdot \mathbf{N}^{T}\right], \qquad (3.28)$$

where, for the specific case of voice and data services in a GSM/UMTS/WLAN scenario, $\mathbf{N}$ is given by (3.12), with $\mathbf{N^T}$ the transpose of vector $\mathbf{N}$, and $\mathbf{a_{k,j}}$ given by (3.13) for $j \in \{v, d\}$ and $k \in \{g, u, w\}$. It also follows that the per-service served number of users (N_j) is the sum of the average number of users over all RATs that uphold such service, i.e. [similar to (2.32)]

$$N_j = \sum\nolimits_{\forall k\,:\,(k,j)\in\mathcal{I}} N_{(j,k)}, \tag{3.29}$$

for $j \in \{v, d\}$ and $N_{(j,k)}$ given in (3.28).

3.6.2.2 Service Unavailability

We are mainly concerned with the problem that a user cannot be served due to the constraints imposed by non-service support, poor coverage or lack of appropriate terminal capabilities. A given user demanding a specific service j can be denied admission in RAT k if: 1) the RAT does not support the specific service; 2) no free resources are available, in which case we refer to this as *blocking*; and 3) no accessible resources are available due to terminal/coverage constraints. In this last case we will refer to *inaccessibility.* We may consequently define the following appropriate probabilities for the blocking and inaccessibility events.

The *Service Unavailability (SU) probability*, $P_{SU,j}$, defined as the probability that a specific service j is denied admission due to blocking and/or inaccessibility, is defined as

$$P_{SU,j} = P_{I,j} + (1 - P_{I,j}) \cdot P_{B,j}, \tag{3.30}$$

with $P_{B,j}$ and $P_{I,j}$ the blocking and inaccessibility probabilities of service j respectively.

The blocking probability of a service j ($P_{B,j}$) is defined as the probability of being in those states $S_{\mathbf{N}} \in \mathcal{S}$ in which the addition of single user with service type j forces a transition to a non-feasible state $S'_{\mathbf{N}} \notin \mathcal{S}$. Let $\mathcal{S}_{B,j}$ be the set of these blocking states, then the blocking probability of service j is

$$P_{B,j} = \sum\nolimits_{S_{\mathbf{N}}\in\mathcal{S}_{B,j}} P_{\mathbf{N}}. \tag{3.31}$$

The SU probability can be computed noting that the served traffic equals the offered traffic which has not been denied service, i.e. $T_j^{served} = T_j(1 - P_{SU,j})$. Since the served traffic equals the average number of users, i.e. $T_j^{served} = N_j$, with N_j given by (3.29), the SU probability yields

$$P_{SU,j} = 1 - N_j/T_j. \tag{3.32}$$

Then, the inaccessibility probability, $P_{I,j}$, is computed from (3.30) once $P_{B,j}$ and $P_{SU,j}$ are obtained with (3.31) and (3.32).

Note that the inaccessibility probability (and hence the SU probability) will largely depend on both coverage and terminal conditions. In fact, (3.30) indicates that even though infinite resources could be offered to a service-type j user, u_j, and thus a negligible blocking probability could be achieved, there might be a non-zero probability that such user cannot access any of its eligible RATs [i.e. those indicated by in (3.11)]. Then, some lower-bound for the SU probability may appear, i.e. $P_{SU,j} \geq P^*_{SU,j}$. While this is true when coverage limitations apply, note however that when only terminal limitation applies, given users u_j necessarily need to have terminal capabilities for accessing at least one RAT supporting service j, this lower-bound yields zero. Considering otherwise could imply, for example, that a particular user requesting a video-call session has a GSM terminal only, which results unrealistic.

3.6.2.3 Throughput

The average j-service throughput can be expressed as

$$\Gamma^j = \mathbb{E}\Big[\sum\nolimits_{\forall k\,:\,(k,j)\in\mathcal{I}} \Gamma_k^j(N_{k,j})\Big], \tag{3.33}$$

which reflects the additive contribution of per-service throughputs in each RAT, $\Gamma_k^j(N_{k,j})$. In turn, this throughput can be defined as the product between the number of users in each RAT ($N_{k,j}$) times the granted bit-rate per user ($R_{k,j}$), i.e.

$$\Gamma_k^j(N_{k,j}) = N_{k,j} \cdot R_{k,j}. \tag{3.34}$$

Note that, whereas voice users in GSM and UMTS along with data users in UMTS use dedicated channels (DCH), and thus a constant bit-rate is granted, for the case of WLAN, data throughput depends on the number of admitted users (recall from Section 3.5.1.3), i.e. $R_{w,d} = R_{w,d}(N_{w,d})$.

3.6.2.4 Average Load in each RAT

The average load in each RAT k can be computed from the steady state probabilities $P_{\mathbf{N}}$, as

$$L_k = \mathbb{E}\big[L^k_{\mathbf{N}}\big] \tag{3.35}$$

where $L^k_{\mathbf{N}}$ should be suitably particularized for each RAT.

3.6.2.5 Erlang Capacity

Finally, in order to compare the presented RAT selection policies in coverage/terminal-constrained scenarios we adopt the *Erlang Capacity* as, e.g. in [6]. The Erlang capacity defines a region comprised by the set of offered traffic loads of each service type, i.e. $T_v = \lambda_v/\mu_v$ and $T_d = \lambda_d/\mu_d$, such that some QoS requirement is fulfilled. Specifically, it is considered the SU probability, given in (3.32), as the QoS requirement to be met. Accordingly, the *Erlang Capacity Region* is given by

$$\mathcal{E} = \big\{(T_v, T_d) \ : \ P_{SU,v} \leq P^*_{SU,v}, P_{SU,d} \leq P^*_{SU,d}\big\}\,, \tag{3.36}$$

where $P^*_{SU,v}$ and $P^*_{SU,d}$ are design parameters indicating the maximum allowable SU probability values for voice and data users respectively. We will refer to the *Erlang Capacity Limit*, as the boundary of the Erlang Capacity Region. Fig. 3.1 illustrates the concept of the Erlang capacity region and limit.

In addition, the *traffic-mix* (or service-mix) is defined as the ratio $\sigma = T_v/(T_v+T_d)$, with $\sigma \in [0,1]$, indicating the dominant offered traffic, i.e voice-dominant if $\sigma > 0.5$ or data-dominant if $\sigma < 0.5$.

3.6.3 Parameter Settings

It is assumed that voice and data traffic exhibit activity factors of $\alpha_v = 0.8$ and $\alpha_d = 0.4$ respectively. For GSM, it is considered that two carriers (with 8 channels per carrier) are devoted for voice services, therefore yielding $C = 2\times 8 - 1 = 15$ channels (where one channel is devoted to signaling purposes). Bit-rates of voice users in GSM are assumed to be $R_{g,v} = 12.2$ kbps. As for UMTS, we adopt some representative numerical values given in [29], where the chip-rate is $W = 3.84$Mcps, voice and data bit-energy-to-noise ratio targets are $\theta_v = 7.9$dB and $\theta_d = 4.5$dB respectively, and granted bit-rates for voice and data users are $R_{u,v} = 12.2$kbps and

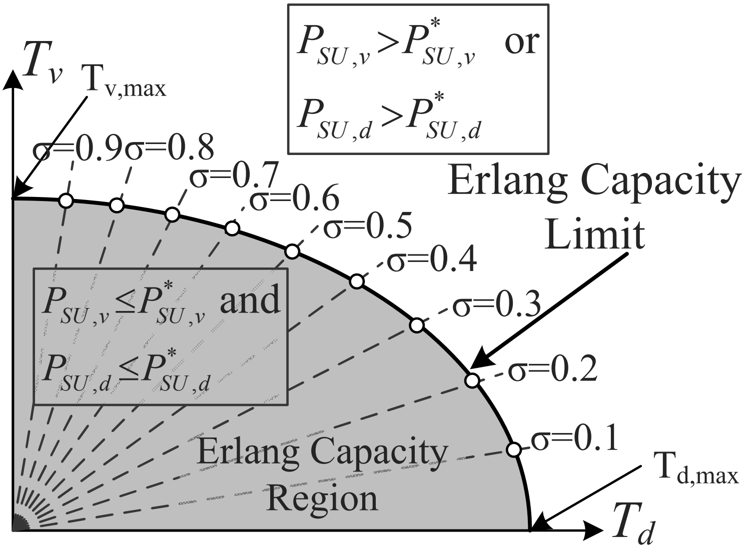


Figure 3.1: Erlang capacity region and limit. Erlang capacity limit for a particular traffic mix (σ) is represented with a circle.

$R_{u,d} = 144$kbps correspondingly. In addition, the average UL inter-cell-to-intra-cell interference is assumed to be $f = 0.73$ and the maximum UL load factor is $\eta_{max} = 0.8$.

In WLAN, parameters are set so that the minimum achieved data rate per user is $R^{min}_{w,d} = 1$Mbps which causes the maximum number of admitted data users in WLAN to be $N^{max}_{w,d} = 22$ ($n^{\max}_{w,d} = 9$). Table 3.6 lists the assumed parameters for numerical evaluation of the WLAN model as described in [28].

With the parameters given above, the RC of each service in each RAT, according to (3.15), (3.18) and (3.21), yields $\Delta\gamma_{g,v} = 1/15$, $\Delta\gamma_{u,v} \approx 1/30$, $\Delta\gamma_{u,d} \approx 1/12$ and $\Delta\gamma_{w,d} = 1/22$.

3.7 Numerical Results

Following subsections present numerical results to evaluate the performance of the proposed models.

Table 3.6: Parameters for numerical evaluation of WLAN model [28]. The assumed data service type corresponds to Best-Effort (BE) traffic (AC0 in IEEE802.11e specifications).

Parameter	Value
BE packet payload size	1500 bytes
Average Channel Bit Rate	11 Mbps
Slot Time	9 μs
SIFS	16 μs
Time required to transmit a RTS	15 μs
Time required to transmit a CTS	10 μs
Time required to transmit an ACK	10 μs
Packet propagation delay	1 μs
Time required to transmit a MAC header	10 μs
Time required to transmit a PHY header	48 μs
Minimum contention window for BE traffic	15
Arbitration Interframe Space Number (AIFSN)	2
Maximum backoff stage	6

3.7.1 Initial Model Assessment

Consider policy SB#1 (i.e. voice to GSM and data to WLAN as preferred option) in a scenario with perfect coverage conditions (i.e. CS3). Fig. 3.2 shows the service distribution (or carried traffic) in each RAT, i.e. voice users in GSM and UMTS [Fig. 3.2(a) and Fig. 3.2(b) respectively] along with data users in UMTS and WLAN [Fig. 3.2(c) and Fig. 3.2(d) resp.]. Provided GSM only handles voice users, the average served voice traffic in GSM (i.e., the average number of served voice users in GSM) is insensitive to the offered data traffic load, thus resulting in the behavior noted in Fig. 3.2(a). Likewise, SB#1 directs data users to WLAN as a preferred option. In this case, given WLAN does not support voice traffic, the carried data traffic in WLAN [see Fig. 3.2(d)] is insensitive to the offered voice traffic. Moreover, as expected, the carried data traffic increases with the offered data traffic. As for the traffic distribution in UMTS, both voice and data traffic that cannot be accommodated by GSM and WLAN respectively is directed to UMTS and thus compete for existing resources in this RAT. Note in Fig. 3.2(b) that the carried voice traffic in UMTS increases with the offered voice traffic but can be slightly reduced if the offered data traffic load increases [particularly noticeable for $T_v = 30$ Erlangs (E)]. Besides, data served traffic in UMTS [Fig. 3.2(c)] increases with rising offered data traffic and decreasing offered voice traffic.

3.7.2 Service Blocking vs. Service Unavailability

Fig. 3.3 shows the SU probability [Fig. 3.3(a) and defined in (3.32)] along with the blocking probability [Fig. 3.3(b), defined in (3.31)] and inaccessibility probability

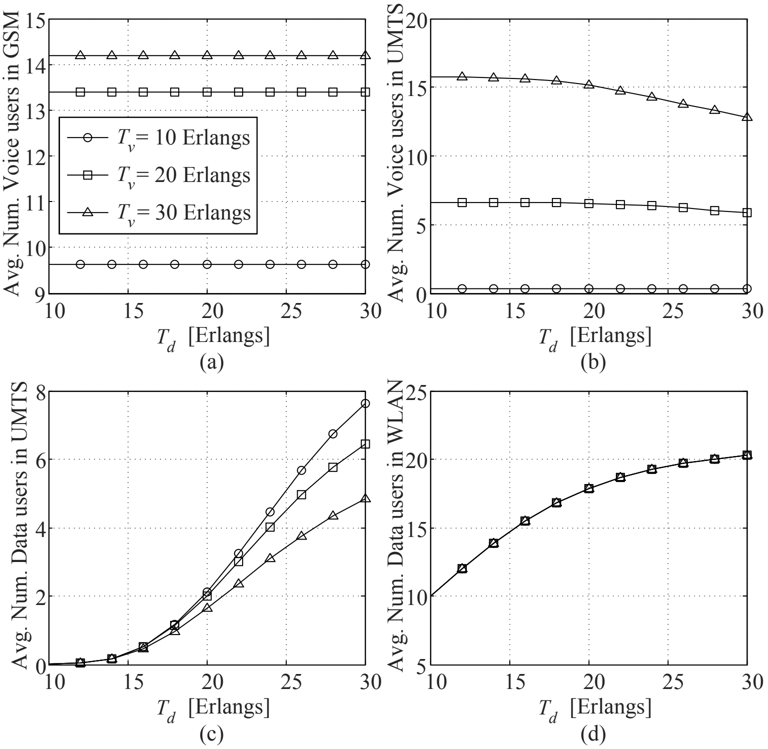


Figure 3.2: User distribution in each RAT when policy SB#1 and perfect coverage (CS3) applies. Legend in (a) applies elsewhere.

[Fig. 3.3(c), computed using (3.30)] for data users when scenarios CS1, CS2 and CS3 apply. As expected, the SU probability is lower-bounded by the inaccessibility probability; in this case $P^*_{I,d} = 0.2$ (refer to Table 3.4) for CS1 and CS2 and $P^*_{I,d} = 0$ for CS3. This means that, regardless of the offered data load, there will be always some probability that a user cannot access data resources due to lack of coverage. In addition, provided CS2 implies an increased WLAN coverage with respect to CS1, lower SU values can be potentially achieved, although in both cases the inaccessibility probability is the same. As for CS3, since all RATs are accessible, data SU in this scenario is lower-bounded by zero. Recall from (3.30) that the SU probability is a contribution of both the blocking probability and the inaccessibility probability. Then, Fig. 3.3 reveals that CS1 is solely affected by inaccessibility (with no blocking) whereas CS3 is solely affected by blocking probability (with no inaccessibility). In-between, CS2 represents a case where both inaccessibility and blocking [for values of $T_d \geq 30$E, see Fig. 3.3(b)] affect the SU probability.

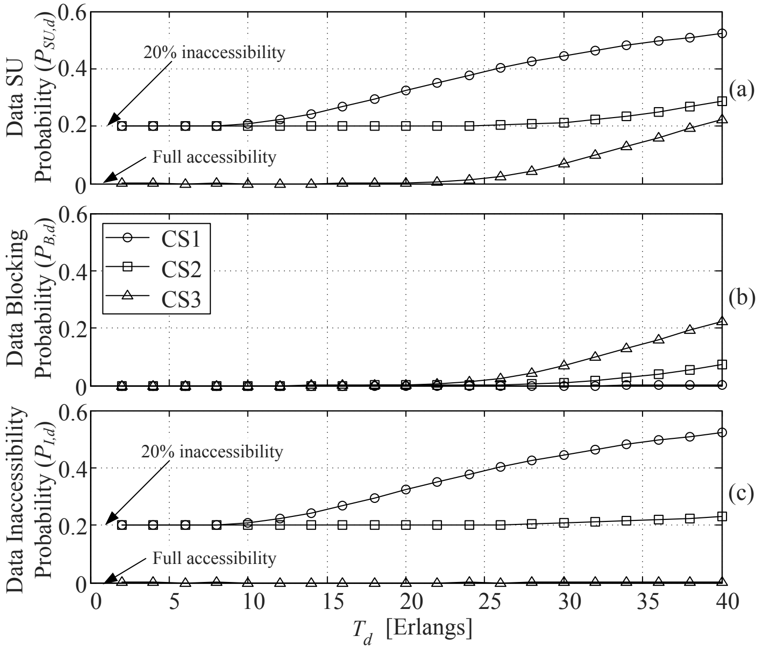


Figure 3.3: Behavior of the (a) service unavailability probability (b) blocking probability and (c) inaccessibility probability for data users in scenarios CS1, CS2 and CS3 for policy SB#1 and $T_v = 10$E. Legend in (b) applies elsewhere.

3.7.3 RAT Selection Comparison with Perfect Coverage/Terminal Conditions

As a first comparison between the different proposed RAT selection policies we consider the Erlang Capacity definition in (3.36) such that the target SU probabilities for both voice and data traffic are at most 5%, i.e. $P^*_{SU,v} = P^*_{SU,d} = 0.05$. Accordingly, Fig. 3.4 shows the Erlang capacity for service-based policies SB#1 to SB#4, along with the LB policy, for traffic-mix values $0 \leq \sigma \leq 1$ considering a 0.1 granularity.

According to Fig. 3.4, and regarding that the larger the Erlang Capacity region the higher the traffic that can be offered while guaranteeing QoS requirements, an overall improvement of both SB#1 and LB policies with respect to the rest is noted throughout the whole span of service-mix values.

LB policy intends to balance the loads among the 3 RATs, thus exhibits a higher flexibility in allocating resources between the RATs that uphold particular services.

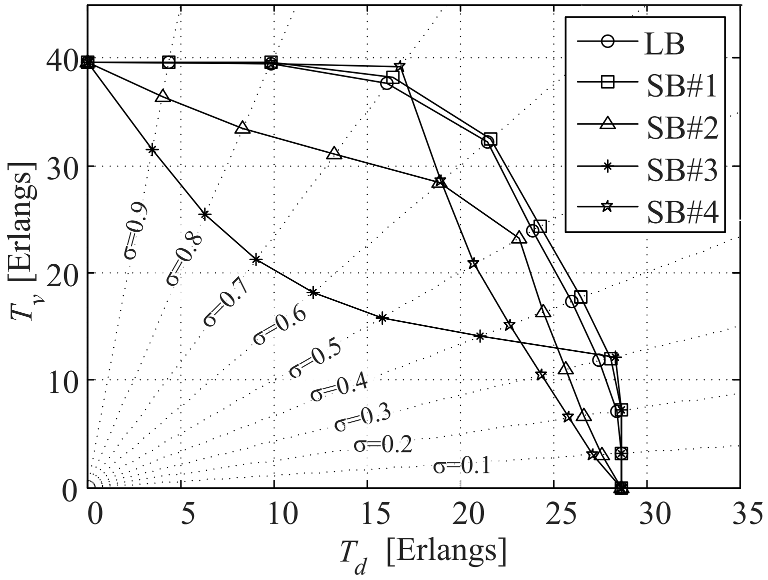


Figure 3.4: Erlang Capacity for the proposed policies with $P_{SU,v} \leq 0.05$ and $P_{SU,d} \leq 0.05$.

Since UMTS is able to handle both services, this policy prevents overload situations in UMTS.

For the case of SB#1, this strategy prevents from directing voice and data traffic to UMTS unless no capacity is available in GSM and WLAN respectively. As shown in Fig. 3.4, this policy is mainly suitable when the traffic-mix is $0.3 \leq \sigma \leq 0.7$, i.e. no high predominance of voice nor data traffic exists.

Oppositely, SB#2 directs both voice and data traffic to UMTS as a first option, driving UMTS to a blocking state where it cannot handle further traffic and, hence, causing an overall degraded operation, i.e. higher SU probability.

Moreover, SB#3, primarily allocating voice users to GSM and data users to UMTS, exhibits a good performance when the dominant traffic is data, particularly for $\sigma \leq 0.3$. Truly, given both voice and data users compete for UMTS resources, directing data users to UMTS and voice users to GSM will benefit the allocation of data traffic. However, if the fraction of voice traffic is increased, $\sigma \geq 0.3$, directing data users to UMTS will prevent voice users overflowing from GSM to UMTS, and given no voice traffic is handled by WLAN, will translate into a poorer performance of SB#3 with respect to the other policies.

Policy SB#4 behaves opposite with respect to policy SB#3. In this case, voice traffic is primarily directed to UMTS while data traffic is offered to WLAN. Consequently, in the case of dominant voice traffic, $\sigma \geq 0.7$, it is preferable to direct voice users to UMTS as a first option, and data users to WLAN. In this way, UMTS resources can be devoted to voice traffic.

As expected, for the cases where $\sigma = 0$ (i.e. $T_v = 0$E) or $\sigma = 1$ (i.e. $T_d = 0$E) all policies converge to the same Erlang capacity value. This is in fact an expected result, see e.g. [3], of considering all resources from available RATs as a whole, which is, on the other hand, the basis of CRRM.

Based on the results in Fig. 3.4 and the subsequent analysis, we may infer the following allocation guiding principles for the perfect-eligibility case:

(R1) Service-greedy, i.e. either voice-greedy or data-greedy, allocation principle is desirable when favorable traffic-mix conditions apply. Hence, voice-greedy (data-greedy) allocation is suitable in voice-dominant (data-dominant) traffic situations.

(R2) Late-contention is preferable to early-contention in UMTS.

In order to assess the benefits of CRRM as opposed to managing the resources of available RATs individually, we define an equivalent Erlang capacity region for this case as in [3]. Thus, the Erlang capacity region of the individual (non-CRRM) operation is obtained as an additive contribution of the Erlang Capacity bounds in each RAT. In this way, Fig. 3.5 shows the individual Erlang Capacity regions for GSM, UMTS and WLAN along with the combined Erlang capacity region when non-CRRM operation is considered. As stated in [3], the combined Erlang Capacity region in Fig. 3.5, obtained from the vector-sum of the Erlang capacity regions of individual RATs, largely depends on the allocation order and the service-mix. Here we consider the maximum possible Erlang capacity region, refer to [3], expressed as

$$\mathcal{E}_{NoCRRM} = \left\{ (T_v, T_d) : (T_v, T_d) = \sum\nolimits_{k=1}^{K} (T_v^{(k)}, T_d^{(k)}) \right\}, \tag{3.37}$$

where $T_j^{(k)}$ indicates the maximum offered j traffic load to RAT k such that some QoS guarantee is met. To compare the non-CRRM operation with the CRRM case studied earlier on, we consider a blocking probability of 0.05 for each service in each RAT as the target requirement to be fulfilled. Note that RAT-service compliance causes both GSM and WLAN Erlang capacity regions to be one-dimensional, as opposed to the two-dimensional UMTS case.

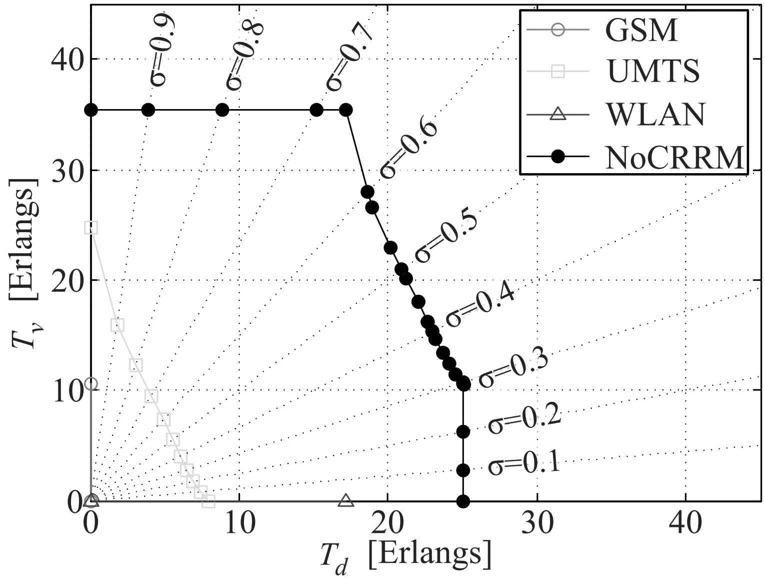


Figure 3.5: Stand-alone Erlang capacity regions for GSM, UMTS and WLAN. The vector-sum Erlang capacity region assuming individually-managed RATs is also plotted (labeled as NoCRRM).

A gain metric considering the use of CRRM (through RAT selection policies) with respect to the non-CRRM case can be defined. Then, the *Erlang Capacity Gain* (ECG) is given by

$$E_G(\sigma) = (\|(T_v, T_d)\|_{\pi} / \|(T_v, T_d)\|_{NoCRRM} - 1) \times 100, \tag{3.38}$$

with $\|(x, y)\|=(x^2 + y^2)^{1/2}$. Fig. 3.6 shows the ECG for the considered RAT selection policies. One may first observe that, while some policies clearly exhibit some gain with respect to the NoCRRM case, i.e. LB and SB#1, other policies reveal poorer performance than the case of NoCRRM. However, as mentioned in [3], the NoCRRM case here considered is the maximum achievable Erlang capacity region, thus the worse case scenario when comparing to the CRRM case. In addition, it is worthwhile noting that the ECG is an additive contribution of the so-called *trunking gain* and the *selection gain*. Trunking gain refers to the capacity gain obtained by considering a pool of resources as a whole rather than the additive contribution of individual resources. Then, the trunking gain is reflected in Fig. 3.6 when σ=0 and $\sigma=1$, yielding the data (14.79%) and voice (11.87%) trunking gains respectively. As for the selection gain, it refers to the ability in allocating resources to the different RATs. As shown in Fig. 3.6, some allocation principles result in improved ECG according to the offered traffic mix. On the contrary, other RAT selection policies may result extremely hazardous in some situations, e.g. SB#3 for σ=0.7 results in 40% degradation with respect to the NoCRRM case.

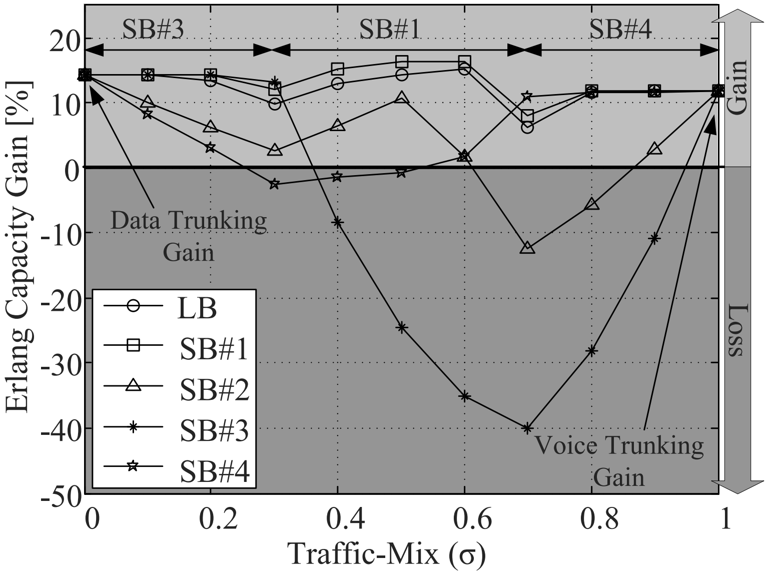


Figure 3.6: Erlang Capacity Gain of the proposed policies with respect to the non-CRRM operation. Graph annotations at the top indicate most suitable RAT selection policy according to Erlang Capacity Gain.

Whereas the Erlang capacity region is able to capture the ability of the considered RATs to efficiently allocate resources, other effects must not be disregarded. Then, in terms of throughput, Fig. 3.7 shows the average data throughput (Γ^d) served by the different RAT selection policies and several offered traffic-mix configurations, $\sigma = \{0.2, 0.5, 0.8\}$, which represent data-dominant, equal service-mix, and voice-dominant scenarios respectively. Throughput values are plotted for each policy as long as offered traffic mix pairs (T_d, T_v) lay in their respective Erlang capacity regions given by Fig. 3.4. For data-dominant situations, Fig. 3.7(a), we observe that for low offered traffic values the data throughput benefits from data users being directed to WLAN, as in SB#1 and SB#4. Conversely, directing data users to UMTS as a preferred option, i.e. SB#2 and SB#3, does not favor data throughput if the offered data traffic is low, but may result in improved throughput performance for higher offered traffic values as shown in Fig. 3.7(a). This is because data users in UMTS achieve a 144kbps data service as opposed to WLAN that ensures a 1 Mbps minimum bit rate per user. In this case, SB#2 results in the policy that achieves higher throughput when $T_d \in [17, 20]$E, and it is subsequently surpassed by policy SB#3, given its data-greedy nature allowing a high dedication of resources towards data users. Policy LB on the other hand, exhibits a poorer performance with respect to SB#1 and SB#4 since it may allocate data users to UMTS in order to keep loads balanced. However, it exhibits an improved performance as the

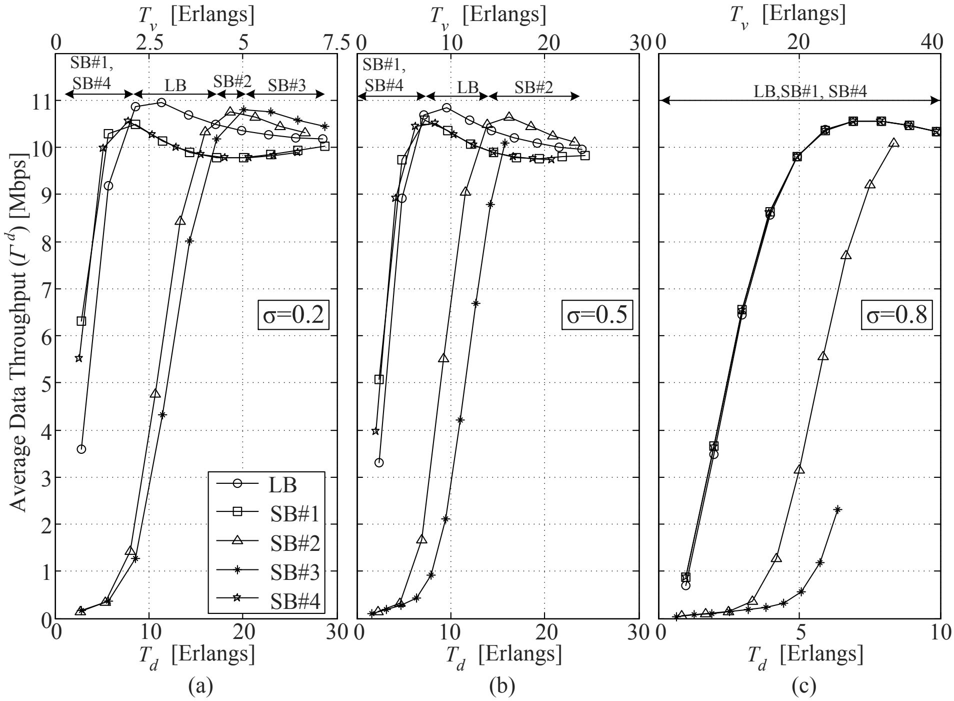


Figure 3.7: Data throughput against offered data load (T_d) and voice load (T_v) for the proposed RAT selection policies in the case of (a) data-dominant services (σ=0.2), (b) equal service-mix (σ=0.5) and (c) voice-dominant services (σ=0.8). Graph annotations indicate the most suitable RAT selection policy in terms of data throughput for a specific range of offered traffic. Legend in (a) applies elsewhere.

data traffic load is increased. For an equal-service mix scenario, see Fig. 3.7(b), a similar behavior is noted for LB, SB#1, SB#2 and SB#4 with respect to the previous case. For the case of SB#3, reduced Erlang capacity, as shown in Fig. 3.4, prevents from using this policy if SU probability guarantees are to be maintained, and thus data throughput above T_d≈15.8E are not even considered although they would result in improved data throughput values as in the data-dominant case. Finally, voice-dominant situations, as reflected in Fig. 3.7(c), suggest the use of policies that do not degrade voice SU in excess, therefore explaining the poorer performance of SB#2 and SB#3 which attempt to use UMTS shared resources. Conversely, LB, SB#1 and SB#4 achieve similar performances since data traffic is mainly directed to WLAN devoting UMTS resources to voice traffic.

From Fig. 3.7 it is worthwhile noting that, although specific policies could be selected according to measured traffic-mix conditions (which could be highly varying along time) the LB policy presents an overall improved performance in all presented

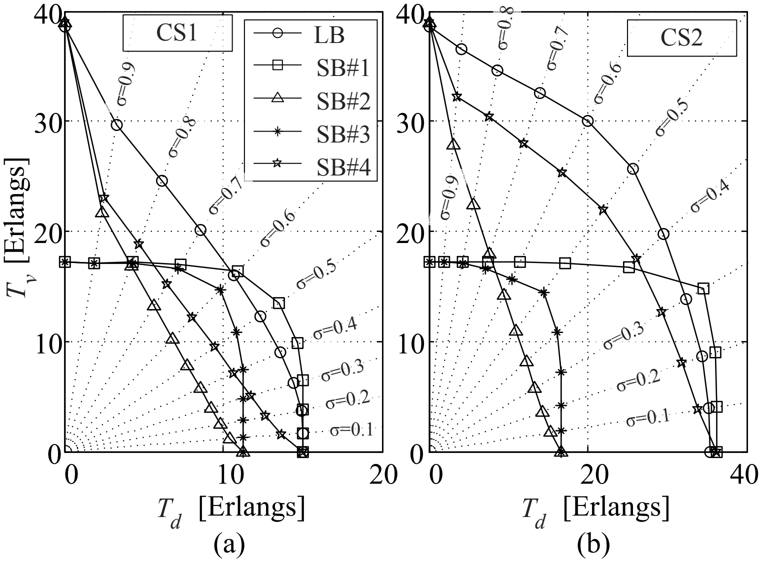


Figure 3.8: Erlang Capacity regions for coverage scenarios (a) CS1 and (b) CS2 for $P_{SU,v} \leq 0.05$ and $P_{SU,d} \leq 0.25$. Legend in (a) also applies in (b).

cases, thus suggesting a wide applicability and use while still maintaining a good performance. Recall that the same conclusions were drawn for the LB policy in Chapter 2 suggesting its applicability.

3.7.4 Non-Perfect Coverage Impact

In this section, the comparison of the proposed RAT selection policies is carried for coverage scenarios CS1, CS2 and CS3.

A good way to assess the combined behavior of voice and data traffic is using the Erlang Capacity region as depicted in Fig. 3.8. Note that, since CS1 and CS2 mainly differ in the ability of WLAN to capture data traffic, it is noteworthy the differences in offered data traffic (x-axis) that both scenarios uphold. Whereas for CS1 a maximum data traffic load of $T_d = 15$E can be offered, CS2 allows up to $T_d = 36.2$E to be offered while still maintaining the required QoS guarantees.

Observing Fig. 3.8 a main guiding principle may be identified for the non-perfect coverage case:

(R3) Worst-coverage-first allocation principle is desirable.

In the following we justify the abovementioned rule [along with rules (R1) and (R2) given in Section 3.7.3] by exploring the behavior of the proposed RAT selection policies.

Rule (R3) is motivated by the fact that if we first attempt admission to the RAT that undergoes the poorest coverage conditions, we could still attempt admission to RATs with improved coverage if first attempt fails. Otherwise, if the opposite rule to (R3) was applied, we could end up with improved coverage RATs at full capacity and with users attempting admission to coverage-limited RATs being denied service. A clear example of this is SB#3, where voice is directed to GSM and data is directed to UMTS. Both GSM and UMTS represent best-coverage-first for voice and data users respectively, i.e. opposite rule to (R3). We observe how the Erlang capacity is severely limited in both offered voice and data traffic dimensions. In addition, SB#1 only follows rule (R3) for data traffic, i.e. it directs voice users to GSM while data users are primarily directed to WLAN. In this case, improvements are only noticed when predominant traffic is data [see Fig. 3.8(a) for $\sigma \leq 0.6$ and Fig. 3.8(b) for $\sigma \leq 0.3$]. Similarly, SB#2 follows rule (R3) for voice users only, by directing those to UMTS whereas data users are sent also to UMTS. In this case, Erlang capacity is improved with respect to SB#1 and SB#3 when offered traffic is predominantly voice, i.e. $\sigma \geq 0.8$ for CS1 [Fig. 3.8(a)] and $\sigma \geq 0.7$ for CS2 [Fig. 3.8(b)]. As for SB#4, it follows rule (R3) since it directs traffic to the correspondent worse coverage RAT (i.e. voice to UMTS and data to WLAN). However, it violates rule (R1) when data traffic is dominant since SB#4 is also a voice-greedy policy. Hence, although Erlang capacity improvement is noted for SB#4 with respect to the other SB policies for voice-dominant situations, the collision between rules (R3) and (R1) causes a reduction in the Erlang capacity in data-dominant cases. This causes SB#1, which follows rule (R3) for data traffic, to outperform SB#4 in data-dominant traffic cases as reflected in Fig. 3.8(a) and Fig. 3.8(b). In addition, we note that rule (R2), i.e. late-contention better than early-contention, does not necessarily hold in a non-perfect coverage case by observing that SB#2 (early-contention) outperforms SB#1 (late-contention) in voice-dominant situations.

Finally, LB seems to achieve an overall good performance throughout the span of traffic-mix values, due to its flexible operation in allocating resources, only surpassed by SB#1 in data-dominant cases.

Summarizing, despite rule (R3) (allocate to worst coverage first) seems to be a suitable allocation principle, it should be used in conjunction with rule (R1). In particular, for voice or data traffic-dominant situations, following rule (R1) is suggested if both rules (R1) and (R3) are contradictory. Since LB effectively balances

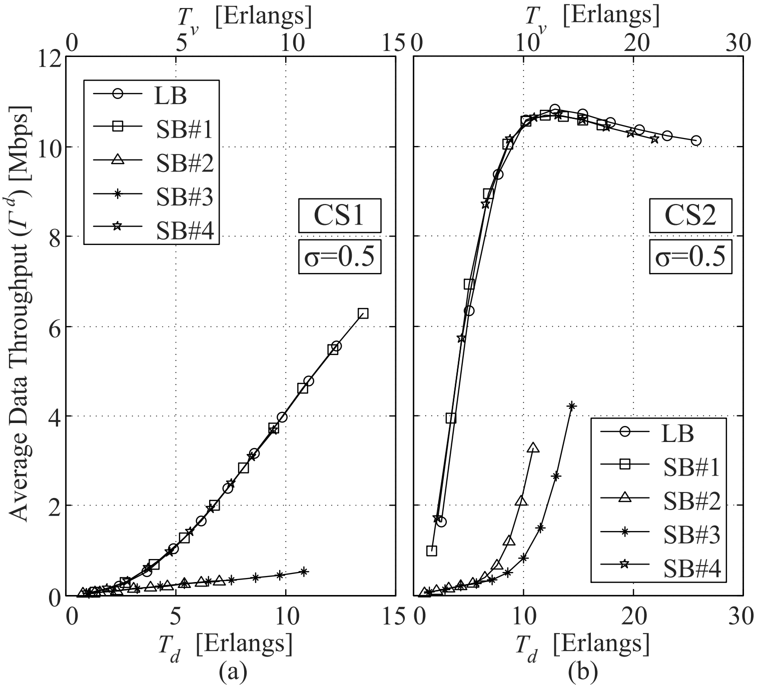


Figure 3.9: Data throughput against offered data load (T_d) and voice load (T_v) for the proposed RAT selection policies in (a) CS1 and (b) CS2 in the case of equal service-mix ($\sigma = 0.5$).

the impact of coverage and traffic-mix conditions, it is well-suited in cases where the traffic-mix is not known a priori or vaguely estimated.

Fig. 3.9 and shows the average data throughput in CS1 and CS2 for $\sigma = 0.5$ (i.e. equal service-mix). We note that directing data users to UMTS as a first option (i.e. policies SB#2 and SB#3) results in a poorer throughput performance with respect to the other policies. First, UMTS provides 144Kbps bit-rate as opposed to 1Mbps (minimum) of WLAN; and second, SB#2 and SB#3 do not comply with (R3) which causes an increased SU. Regarding SB#1 and SB#4, since they both direct data users to WLAN, which complies with (R3) and allows also a higher throughput-per-user, similar throughput values are attained in their respective SU probability ranges. LB policy on the other side performs similar to SB#1 and SB#4. Since the coverage limitation is on WLAN, the LB policy will try to allocate as many data users in this RAT in order to balance loads.

3.7.5 Multi-Mode Terminal Heterogeneity Impact

Assuming full coverage conditions, i.e. $P_c\{r_k\} = 1\ \forall k$, we now face the problem of users with terminals that may not support particular RATs. In this case, voice users may only be capable of accessing GSM whereas data users may only be capable of accessing UMTS as reflected in Table 3.5. The study of this scenario is analogous to the coverage-limited scenario cases studied in the previous section.

Fig. 3.10 shows the Erlang capacity limits for the considered terminal scenarios TS1 [Fig. 3.10(a)] and TS2 [Fig. 3.10(b)], where the QoS constraints are set to $P_{SU,v}$=$P_{SU,d}$≤0.05. In this case, and oppositely to the coverage case (see Fig. 3.3), the inaccessibility concept does not apply, in the sense that a voice (data) user is assumed to have a terminal with accessibility to at least one RAT supporting voice (data) services. Note that considering otherwise would be unrealistic. This means that QoS constraints can be set as low as desired provided $P_{SU,v}, P_{SU,d}>0$. A comparison of Fig. 3.10 with Fig. 3.8 (where the coverage scenarios are depicted) indicates similar trends and parallelism between coverage and terminal heterogeneity limitations on the Erlang capacity for the considered RATs. Hence, similar conclusions may be extracted, i.e., directing multi-mode voice (data) users to UMTS (WLAN) seems to be a better allocation principle than operating in the opposite way. This enables to identify the following rule [analogous to (R3)]:

(R4) Terminal-limited-RAT-first allocation principle is desirable.

In addition to (R4), also rule (R1) must not be disregarded. Then, when data traffic is dominant, policies SB#1 and SB#4 achieve good performance with an improvement of SB#1 over SB#4 given the former prevents from the use of UMTS resources for voice users whereas the latter, SB#4, is a voice-greedy policy which monopolizes UMTS resources. In the cases where voice traffic is predominant, policies SB#2 and SB#4 follow rule (R4) by directing voice users to UMTS as a preferred option. In this case, both policies achieve similar performances given voice users are able to occupy UMTS resources in SB#2 provided voice traffic is dominant, and, on the other hand, SB#4 is a voice-greedy policy. Lastly, the LB policy achieves an overall good performance as already stated in previous sections.

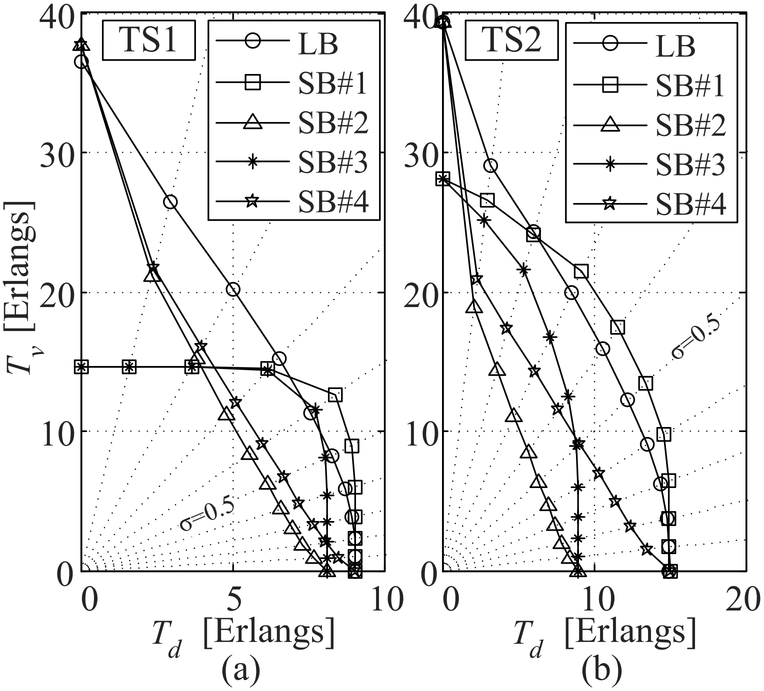


Figure 3.10: Erlang Capacity regions for terminal scenarios (a) TS1 and (b) TS2 for $P_{SU,v} \leq 0.05$ and $P_{SU,d} \leq 0.05$.

3.8 Chapter Summary

This chapter has presented a Markovian framework for the evaluation of multi-service allocation principles in multi-RAT heterogeneous networks. The proposed model offers a wide flexibility and applicability with respect to existing works by considering diverse access-limited scenarios: these are, but not limited to, multi-access coverage overlapping conditions, non-homogeneous user distribution and the ability of efficient coverage and multi-mode terminal availability. In this sense, a probabilistic approach is considered to account for both coverage and terminal heterogeneity scenarios allowing the definition of a wide range of access-limited scenarios.

Under this framework, several RAT selection policies have been defined, considering various service-based policies and a load-balancing policy. Our particular study case focuses on the allocation of voice and data services in a heterogeneous network composed by GSM, UMTS and WLAN RATs, but may be easily extended to include other technologies.

A thorough evaluation of the proposed model was carried our in diverse scenarios accounting for coverage and terminal heterogeneity. Three main factors have been identified to largely influence the performance of the RAT selection process:

- *Traffic-mix*: It affects the suitability of a particular RAT selection policy when allocating multiple services on multiple RATs. No single SB policy outperforms others throughout the whole span of traffic-mix values.
- *Resource contention*: Different RATs upholding a particular number of services will lead to different contention situations among services. In these situations, service-greedy situations may appear and must be taken into consideration when allocating multiple services on multiple RATs. For example, in our specific study, solely UMTS provided resource contention between voice and data users, thus yielding data-greedy (SB#3) and voice-greedy (SB#4) allocation principles.
- *RAT-Service eligibility*: Both coverage availability and terminal support determine if a particular RAT is "eligible" in the RAT selection procedure. If only one RAT is eligible the RAT selection process is straightforward. On the other hand, if more than one candidate RAT is eligible, it is useful to know which RAT undergoes improved or degraded eligibility.

Based on the influencing parameters described above, specific guidelines regarding the evaluated scenario can be extracted from the performance evaluation figures:

- When perfect eligibility applies (i.e. CS3 or TS3), service-greedy allocation results in a suitable allocation principle (in terms of Erlang Capacity) if according traffic-mix conditions are favorable [essentially rule (R1)]. Then, see Fig. 3.4, voice-greedy and data-greedy policies (i.e. SB#4 and SB#3 respectively) are suggested in voice-dominant and data-dominant conditions respectively.
- If perfect accessibility applies, late-contention (SB#1) is better than early-contention (SB#2) in terms of Erlang capacity (see Fig. 3.4). This was captured by rule (R2).
- In terms of data throughput, a trade-off arises between higher throughput-per-user achieved in WLAN and the contention for UMTS resources against voice users. For perfect eligibility conditions, policies SB#1 and SB#4, which direct data users to WLAN in the first place, are suitable when offered data traffic can be handled by WLAN, thus UMTS resources can be devoted to

voice users if necessary. On the contrary, if data traffic is large enough to require UMTS resources, then policies SB#2 and SB#3 favor the competition for UMTS resources thus being more suitable in this case (see Fig. 3.7). As for coverage-limited scenarios, allocation principles that comply with rule (R3) achieve improved throughput performance (see Fig. 3.9).

- In limited-access scenarios the rule *access-limited-first*, which corresponds to (R3) in the coverage-limited case and to (R4) in the terminal-limited case, provides a good guiding principle for the allocation of multiple services to multiple RATs. This rule should be used in combination with rule (R1) making sure that favorable service-mixes apply.

- Finally, LB policy exhibits an overall improved behavior regardless of the traffic-mix and RAT eligibility. It is then suggested when the service distribution is unknown or vaguely estimated.

Bibliography

[1] O. Falowo and H. Chan, “Joint call admission control algorithms: Requirements, approaches, and design considerations,” *Computer Communications*, vol. 31, no. 6, pp. 1200–1217, April 2008.

[2] A. Tolli, P. Hakalin, and H. Holma, “Performance evaluation of common radio resource management (CRRM),” in *Communications. IEEE International Conference on*, 2002, pp. 3429–33 vol.5.

[3] I. Koo, A. Furuskar, J. Zander, and Kiseon Kim, “Erlang capacity of multiaccess systems with service-based access selection,” *IEEE Commun. Lett.*, vol. 8, no. 11, pp. 662–664, 2004.

[4] J. Pérez-Romero, O. Sallent, and R. Agustí, “On the optimum traffic allocation in heterogeneous CDMA/TDMA networks,” *IEEE Trans. Wireless Commun.*, vol. 6, no. 9, pp. 3170–3174, 2007.

[5] Remco Litjens, Ljupco Jorguseski, and Mariya Popova, “Performance modelling and evaluation of wireless multi-access networks,” in *Next Generation Teletraffic and Wired/Wireless Advanced Networking*, pp. 194–209. 2007.

[6] A. Furuskar and J. Zander, “Multiservice allocation for multiaccess wireless systems,” *IEEE Trans. Wireless Commun.*, vol. 4, no. 1, pp. 174–184, 2005.

[7] I. Cananea, D. Mariz, J. Kelner, D. Sadok, and G. Fodor, “An on-line access selection algorithm for ABC networks supporting elastic services,” in *Wireless Communications and Networking Conference. IEEE*, Apr. 2008, pp. 2033–38.

[8] J. Pérez-Romero, O. Sallent, and R. Agustí, “Network controlled cell breathing in multi-service heterogeneous CDMA/TDMA scenarios,” in *Vehicular Technology Conference, 2006. VTC-2006 Fall. 2006 IEEE 64th*, 2006.

[9] Susan Lincke-Salecker and Cynthia S. Hood, “Integrated networks that overflow speech and data between component networks,” *International Journal of Network Management*, vol. 12, no. 4, pp. 235–257, 2002.

[10] 3GPP, “Improvement of RRM across RNS and RNS/BSS,” TR 25.881 v5.0.0, 3rd Generation Partnership Project (3GPP), 2001.

[11] 3GPP, "Improvement of RRM across RNS and RNS/BSS (post rel-5) (release 6)," TR 25.891 v0.3.0, 3rd Generation Partnership Project (3GPP), 2003.

[12] O. E. Falowo, N. Ventura, and H. A. Chan, "Effect of mobile terminal heterogeneity on connection-level QoS in next generation wireless networks," in *Electrical and Computer Engineering. Canadian Conference on*, 2009, pp. 931–935.

[13] Cheng-Fu Chou, Ching-Ju Lin, and Chung-Chieh Tsai, "Traffic-aware resource management in heterogeneous cellular networks," in *Wireless Networks, Communications and Mobile Computing, 2005 International Conference on*, December 2005, pp. 762–767 vol.1.

[14] X. Gelabert, J. Pérez-Romero, O. Sallent, and R. Agustí, "A Markovian approach to radio access technology selection in heterogeneous multiaccess/multiservice wireless networks," *Mobile Computing, IEEE Transactions on*, vol. 7, no. 10, 2008.

[15] X. Gelabert, J. Pérez-Romero, O. Sallent, and R. Agustí, "A 4-Dimensional Markov Model for the Evaluation of Radio Access Technology Selection Strategies in Multiservice Scenarios," in *Proc. 64th Semi-annual IEEE Vehicular Technology Conference Fall (VTC-Fall'06)*, Montreal, Canada, September 25-28, 2006.

[16] Yi-Bing Lin, Wei-Ru Lai, and Rong-Jaye Chen, "Performance analysis for dual band PCS networks," *IEEE Trans. Comput.*, vol. 49, no. 2, pp. 148–159, 2000.

[17] Yuguang Fang, Imrich Chlamtac, and Yi-Bing Lin, "Modeling pcs networks under general call holding time and cell residence time distributions," *IEEE/ACM Trans. Netw.*, vol. 5, no. 6, pp. 893–906, December 1997.

[18] Daehyoung Hong and S. S. Rappaport, "Traffic model and performance analysis for cellular mobile radio telephone systems with prioritized and nonprioritized handoff procedures," *IEEE Trans. Veh. Technol.*, vol. 35, no. 3, pp. 77–92, 1986.

[19] Qing-An Zeng, Kaiji Mukumoto, and Akira Fukuda, "Performance analysis of mobile cellular radio systems with two-level priority reservation handoff procedure," *IEICE transactions on communications*, vol. 80, no. 4, pp. 598–607, 1997.

[20] S. Nanda, "Teletraffic models for urban and suburban microcells: cell sizes and handoff rates," *IEEE Trans. Veh. Technol.*, vol. 42, no. 4, pp. 673–682, 1993.

[21] Jingao Wang, Qing-An Zeng, and D. P. Agrawal, "Performance analysis of a preemptive and priority reservation handoff scheme for integrated service-based wireless mobile networks," *IEEE Trans. Mobile Comput.*, vol. 2, no. 1, pp. 65–75, 2003.

[22] E. Gustafsson and A. Jonsson, "Always best connected," *Wireless Communications, IEEE [see also IEEE Personal Communications]*, vol. 10, no. 1, pp. 49–55, 2003.

[23] Gunter Bolch, Stefan Greiner, Hermann de Meer, and Kishor S. Trivedi, *Queueing Networks and Markov Chains: Modeling and Performance Evaluation with Computer Science Applications*, WileyBlackwell, 2nd edition edition, May 2006.

[24] Leonard Kleinrock, *Queueing Systems. Volume 1: Theory*, Wiley-Interscience, 1 edition, January 1975.

[25] Eitan Altman, Tijani Chahed, and Salah E. Elayoubi, "Joint uplink and downlink capacity considerations in admission control in multiservice CDMA/HSDPA systems," in *Performance evaluation methodologies and tools. Proc. 2nd int. conf. on*, 2007.

[26] C. Tarhini and T. Chahed, "System capacity in OFDMA-based WiMAX," in *Systems and Networks Communications, 2006. ICSNC '06. International Conference on*, December 2006, p. 70.

[27] Jordi Pérez-Romero, Oriol Sallent, Ramon Agustí, and Miguel Ángel Díaz-Guerra, *Radio Resource Management Strategies in UMTS*, John Wiley & Sons, 2005.

[28] Yu-Liang Kuo, Chi-Hung Lu, E. H. K. Wu, and Gen-Huey Chen, "An admission control strategy for differentiated services in ieee 802.11," in *Global Telecommunications Conference, 2003. GLOBECOM '03. IEEE*, December 2003, pp. 707–712 Vol.2.

[29] J. Kelif, E. Altman, and I. Koukoutsidis, "Admission and GoS control in a multiservice

WCDMA system," *Computer Networks*, vol. 51, no. 3, pp. 699–711, February 2007.

Chapter 4

Radio Access Congestion in Multi-Access/Multi-Service Wireless Networks

This chapter addresses the problem of radio access congestion control and resource allocation in scenarios where multiple available Radio Access Technologies (RATs) support a wide range of services over a given coverage area. A key issue in these networks is to select the most appropriate RAT at call/session establishment according to some specified user/operator criteria. In this sense, as shown in Chapters 2 and 3, a wide range of high-level policies can be defined providing the most favorable resource allocation. Regardless of having efficient RAT selection policies which may ensure some initial Quality of Service (QoS) requirements, intrinsic network dynamics (e.g. mobility, user activity, interference rise, etc.) can cause potential QoS failures leading to a degraded network performance, and hence, to radio access congestion. This chapter is devoted to study the impact of radio access congestion on a number of RAT selection policies. Consequently, a congestion probability model is developed to capture the statistical behavior of radio access congestion events. In addition, as in Chapter 3, a general Markovian framework is adopted for evaluating the allocation of multiple services into multiple RATs by means of high-level policy definitions. As a difference from Chapter 3, both uplink and downlink considerations are made in this chapter. Specific RAT selection policies are defined according to several criteria and its performance evaluated in a TDMA/WCDMA multi-RAT scenario supporting voice and data services. Moreover, the use of congestion probability information as a possible allocation principle for RAT selection is also evaluated, which, in the assumed scenario, results in the most favorable allocation policy.

4.1 Radio Access Congestion: Motivation and Problem Statement

Among the well-known RRM strategies, Congestion Control (CC) is the function devoted to overcome potential Quality of Service (QoS) failures at the radio interface layer due to the intrinsic dynamics of the network (e.g. mobility, interference rise, traffic variability, etc.) [1]. Regardless of having a strict CAC mechanism, which may ensure some QoS requirements at call/session establishment, if the dynamics of certain network parameters suffers from high random behavior, the network may experience high-load/high-interference situations which in turn may degrade the QoS perceived by users. Consequently, CC strategies are designed so as to minimize the negative impact of radio access congestion on the network performance. Typically, [1], CC mechanisms involve three phases, namely: congestion detection (CD), congestion resolution (CR) and congestion recovery (CRV). CD is responsible for monitoring the network status in order to correctly identify a congestion situation by means of RAT-specific measurements. On the other hand, CR actuates over a set of RAT-specific parameters in order to reduce the load and consequently the congestion situation. Finally, CRV will attempt to restore the transmission parameters that were set before the congestion was triggered.

In a B3G scenario with multiple RATs and multiple services, congestion situations in each RAT can be statistically characterized by means of a *congestion probability* given the number of allocated users. Then, provided a set of initial RAT selection policies, the evaluation of the congestion impact of these policies is of special interest. Specially because favorable user allocations according to the congestion probability in each of the existing RATs can be found. Consequently, the adopted approach in this chapter is *proactive*, in the sense that potential congestion problems are anticipated by means of efficient initial RAT selection, as opposed to *reactive*, i.e. acting after congestion is detected, which would include CR mechanisms such as, e.g., those in [2, 3]. In addition, if the congestion probability is known a priori (through models, real measurement campaigns, etc.) for a given user distribution, an initial RAT selection policy can be designed using such information so that the resulting congestion probability is kept as low as possible.

Under the above framework, this chapter intends to define and assess the impact of radio access congestion in RATs with different underlying technologies, such as TDMA and WCDMA, that are subject to different traffic load conditions. The main objective is to determine user allocation principles at call/session establishment, i.e. initial RAT selection policies, which are favorable in terms of potential congestion eventualities. Besides the valuable interest of determining such initial

RAT selection policies, this will provide a solid basis for prospective evaluations contemplating the support of inter-RAT handovers (or VHOs as defined in Section 1.2.1.2), which has not been considered in this study for the sake of model tractability. In addition, this chapter proposes a novel initial RAT selection policy which allocates users so as to minimize the resulting congestion probability. Further detail on the contributions of this chapter is presented in the next section.

4.1.1 Related Work and Main Contributions

The Third Generation Partnership Project (3GPP) standardization body identified some CRRM architectures and procedures for integrated GSM/EDGE and UMTS network operation in [4, 5].

Despite the large amount of work devoted to CRRM algorithms in recent years (specially in what refers to RAT selection, see e.g. [6–10]), to the best of the author's knowledge, very few contributions have been made towards radio access common congestion control strategies. The authors in [11] propose several schemes for controlling and distributing the network traffic over two RATs. Cost metrics are assigned to each service in each RAT, and load control is evaluated by means of initial RAT selection and VHO. In [2, 3], the author presented a framework for managing congestion situations in heterogeneous networks. In particular, practical methodologies for congestion detection in GSM/EDGE and UMTS networks are provided along with VHO and bit-rate reduction techniques so as to lessen the congestion status of the network.

While aforementioned approaches usually rely on extensive system-level simulations, a theoretical approach to the problem seems to be uncovered, and thus constitutes a novel contribution. It could then be of great interest to gain insight into the general problem as well as inspire the definition of practical and efficient congestion control strategies. In this sense, the main contributions of this chapter include:

- To propose a general Markovian framework for the allocation of multiple services into multiple RATs by means of simple initial RAT selection policies. This is covered in section 4.2 and, along with the inclusion of both uplink (UL) and downlink (DL) procedures, extends previous work done by the authors in Chapters 2 and 3.

- To define a statistical model for the characterization of radio access congestion in B3G scenarios, which is provided in section 4.3. This model takes into account the congestion that may eventually arise in both the uplink and downlink provided a particular distribution of services is known. Such service distribution will be provided by the Markov allocation framework given in section 4.2.

- To describe several initial RAT selection policies responding to different allocation principles within the specified multi-service/multi-RAT framework. In particular, a novel initial RAT selection policy that takes advantage of congestion information is presented, which can be found in section 4.4.

- For a particular multi-RAT scenario, considering TDMA and WCDMA access networks along with generic voice and data services, this chapter presents RAT-specific analytical expressions to compute the congestion probability in the aforementioned RATs which is detailed in section 4.5. These congestion-related expressions are formulated considering well-known (see, for example, [12] and [13]) TDMA and WCDMA expressions.

4.2 General Framework For Multi-Service Allocation in Multi-Access Networks

The problem of resource allocation in multi-access systems may be approached by means of multi-dimensional Markov models where each dimension corresponds to the allocation of a particular service type into a given Radio Access Technology (RAT). In particular, the allocation framework presented in the following is based on previous work in Chapters 2 and 3; to which the interested reader can refer to for model validation and further details. The model definition involves, in the first place, the identification of the state space followed by the definition of the state transition rates which will eventually lead to the steady state balance equations provided at the end of this section. Given that the focus of this chapter is to evaluate and determine the initial RAT selection policy impact on the radio access congestion, considering the static case (i.e. no mobility) seems to be adequate for such evaluation while maintaining the model complexity at a tractable level. In addition, it will be assumed that served users will not suffer from call/session disruptions due to the poor quality or link failures in the selected RAT.

4.2.1 Defining the Markov State Space

Let us consider a number of J different traffic classes. A total of K co-sited Radio Access Technologies (RATs) are deployed, each one of them to support either all J traffic classes or a subset of them. So as to account for RATs that do not uphold particular traffic classes, a $K{\times}J$ compatibility matrix, denoted as $\mathbf{B}$, may be defined with elements b_{kj}=1 if RAT k supports traffic type j and b_{kj}=0 otherwise.

Based on matrix $\mathbf{B}$, the number of supported services by a given RAT k, J_k, can be computed as $J_k = \sum_{j=1}^{J} b_{kj}$. Therefore the Markov state dimension, M, that accounts for the allocation of each supported service into each RAT may be computed as $M = \sum_{k=1}^{K} J_k$.

We may now define the row vector

$$\boldsymbol{N}_k = [N_{k,1}, N_{k,2}, \ldots, N_{k,l}, \ldots, N_{k,J_k}] \in \mathbb{Z}_+^{J_k} \,, \tag{4.1}$$

with elements $N_{k,l}$ denoting the number of allocated users in RAT k with supported service l. Note that index l, with $l = 1, 2, \ldots, J_k$, corresponds to the l-*th* supported service in RAT k, while j is the available service index. For convenience, let us define a *mapping* function g_k that returns the available traffic class index j given the l-*th* supported traffic class in RAT k, i.e. $j = g_k(l)$. Furthermore, the inverse mapping function is also defined, $l = g_k^{-1}(j)$, providing the l-*th* supported traffic class in RAT k given available traffic class index j. It is worthwhile noticing that if a RAT k supports all service types, then $l = j$.

Taking into account the number of available RATs, the number of users of each supported service in each RAT may be defined as a row vector

$$\boldsymbol{N} = [\boldsymbol{N}_1, \boldsymbol{N}_2, \ldots, \boldsymbol{N}_k, \ldots, \boldsymbol{N}_K] \in \mathbb{Z}_+^{M} \,, \tag{4.2}$$

where $\boldsymbol{N}$ may be used as the index to uniquely define each state, hereon denoted as $S_{\boldsymbol{N}}$, in the Markov chain model.

Assuming the capacity of a particular RAT k, defined as the maximum allowable number of users of each service type it may handle, is upper-bounded (i.e. *hard capacity* is assumed); a finite number of states $S_{\boldsymbol{N}}$ exists. This limit is usually set by RAT-specific Call Admission Control (CAC) procedures that determine if a new user should be admitted or not in the system so that minimum QoS requirements of already accepted users are guaranteed. In terms of the number of states, we may define the set of feasible states in RAT k, $\mathcal{S}^k$, as [similar to (3.3)]

$$\mathcal{S}^k = \left\{ S_{\boldsymbol{N}} : 0 \le f_{\boldsymbol{N}_k}^{k} \le 1 \right\} \,, \tag{4.3}$$

where $f^k_{\boldsymbol{N}_k}$ is defined as the *feasibility condition* which accounts for the CAC procedures in RAT k by defining a given state $S_{\boldsymbol{N}}$ as *feasible* in RAT k provided $0 \leq f^k_{\boldsymbol{N}_k} \leq 1$.

Finally, a given state $S_{\boldsymbol{N}}$ is said to be feasible if it satisfies $S_{\boldsymbol{N}} \in \mathcal{S}$ with $\mathcal{S} = \bigcap_{k=1}^{K} \mathcal{S}^k$, i.e. if it is feasible in all RATs.

4.2.2 Defining State Transitions

Transitions between states $S_{\boldsymbol{N}} \in \mathcal{S}$ in the resulting M-dimensional Markov chain happen due to service arrival rates, i.e. $\boldsymbol{\lambda} = [\lambda_1, \lambda_2, \ldots, \lambda_j, \ldots \lambda_J]$, or due to service departure rates $\boldsymbol{\mu} = [\mu_1, \mu_2, \ldots, \mu_j, \ldots, \mu_J]$. As widely assumed in the literature, see e.g. [14–18], arrival rates λ_j are Poisson-distributed and service times follow an exponential distribution with mean service time $1/\mu_j$ [19]. Since not all services may be supported by all RATs we define the supported arrival rates into RAT k, $\boldsymbol{\lambda}_k$, as

$$\boldsymbol{\lambda}_k = [\lambda_{k,1}, \lambda_{k,2}, \ldots, \lambda_{k,l}, \ldots, \lambda_{k,J_k}] \in \mathbb{R}_+^{J_k}\,, \tag{4.4}$$

with $\lambda_{k,l}$ the arrival rate of the *l-th* supported service type in RAT k. Note that $\boldsymbol{\lambda}_k$ is a subset of $\boldsymbol{\lambda}$ determined by compatibility matrix $\mathbf{B}$. Finally, the supported arrival rates of each traffic class into each RAT is captured by row vector $\boldsymbol{\lambda}_u = [\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2, \ldots, \boldsymbol{\lambda}_k, \ldots, \boldsymbol{\lambda}_K] \in \mathbb{R}_+^M$.

A particular traffic allocation policy, referred to as $\pi_{\boldsymbol{N}}$, is then responsible of determining, in each state $S_{\boldsymbol{N}} \in \mathcal{S}$, the specific transition arrival rates of each service type into each of the available RATs, $\boldsymbol{\lambda}_\pi$, thus defining the following function:

$$\begin{array}{rccc} \pi_{\boldsymbol{N}}: & \mathbb{R}_+^J & \longrightarrow & \mathbb{R}_+^M \\ & \boldsymbol{\lambda} & \longrightarrow & \boldsymbol{\lambda}_\pi \end{array}\,, \tag{4.5}$$

where vector $\boldsymbol{\lambda}_\pi$ contains elements $\lambda_{(\pi,k,l)}$ denoting the transition arrival rate of supported service l into RAT k due to policy $\pi_{\boldsymbol{N}}$ in state $S_{\boldsymbol{N}}$.

A specific policy π can be implemented by means of a *policy decision function*, $\boldsymbol{\Theta}_{\boldsymbol{N}} \in \mathbb{R}^M$, with elements $\Theta_{(\boldsymbol{N},k,l)} \in [0,1]$ (called *policy actions*) determining the fraction of supported traffic l into RAT k in a state $S_{\boldsymbol{N}}$, i.e.

$$\lambda_{(\pi,k,l)} = \Theta_{(\boldsymbol{N},k,l)} \lambda_{k,l}\,. \tag{4.6}$$

4.2.3 Defining the Steady State Balance Equations (SSBEs)

In equilibrium, the SSBE for state $S_{\boldsymbol{N}} \in \mathcal{S}$ results from equaling the inflow rate to state $S_{\boldsymbol{N}}$ to the outflow rate of state $S_{\boldsymbol{N}}$ [19]:

$$\begin{gathered} P_{\boldsymbol{N}} \left[\sum_{k=1}^{K} \sum_{l=1}^{J_k} \lambda_{k,l} \Theta_{(\boldsymbol{N},k,l)} \delta_{(\boldsymbol{N}+\boldsymbol{a}_{k,l})} + N_{k,l} \mu_{k,l} \delta_{(\boldsymbol{N}-\boldsymbol{a}_{k,l})} \right] = \\ \sum_{k=1}^{K} \sum_{l=1}^{J_k} \lambda_{k,l} \Theta_{(\boldsymbol{N}-\boldsymbol{a}_{k,l},k,l)} P_{(\boldsymbol{N}-\boldsymbol{a}_{k,l})} \delta_{(\boldsymbol{N}-\boldsymbol{a}_{k,l})} + (N_{k,l}+1) \mu_{k,l} P_{(\boldsymbol{N}+\boldsymbol{a}_{k,l})} \delta_{(\boldsymbol{N}+\boldsymbol{a}_{k,l})} \end{gathered}, \tag{4.7}$$

where $P_{\boldsymbol{N}}$ is the probability of being in state $S_{\boldsymbol{N}}$, and $\boldsymbol{a}_{k,l} \in \mathbb{Z}_{+}^{M}$ is a row vector containing all zeros except for the *l-th* supported service in RAT k element which is 1. In addition, $\delta_{\boldsymbol{N}}$ is an indicator function which will guarantee that non-feasible states, i.e. $S_{\boldsymbol{N}} \notin \mathcal{S}$, are not taken into account; thus, $\delta_{(\boldsymbol{N})} = 1$ if $S_{\boldsymbol{N}} \in \mathcal{S}$ and $\delta_{(\boldsymbol{N})} = 0$ otherwise.

Once the SSBEs are determined for all states $S_{\boldsymbol{N}} \in \mathcal{S}$, numerical methods may be applied to solve the resulting system of equations given by the SSBEs plus the normalization constraint $\sum_{S_{\boldsymbol{N}} \in S} P_{\boldsymbol{N}} = 1$.

The interested reader is referred to, e.g., [20] and [21] for further details on the numerical solution of Markov chains.

4.3 Congestion Probability Model

The Markov model proposed in the previous section provides, given total traffic arrival rates $\boldsymbol{\lambda}$ and a given RAT selection policy π, the probability, $P_{\boldsymbol{N}}$, of having a particular number of admitted users denoted by vector $\boldsymbol{N}$. Then, the congestion probability in each RAT k for a given RAT selection policy π may be formulated as follows:

$$P_c^{k,\pi} = \sum\nolimits_{S_{\boldsymbol{N}} \in \mathcal{S}} P_c^k(\boldsymbol{N}_k) \cdot P_{\boldsymbol{N}} \,, \tag{4.8}$$

where $P_c^k(\boldsymbol{N}_k)$, the congestion probability in RAT k given $\boldsymbol{N}_k$ allocated users, is averaged over all state probabilities $P_{\boldsymbol{N}}$ in the state space $\mathcal{S}$. A detailed explanation on how $P_c^k(\boldsymbol{N}_k)$ is computed will be provided in section 4.5.2 for the case of TDMA and WCDMA technologies. At this point it should be noted that $P_c^k(\boldsymbol{N}_k)$ highly depends on the underlying access scheme of the considered RAT k, which may lead to different definitions of radio access congestion and thus, different definitions of congestion probability. In this sense, a TDMA based scheme can undergo congestion situations due to, e.g., excessive timeslot reuse. On the contrary, a RAT based on WCDMA may be congested if, e.g., excessive interference is detected, hence degrading the system performance. Moreover, it can be foreseen that the higher the number of allocated users, $\boldsymbol{N}_k$, the higher the congestion probability; nevertheless, users demanding different services may have different impacts on the perceived congestion as it will be shown in section 4.5.2.

In general, the congestion probability in RAT k given $\boldsymbol{N}_k$ allocated users, $P_c^k(\boldsymbol{N}_k)$, may arise independently due to congestion measured in the UL and/or due to congestion measured in the DL. Consequently one can express this probability as

$$P_c^k(\boldsymbol{N}_k) = 1 - [1 - P_c^{\mathrm{UL},k}(\boldsymbol{N}_k)][1 - P_c^{\mathrm{DL},k}(\boldsymbol{N}_k)] \,, \tag{4.9}$$

where $P_c^{\mathrm{UL},k}(\boldsymbol{N}_k)$ and $P_c^{\mathrm{DL},k}(\boldsymbol{N}_k)$ indicate UL and DL congestion probability respectively.

Provided the congestion probability in RAT k given policy π in (4.8), the overall congestion probability in the multi-RAT scenario, i.e. the probability that there exists congestion in at least one RAT, P_c^π, can be then written as

$$P_c^\pi = 1 - \prod\nolimits_{k=1}^{K} (1 - P_c^{k,\pi}) \,, \tag{4.10}$$

where it is assumed that congestion probabilities arising in each RAT are independent.

4.4 Defining RAT Selection Policies

With previous definitions, a number of RAT selection policies can be characterized by means of appropriately defining policy decision function $\boldsymbol{\Theta}_{\boldsymbol{N}}$ through its elements $\Theta_{(\boldsymbol{N},k,l)}$. The outcome of a particular RAT selection policy π for the allocation of a j-service type user at a given state $S_{\boldsymbol{N}} \in \mathcal{S}$ can be represented by the set $\mathcal{R}_j^\pi$ containing the selected RAT(s). Note that elements in $\mathcal{R}_j^\pi$ are those RAT(s) selected by the policy provided service class j is supported and that the allocation of such service is possible, i.e. admission control is successful. Mathematically, it is required that $\mathcal{R}_j^\pi \in \mathcal{A}_j$ with

$$\mathcal{A}_j = \left\{ k \in [1, K] : b_{kj} \neq 0, S_{(\boldsymbol{N}_k + \boldsymbol{e}_{k,l})} \in \mathcal{S} \right\} \,, \tag{4.11}$$

where $\boldsymbol{e}_{k,l}\in\mathbb{Z}^{J_k}$ is a vector containing all zeros except for the *l-th* component, which is 1. Moreover, refer to section 4.2.1, $b_{kj}\neq 0$ indicates that service j is supported by RAT k.

Note that, if $|\mathcal{R}_j^{\pi}|$ denotes the cardinality of $\mathcal{R}_j^{\pi}$ (i.e. the number of elements in $\mathcal{R}_j^{\pi}$), then in the singular case that set $\mathcal{R}_j^{\pi}$ contains more than one selected RAT, i.e. $|\mathcal{R}_j^{\pi}| > 1$, it is assumed that a RAT is chosen among those according to some preestablished prioritization scheme, particular to each policy. In addition, it should be noticed that the RAT selection policy definition given by (4.11) implicitly states that, if the *preferred* RAT according to a given policy cannot admit further service class j users, then alternative RATs are sequentially selected according to a prioritized order determined by the policy.

Subsequently, some specific RAT selection policies are presented and defined in terms of $\mathcal{R}_j^{\pi}$. Such RAT selection policies are suggested and justified by the authors in Chapter 2, to which the interested reader is referred for further details.

4.4.1 Load-Balancing (LB) RAT Selection Policy

This policy allocates a given user demanding a particular traffic class to the RAT that undergoes a lower load status. Naturally, this implies defining appropriate load metrics for each of the considered RATs. Let the load of a RAT k at a given state $S_{\boldsymbol{N}}\in\mathcal{S}$ be denoted as $L^k_{\boldsymbol{N}_k}$, where it explicitly states that the load value in RAT k will depend, among other RAT-specific parameters, on the number of users of each service class in RAT k, i.e. $\boldsymbol{N}_k$.

At a specified state $S_{\boldsymbol{N}}\in\mathcal{S}$, a service class j user will be allocated to the RAT k that, including this new service class j user, exhibits a lower load. Accordingly, we may define

$$\mathcal{R}_j^{LB} = \underset{k\in\mathcal{A}_j}{\arg\min}\left\{L^k_{(\boldsymbol{N}_k+\boldsymbol{e}_{k,l})}\right\}\,, \tag{4.12}$$

where elements in $\mathcal{R}_j^{LB}$ are those RAT(s) with lowest load provided service class j is supported and that the allocation of supported service class l is possible.

Then, LB policy decision towards supported traffic l into RAT k in state $S_{\boldsymbol{N}}$ can be defined in terms of $\mathcal{R}_j^{LB}$ as

$$\Theta_{(\boldsymbol{N},k,l)} = \begin{cases} \frac{1}{|\mathcal{R}_j^{LB}|} & \text{if } k\in\mathcal{R}_j^{LB} \text{ with } j=g_k(l) \\ 0 & \text{otherwise} \end{cases}\,, \tag{4.13}$$

where it is implicitly assumed that, in the singular case of $|\mathcal{R}_j^{LB}|>1$, i.e. the minimum load is achieved in more than one RAT, selection is done randomly among those RATs, with probability $1/|\mathcal{R}_j^{LB}|$.

4.4.2 Service-Based (SB) RAT Selection Policy

This policy selects a particular RAT k is based on the demanding user service type j according to some predefined preference scheme. Consider that for each available service type j we may build a row vector $\boldsymbol{s}_j = \left[s_{(1,j)}, s_{(2,j)}, \ldots, s_{(k,j)}, \ldots, s_{(K,j)}\right] \in \mathbb{Z}^K$. Each element $s_{(k,j)}$ in this vector takes a non-negative integer value indicating the preference of service j on RAT k. The higher the value of $s_{(k,j)}$ the higher the priority.

At a specified state $S_{\boldsymbol{N}} \in \mathcal{S}$, a service class j user will be allocated to the RAT k according to

$$\mathcal{R}_j^{SB} = \underset{k \in \mathcal{A}_j}{\arg\max} \left\{ s_{(k,j)} \right\} \,, \tag{4.14}$$

where $\mathcal{R}_j^{SB}$ contains the RAT with highest priority value provided it is capable of supporting and admitting the new service class j user. It is worthwhile mentioning that, for this particular policy, we will always have $|\mathcal{R}_j^{SB}| = 1$, i.e. the policy outcome is always one RAT.

In this case, the SB policy decision with respect to supported traffic l into RAT k at a given state $S_{\boldsymbol{N}}$ can be defined in terms of $\mathcal{R}_j^{SB}$ as

$$\Theta_{(\boldsymbol{N},k,l)} = \begin{cases} 1 & \text{if } k \in \mathcal{R}_j^{SB} \text{ with } j = g_k(l) \\ 0 & \text{otherwise} \end{cases} \,. \tag{4.15}$$

4.4.3 Congestion-Aware (CA) RAT Selection Policy

The statistical model for congestion probability presented in section 4.3 can be used as a valuable input for deciding the most appropriate RAT for a specific service request. In this way, a particular user may be allocated to the RAT that will be less likely to fall in a high congestion state. In this case, the set of selected RAT(s) are

$$\mathcal{R}_j^{CA} = \underset{k \in \mathcal{A}_j}{\arg\min} \left\{ P_c^k(\boldsymbol{N}_k + \boldsymbol{e}_{k,l}) \right\} \,, \tag{4.16}$$

where the RAT(s) with lowest congestion probability is selected provided service class j is supported and that the allocation of supported service class l is possible.

Similar to the LB case in (4.13), CA policy decision with respect to supported traffic l into RAT k at a given state $S_{\boldsymbol{N}}$ can be defined in terms of $\mathcal{R}_j^{CA}$ as

$$\Theta_{(\boldsymbol{N},k,l)} = \begin{cases} \frac{1}{|\mathcal{R}_j^{CA}|} & \text{if } k \in \mathcal{R}_j^{CA} \text{ with } j = g_k(l) \\ 0 & \text{otherwise} \end{cases} \,, \tag{4.17}$$

where it is also assumed that, in the particular case of $|\mathcal{R}_j^{CA}| > 1$, i.e. the minimum congestion probability is achieved in more than one RAT, the RAT is selected randomly among those, with probability $1/|\mathcal{R}_j^{CA}|$.

According to the above definitions, policy decisions in the SSBEs will be performed using (4.13), (4.15) and (4.17) into (4.7), so as to account for the different user allocation strategies.

4.5 Case Study: Voice and Data Services in a TDMA/WCDMA Scenario

In this section, the state space and the congestion probabilities assuming two generic service types, voice and data, and two RATs, TDMA and WCDMA, are derived. Both RATs are considered to support both services in a single-cell scenario, thus a 4-Dimensional ($M = 4$) Markov model arises. In the remainder, as in Chapter 2, let v and d represent voice and data indexes along with t and w represent TDMA and WCDMA indexes.

4.5.1 Markov State Space

A boundary on the number of states comprised by the Markov chain is set by means of imposing appropriate Call Admission Control (CAC) mechanisms in each RAT.

4.5.1.1 TDMA Case

For Time Division Multiple Access (TDMA) based systems, a total amount of C channels (or timeslots, TSL) are to be shared among voice and data users within a time frame. It is further assumed that no service has priority over any other. Typically, but not necessarily, voice users occupy a whole TSL throughout the duration of a call, which will be assumed in this chapter. As for data users, a given TSL may be shared by up to n_C data transmissions corresponding to different users by means of efficient time scheduling. Moreover, a given data user may be granted several TSLs in order to increase its achievable bit rate. This feature is commonly referred to as multi-slot capability, and constitutes a differentiating element from

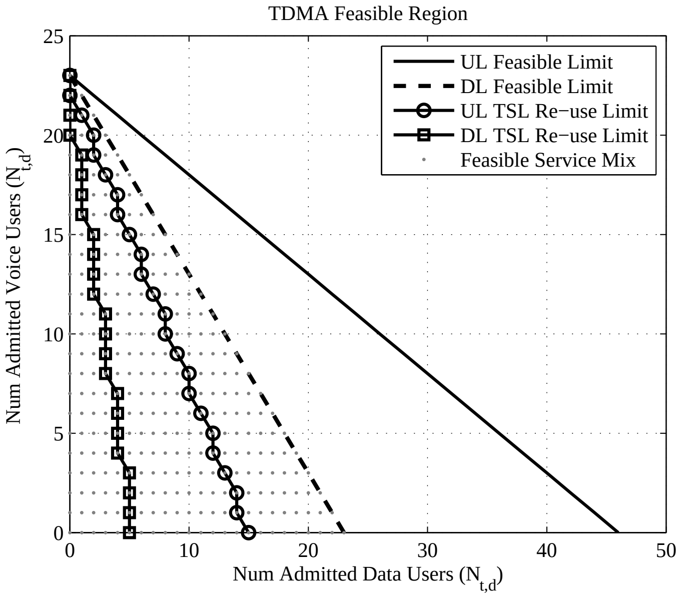


Figure 4.1: TDMA feasible region for $C = 23$, $n_C^{\mathrm{UL}} = 3$, $n_C^{\mathrm{DL}} = 4$, $q^{\mathrm{UL}} = 3$, $q^{\mathrm{DL}} = 8$ and $\alpha_d^{\mathrm{UL}} = \alpha_d^{\mathrm{DL}} = 0.5$.

Chapter 2, where a single TSL was allocated to data users. In order to capture this multi-slot capability, it is assumed that the number of allocated TSLs to a particular data user is q.

The CAC mechanism may be then expressed in mathematical terms considering the total amount of resources, i.e. C TSLs, to be shared among the simultaneous voice and data users in the system. This will define the set of feasible number of admitted voice and data users in the UL and DL, $\mathcal{S}^{\mathrm{UL},t}$ and $\mathcal{S}^{\mathrm{DL},t}$, as:

$$\begin{aligned} \mathcal{S}^{\mathrm{UL},t} &= \left\{ \boldsymbol{N}_t : 0 \leq \frac{N_{t,v}}{C} + \frac{N_{t,d}\alpha_d^{\mathrm{UL}} q^{\mathrm{UL}}}{n_C^{\mathrm{UL}} C} \leq 1 \right\} \\ \mathcal{S}^{\mathrm{DL},t} &= \left\{ \boldsymbol{N}_t : 0 \leq \frac{N_{t,v}}{C} + \frac{N_{t,d}\alpha_d^{\mathrm{DL}} q^{\mathrm{DL}}}{n_C^{\mathrm{DL}} C} \leq 1 \right\}, \end{aligned} \tag{4.18}$$

where $N_{t,v}$ ($N_{t,d}$) is the number of admitted voice (data) users in TDMA and α_d^{UL} (α_d^{DL}) is the UL (DL) activity factor of data users. In addition, the number of allocated TSLs to a particular data user in the UL (DL) is represented by q^{UL} (q^{DL}), and the maximum number of users sharing a same TSL in the UL (DL) is denoted by n_C^{UL} (n_C^{DL}). The set of feasible number of users satisfying both conditions in (4.18), i.e. $\mathcal{S}^t = \mathcal{S}^{\mathrm{UL},t} \cap \mathcal{S}^{\mathrm{DL},t}$, constitutes the so-called TDMA admission region. In Fig.4.1 the admission limits imposed by the UL and DL are shown along with the limits above which TSL reuse exists in the UL and DL directions.

4.5.1.2 WCDMA Case

As for WCDMA based systems, CAC procedures also comprise both UL and DL mechanisms. In WCDMA, the *uplink load factor*, η^{UL}, can be used to determine if a new user should be admitted in the system by setting a maximum value, η^{UL}_{max}, which cannot be exceeded. Similar to what was shown for TDMA, an UL WCDMA feasibility set can be defined as

$$\mathcal{S}^{\text{UL},w} = \left\{ \boldsymbol{N}_w : 0 \leq \eta^{\text{UL}}_{\boldsymbol{N}_w} \leq \eta^{\text{UL}}_{max} \right\} , \tag{4.19}$$

where the UL load factor in WCDMA is defined as [13]:

$$\eta^{\text{UL}}_{\boldsymbol{N}_w} = \sum\nolimits_{i=1}^{N_{w,v}} \alpha^{\text{UL}}_v / A^{\text{UL}}_{v,i} + \sum\nolimits_{i=1}^{N_{w,d}} \alpha^{\text{UL}}_d / A^{\text{UL}}_{d,i} \, , \tag{4.20}$$

with

$$A^{\text{UL}}_{v,i} = \frac{W/R^{\text{UL}}_{bv,i}}{\theta^{\text{UL}}_{v,i}} + 1 \quad \text{and} \quad A^{\text{UL}}_{d,i} = \frac{W/R^{\text{UL}}_{bd,i}}{\theta^{\text{UL}}_{d,i}} + 1 \, , \tag{4.21}$$

where W is the chip rate; $R^{\text{UL}}_{bv,i}$ ($R^{\text{UL}}_{bd,i}$) is the UL bit-rate granted to voice (data) users; and $\theta^{\text{UL}}_{v,i}$ ($\theta^{\text{UL}}_{d,i}$) is the UL target bit-energy-to-noise-density ratio after despreading and decoding for voice (data) users. Additionally, α^{UL}_v is the UL activity factor for voice users.

In the DL, it is assumed a CAC algorithm based on the availability of Orthogonal Variable Spreading Factor (OVSF) codes as, e.g., in [13]. Then, assuming a same OVSF code can be shared among different data users due to inactivity periods, the number of dedicated channels will have to fulfill the relationship [13]

$$\mathcal{C}_{\boldsymbol{N}_w} = \sum\nolimits_{i=1}^{N_{w,v}} \frac{1}{SF^{\text{DL}}_{v,i}} + \sum\nolimits_{i=1}^{N_{w,d}} \alpha^{\text{DL}}_d \frac{1}{SF^{\text{DL}}_{d,i}} \leq \mathcal{C}_{max} \, , \tag{4.22}$$

with $SF^{\text{DL}}_{v,i}$ ($SF^{\text{DL}}_{d,i}$) the DL spreading factor for voice (data) users, and $0 \leq \mathcal{C}_{max} \leq 1$. Then, the feasible set of voice and data users in WCDMA DL is expressed as

$$\mathcal{S}^{\text{DL},w} = \{ \boldsymbol{N}_w : 0 \leq \mathcal{C}_{\boldsymbol{N}_w} \leq \mathcal{C}_{max} \} \, . \tag{4.23}$$

The resulting WCDMA total feasible set, $\mathcal{S}^w$, must satisfy both UL and DL feasibility conditions defined in (4.19) and (4.23), i.e. $\mathcal{S}^w = \mathcal{S}^{\text{UL},w} \cap \mathcal{S}^{\text{DL},w}$, which results in the admission region depicted by the dotted area in Fig. 4.2.

Finally, the global Markov state feasibility space, $\mathcal{S}$, can be defined as, from the feasible sets in each RAT as:

$$\mathcal{S} = \left\{ \boldsymbol{N} : \boldsymbol{N}_t \in \mathcal{S}^t, \boldsymbol{N}_w \in \mathcal{S}^w \right\} \, . \tag{4.24}$$

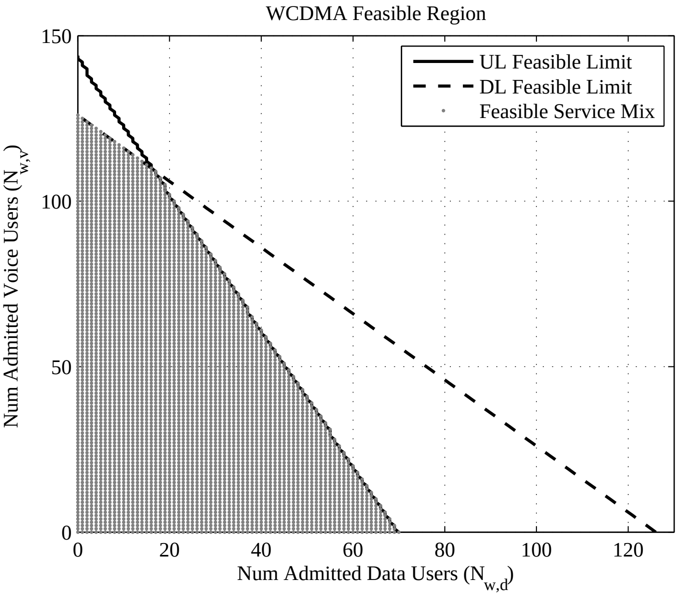


Figure 4.2: WCDMA feasible region for $W = 3.84$ Mcps, $R^{\mathrm{UL}}_{bv,i} = 12.2$ kbps, $R^{\mathrm{UL}}_{bd,i} = 32$ kbps, $\theta^{\mathrm{UL}}_{v,i} = 6$ dB, $\theta^{\mathrm{UL}}_{d,i} = 5$ dB, $\alpha^{\mathrm{UL}}_{v} = \alpha^{\mathrm{UL}}_{d} = \alpha^{\mathrm{DL}}_{d} = 0.5$, $\eta^{\mathrm{UL}}_{max} = 0.9$, $\mathcal{C}_{max} = 63/64$, $SF^{\mathrm{DL}}_{v,i} = 128$ and $SF^{\mathrm{DL}}_{d,i} = 64$.

4.5.2 Congestion Probability Cases

In the following, the congestion probability is derived for RATs WCDMA and TDMA.

4.5.2.1 WCDMA Congestion Probability

In WCDMA, radio congestion events may arise in the UL and/or in the DL, thus enabling the congestion probability in WCDMA to be formulated as

$$P^{w}_{c}(\boldsymbol{N}_w) = 1 - [1 - P^{UL,w}_{c}(\boldsymbol{N}_w)][1 - P^{DL,w}_{c}(\boldsymbol{N}_w)]\,, \tag{4.25}$$

where the assumption that the congestion probability in the UL and DL are independent has been made.

4.5.2.1.1 Uplink Case It is well known, see e.g. [13], that user traffic activity, characterized by an activity factor α, may be modeled by a binomial distribution in order to determine the probability of having n simultaneously transmitting users given N admitted users:

$$P_\alpha(n|N) = \binom{N}{n} \alpha^n \left(1-\alpha\right)^{N-n} \; . \tag{4.26}$$

Then, assuming independence between voice and data call/session generation, the probability of $n_{w,v}$ voice users and $n_{w,d}$ data users simultaneously transmitting in WCDMA when $N_{w,v}$ and $N_{w,d}$ voice and data users are admitted in the system, $P_{\boldsymbol{\alpha}^{\mathrm{UL}}}(\boldsymbol{n}_w|\boldsymbol{N}_w)$, can be formulated as

$$P_{\boldsymbol{\alpha}^{\mathrm{UL}}}(\boldsymbol{n}_w|\boldsymbol{N}_w) = P_{\alpha_v^{\mathrm{UL}}}(n_{w,v}|N_{w,v}) \cdot P_{\alpha_d^{\mathrm{UL}}}(n_{w,d}|N_{w,d}) \, , \tag{4.27}$$

where vector $\boldsymbol{\alpha}^{\mathrm{UL}} = (\alpha_v^{\mathrm{UL}}, \alpha_d^{\mathrm{UL}})$ denotes UL voice and data activity factors.

In WCDMA, congestion situations may be detected in the uplink by means of the UL load factor η^{UL} whenever its value exceeds a given threshold η_c^{UL} during a certain percentage of frames within a period of time [2]. Thus, we may define the WCDMA UL congestion probability, $P_c^{\mathrm{UL},w}(\boldsymbol{N}_w)$, as

$$P_c^{\mathrm{UL},w}(\boldsymbol{N}_w) = \sum_{n_{w,v}=0}^{N_{w,v}} \sum_{n_{w,d}=0}^{N_{w,d}} \Pr\left(\eta_{(\boldsymbol{n}_w)}^{\mathrm{UL}} > \eta_c^{\mathrm{UL}}\right) \cdot P_{\boldsymbol{\alpha}^{\mathrm{UL}}}(\boldsymbol{n}_w|\boldsymbol{N}_w), \tag{4.28}$$

where, we can express the UL load factor in a single-cell scenario when $n_{w,v}$ simultaneous voice users and $n_{w,d}$ simultaneous data users are in the system, $\eta_{(\boldsymbol{n}_w)}^{\mathrm{UL}}$, as [13]:

$$\eta_{(\boldsymbol{n}_w)}^{\mathrm{UL}} = \sum\nolimits_{i=1}^{n_{w,v}} 1/A_{v,i}^{\mathrm{UL}} + \sum\nolimits_{i=1}^{n_{w,d}} 1/A_{d,i}^{\mathrm{UL}} \, , \tag{4.29}$$

with $A_{v,i}^{\mathrm{UL}}$ and $A_{d,i}^{\mathrm{UL}}$ defined in (4.21).

Consequently, and assuming users of same service type to have equal allocated bit rates and bit energy to noise density ratio requirements, i.e. $A_{v,i}^{\mathrm{UL}} = A_v^{\mathrm{UL}}$ and $A_{d,i}^{\mathrm{UL}} = A_d^{\mathrm{UL}}$ $\forall i$, it may be written

$$\Pr\left(\eta_{(\boldsymbol{n}_w)}^{\mathrm{UL}} > \eta_c^{\mathrm{UL}}\right) = \left\{ \begin{array}{ll} 0 & \text{if} \quad \boldsymbol{n}_w \cdot (\boldsymbol{A}^{\mathrm{UL}})^T \leq \eta_c \\ 1 & \text{if} \quad \boldsymbol{n}_w \cdot (\boldsymbol{A}^{\mathrm{UL}})^T > \eta_c \end{array} \right. \; , \tag{4.30}$$

where it has been defined $\boldsymbol{A}^{\mathrm{UL}} = (A_v^{\mathrm{UL}^{-1}}, A_d^{\mathrm{UL}^{-1}})$.

Given the WCDMA system parameters provided in Table 4.1, the congestion probability in the UL with respect to the number of admitted voice and data users defined in (4.28) is illustrated in Fig. 4.3, where congestion probabilities are represented in the form of a gray-scale with the color-bar indicating probability values on the right of the graph.

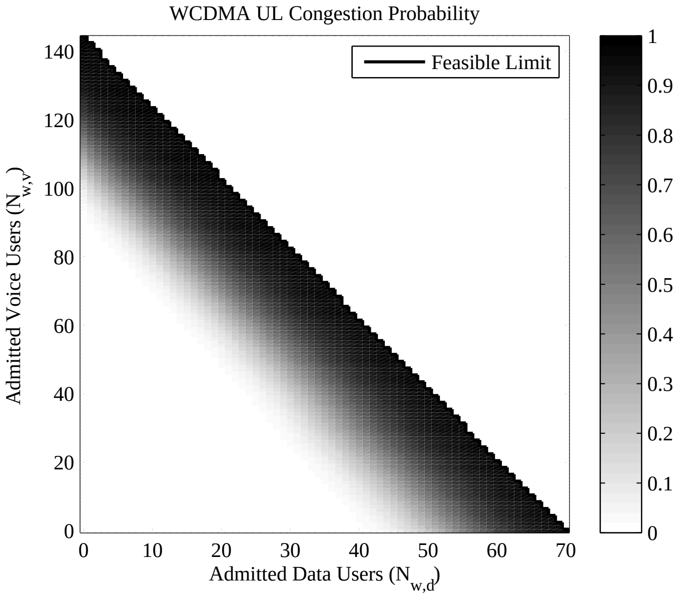


Figure 4.3: WCDMA UL Congestion Probability.

Table 4.1: WCDMA System Parameters For Numerical Evaluation of UL Congestion Probability.

Parameter	Symbol	Value
Chip rate	W	3.84 Mcps
Load factor congestion threshold	η_c^{UL}	0.7
Maximum load factor	η_{max}^{UL}	0.9
E_b/N_0[a] for voice (data) traffic	θ_v^{UL} (θ_d^{UL})	6 (5) dB
Bit-rate for voice (data) traffic	$R_{b,v}^{\mathrm{UL}}$ ($R_{b,d}^{\mathrm{UL}}$)	12.2 (32) kbps
Voice (data) activity factor	α_v^{UL} (α_d^{UL})	0.5 (0.5)

[a] Bit energy to noise density ratio.

4.5.2.1.2 Downlink Case In WCDMA systems, the total DL transmitted power needed to satisfy all $n_{w,v}$ and $n_{w,d}$ simultaneous voice and data users in a single-cell scenario is [13]

$$P_T^{\mathrm{DL}}(\boldsymbol{n_w}) = \frac{P_p + P_N\left(\sum_{i=1}^{n_{w,v}} \frac{L_{p,i}}{A_{v,i}^{\mathrm{DL}}+\rho} + \sum_{i=1}^{n_{w,d}} \frac{L_{p,i}}{A_{d,i}^{\mathrm{DL}}+\rho}\right)}{1-\eta_{(\boldsymbol{n}_w)}^{\mathrm{DL}}} \leq P_{T,\max}^{\mathrm{DL}}, \tag{4.31}$$

with

$$A_{v,i}^{\mathrm{DL}} = \frac{W/R_{bv,i}^{\mathrm{DL}}}{\theta_{v,i}^{\mathrm{DL}}} \quad \text{and} \quad A_{d,i}^{\mathrm{DL}} = \frac{W/R_{bd,i}^{\mathrm{DL}}}{\theta_{d,i}^{\mathrm{DL}}}, \tag{4.32}$$

along with

$$\eta^{\mathrm{DL}}_{(\boldsymbol{n}_w)} = \sum\nolimits_{i=1}^{n_{w,v}} \frac{\rho}{A^{\mathrm{DL}}_{v,i} + \rho} + \sum\nolimits_{i=1}^{n_{w,d}} \frac{\rho}{A^{\mathrm{DL}}_{d,i} + \rho}\,, \tag{4.33}$$

where P_p and P_N are the pilot and thermal noise powers respectively, ρ is the DL orthogonality factor, and $P^{\mathrm{DL}}_{T,\max}$ is the maximum total DL power available at the base station. W is the WCDMA chip-rate, and user requirements are in the form of requested bit rates for voice and data services, $R^{\mathrm{DL}}_{bv,i}$ and $R^{\mathrm{DL}}_{bd,i}$, along with target bit-energy-to-noise-density ratios for voice and data services, $\theta^{\mathrm{DL}}_{v,i}$ and $\theta^{\mathrm{DL}}_{d,i}$. The path-loss experienced by each user, $L_{p,i}$, may be characterized by considering macro or micro-cell environments with shadowing effects modeled by means of log-normal variation [13].

Consequently, for $n_{w,v}$ and $n_{w,d}$ simultaneous users, a congestion situation may be detected whenever the total DL power exceeds a given power threshold $P^{\mathrm{DL}}_{T,c} < P^{\mathrm{DL}}_{T,\max}$. Then, we may write the congestion probability in the DL as shown in (4.34), which depends on the path-loss distribution and consequently on the user′s geographical location.

$$\begin{aligned} \Pr(P^{\mathrm{DL}}_{T}(\boldsymbol{n_w}) > P^{\mathrm{DL}}_{T,c}) \;&=\; \Pr\left\{ \left(\sum_{i=1}^{n_{w,v}} \frac{L_{p,i}}{A^{\mathrm{DL}}_{v,i}+\rho} + \sum_{i=1}^{n_{w,d}} \frac{L_{p,i}}{A^{\mathrm{DL}}_{d,i}+\rho} \right) > \frac{P^{\mathrm{DL}}_{T,c}\left(1-\eta^{\mathrm{DL}}_{(\boldsymbol{n}_w)}\right)-P_p}{P_N} \right\} \\ &\triangleq\; \Pr\left(\gamma_{\boldsymbol{n_w}} > \gamma^{*}_{\boldsymbol{n_w}}\right) \end{aligned} \tag{4.34}$$

For a given number of admitted voice and data users, the DL congestion probability due to power constraints can be written as

$$P_c^{\mathrm{DL},w}(\boldsymbol{N}_w) = \sum_{n_{w,v}=0}^{N_{w,v}} \sum_{n_{w,d}=0}^{N_{w,d}} \Pr\left(\gamma_{\boldsymbol{n}_w} > \gamma^{*}_{\boldsymbol{n}_w}\right) \cdot P_{\boldsymbol{\alpha}^{\mathrm{DL}}}(\boldsymbol{n}_w|\boldsymbol{N}_w). \tag{4.35}$$

The congestion probability in the DL as a function of the number of active voice and data users defined in (4.35) is illustrated in Fig. 4.4, where the terms $\Pr(\gamma_{\boldsymbol{n}_w} > \gamma^{*}_{\boldsymbol{n}_w})$ and $P_{\boldsymbol{\alpha}^{\mathrm{DL}}}(\boldsymbol{n}_w|\boldsymbol{N}_w)$ are computed given the WCDMA system parameters provided in Table 4.2. It can be seen that the strain in the DL congestion probability is set on the number of data users in the system due to their higher demands in terms of bit rates, which is translated into higher power demands and, consequently, higher congestion chances.

The total congestion probability in WCDMA considering the uplink and downlink, i.e. the probability of being in congestion either in the UL or in the DL, can be then computed according to (4.25) and represented as illustrated in Fig. 4.5.

Table 4.2: WCDMA System Parameters For Numerical Evaluation of DL Congestion Probability.

Parameter	Symbol	Value
Cell radius	R	1000 m
User spatial distr.	-	Homogeneous
Propagation model	-	Macrocell
Log-normal shadowing std. deviation	σ	10 dB
Min. coupling loss	MCL	70 dB
Carrier frequency	f_c	1800 Mhz
Pilot power	P_p	30 dBm
Thermal noise power	P_N	-100 dBm
Orthogonality factor	ρ	0.6
Chip rate	W	3.84 Mcps
DL power congestion threshold	$P_{T,c}^{\mathrm{DL}}$	35 dBm
Max. DL transmitted power	$P_{T,max}^{\mathrm{DL}}$	43 dBm
E_b/N_0[a] for voice (data) traffic	θ_v^{DL} (θ_d^{DL})	6 (7) dB
Bit-rate for voice (data) traffic	$R_{b,v}^{\mathrm{DL}}$ ($R_{b,d}^{\mathrm{DL}}$)	12.2 (64) kbps
Voice (data) spreading factor	SF_v^{DL} (SF_d^{DL})	128 (64)
Max. OVSF Code Condition Limit	$\mathcal{C}_{max}$	63/64[b]
Voice (data) activity factor	α_v^{DL} (α_d^{DL})	0.5 (0.5)

[a] Bit energy to noise density ratio.

[b] For the case of a UTRAN R99 WCDMA cell with one code for the CPICH channel, another for the Primary CCPCH carrying the BCH and a couple of Secondary CCPCH channels carrying the FACH and the PCH, all of them with $\check{SF} = 256$, the fraction reserved codes for the common and shared channels would be $R_C = 4\cdot(1/256) = 1/64$ thus yielding $\mathcal{C}_{max} = 1 - R_C = 63/64$ [13].

4.5.2.2 TDMA Congestion Probability

As suggested in [2, 3], a possible indicator of congestion in TDMA access technologies can be based on the effect of timeslot sharing among data users. Accordingly, the congestion probability in TDMA, given $n_{t,v}$ simultaneous voice users are assigned τ_v TSLs and $n_{t,d}$ simultaneous data users are assigned τ_d data TSL, can be

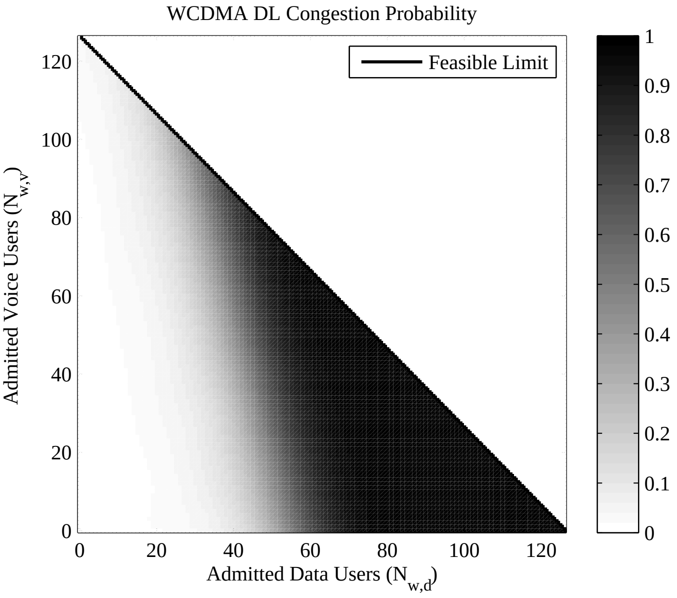


Figure 4.4: WCDMA DL Congestion Probability.

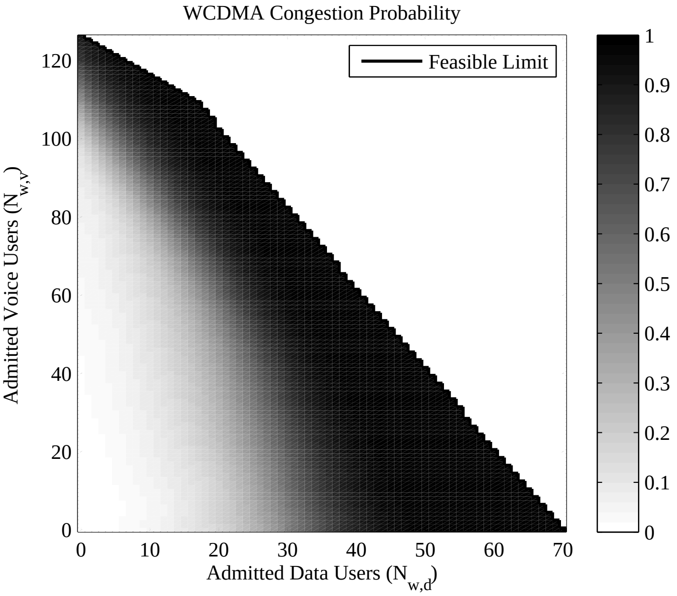


Figure 4.5: WCDMA Congestion Probability.

expressed as

$$\Pr\left(\xi_{(\tau_v,\tau_d)} < \xi_c\right) = \begin{cases} 1 & \text{if } \xi_{(\tau_v,\tau_d)} < \xi_c \\ 0 & \text{if } \xi_{(\tau_v,\tau_d)} \geq \xi_c \end{cases}, \tag{4.36}$$

where, if C is the total number of available TSLs,

$$\xi_{(\tau_v,\tau_d)} = \left\{ \begin{array}{cc} 1 & \text{if } 0 \le \tau_d \le C-\tau_v \\ \frac{C-\tau_v}{\tau_d} & \text{if } \tau_d > C-\tau_v \end{array} \right. , \tag{4.37}$$

is the *reduction factor* [12] that accounts for the effect of TSL sharing among data users in a TDMA system, like e.g. GSM/EDGE. It follows from (4.37) that ξ takes values between 0 and 1, meaning a very saturated network for ξ close to 0 (high TSL sharing), and a low loaded network for ξ close to 1 (low TSL sharing). According to (4.36), congestion is detected if the reduction factor ξ falls below a given threshold ξ_c.

Then, given a number of $N_{t,v}$ and $N_{t,d}$ admitted users in TDMA, we may compute the probability of $n_{t,v}$ voice users and $n_{t,d}$ data users simultaneously transmitting in TDMA in the same way as (4.27). The resulting simultaneous users will be assigned a number of TSLs, in both the UL and DL, given by q^{UL} and q^{DL} respectively, allowing the following expressions for congestion probability in each link to be defined as

$$P_c^{\mathrm{UL},t}(\boldsymbol{N}_t) = \sum_{n_{t,v}=0}^{N_{t,v}} \sum_{n_{t,d}=0}^{N_{t,d}} \Pr\left(\xi^{\mathrm{UL}}_{(n_{t,v},n_{t,d}q^{\mathrm{UL}})} < \xi_c^{\mathrm{UL}}\right) \cdot P_{\boldsymbol{\alpha}^{\mathrm{UL}}}(\boldsymbol{n}_t|\boldsymbol{N}_t), \tag{4.38}$$

along with

$$P_c^{\mathrm{DL},t}(\boldsymbol{N}_t) = \sum_{n_{t,v}=0}^{N_{t,v}} \sum_{n_{t,d}=0}^{N_{t,d}} \Pr\left(\xi^{\mathrm{DL}}_{(n_{t,v},n_{t,d}q^{\mathrm{DL}})} < \xi_c^{\mathrm{DL}}\right) \cdot P_{\boldsymbol{\alpha}^{\mathrm{DL}}}(\boldsymbol{n}_t|\boldsymbol{N}_t), \tag{4.39}$$

where it has been assumed that simultaneous voice users always occupy a single TSL, i.e. $\tau_v = n_{t,v}$, and simultaneous data users are assigned $\tau_d = n_{t,d}q$ TSLs. Finally, it may further be considered that voice users will occupy the entire TSL throughout the duration of the call, which basically means that the activity factor for voice users, α_v, is equal to 1, thus yielding $n_{t,v} = N_{t,v}$. In this case, (4.38) and (4.39) become

$$P_c^{\mathrm{UL},t}(\boldsymbol{N}_t) = \sum_{n_{t,d}=0}^{N_{t,d}} \Pr\left(\xi^{\mathrm{UL}}_{(N_{t,v},n_{t,d}q^{\mathrm{UL}})} < \xi_c^{\mathrm{UL}}\right) \cdot P_{\alpha_d^{\mathrm{UL}}}(n_{t,d}|N_{t,d}), \tag{4.40}$$

along with

$$P_c^{\mathrm{DL},t}(\boldsymbol{N}_t) = \sum_{n_{t,d}=0}^{N_{t,d}} \Pr\left(\xi^{\mathrm{DL}}_{(N_{t,v},n_{t,d}q^{\mathrm{DL}})} < \xi_c^{\mathrm{DL}}\right) \cdot P_{\alpha_d^{\mathrm{DL}}}(n_{t,d}|N_{t,d}). \tag{4.41}$$

Table 4.3: TDMA System Parameters For Numerical Evaluation of Congestion Probability.

Parameter	Symbol	Value
Total number of available channels	C	23
Max. number of data users per TSL in UL (DL)	n_C^{UL} (n_C^{DL})	3 (4)
Reduction factor congestion threshold	$\xi_c^{\mathrm{UL}} = \xi_c^{\mathrm{DL}}$	0.35
Max. number of TSL per data user in UL (DL)	q^{UL} (q^{DL})	3 (8)
Voice activity factor	$\alpha_v^{\mathrm{UL}} = \alpha_v^{\mathrm{DL}}$	1
Data activity factor	$\alpha_d^{\mathrm{UL}} = \alpha_d^{\mathrm{DL}}$	0.5

Finally, the total congestion probability in TDMA due to UL and DL contributions yields

$$P_c^t(\boldsymbol{N}_t) = 1 - [1 - P_c^{\mathrm{UL},t}(\boldsymbol{N}_t)][1 - P_c^{\mathrm{DL},t}(\boldsymbol{N}_t)], \tag{4.42}$$

where the assumption that the congestion probabilities in UL and DL are independent is made.

For illustrative purposes, considering TDMA system parameters provided in Table 4.3, the total TDMA congestion probability as defined in (4.42) is given in Fig. 4.6. It should be observed that with the chosen parameters in Table 4.3, data users are granted with a higher amount of TSL in DL than in the UL (8 as opposed to 3). This asymmetry in channel allocation causes the congestion probability to be higher in the DL case, where TSL reuse is potentially higher, than in the UL case. As a consequence, the total congestion probability in TDMA illustrated in Fig. 4.6 will be mainly DL-driven.

4.6 Model Evaluation

In this section, numerical evaluation of the proposed model is carried out. The reader is referred to Chapter 2 for details concerning the validation of the Markov allocation model. Numerical results for the model evaluation of congestion probabilities are computed in two steps. First, the evaluation of the allocation model (presented in Section 4.2) determines the probability, $P_{\boldsymbol{N}}$, of having a given number of users in each RAT according to a particular RAT selection policy π. Once the statistical distribution of users is known, i.e. $P_{\boldsymbol{N}}$, congestion probabilities in each RAT, given by (4.8), and the overall congestion probability, given by (4.10),

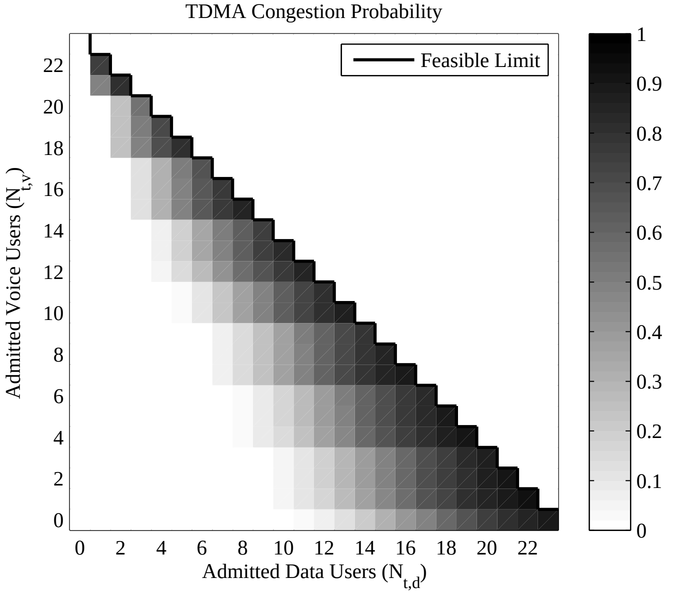


Figure 4.6: TDMA Congestion Probability.

can be computed. In addition, $P_{\boldsymbol{N}}$ obtained from the allocation model will be used to compute the performance metrics, as defined later in Section 4.6.2.

Several RAT selection policies are evaluated in the following in order to assess the performance of the proposed model considering voice and data services to be allocated in TDMA and WCDMA RATs. In particular, LB and CA policies introduced in sections 4.4.1 and 4.4.3 respectively are evaluated. In addition, an illustrative example of a SB policy, introduced in section 4.4.2, is considered aiming to prioritize voice users to be allocated to TDMA and data users to be allocated to WCDMA. It is worthwhile mentioning that the performance evaluation in this chapter is carried out for a particular illustrative case with specific parameter settings for the allocation and congestion models.

In this chapter, and without loss of generality, it is assumed that the load metrics used in the LB policy operation (see section 4.4.1) for TDMA and WCDMA systems are taken in the uplink direction as in [22]. Nevertheless, other configurations are possible in the described model.

According to the above, in TDMA-based systems, such as e.g. GSM/EDGE, the *TSL utilization factor*, [12], can be used to measure the load in a given state $S_{\boldsymbol{N}} \in \mathcal{S}$.

It is defined as the ratio between the number of occupied TSLs over the number of available TSLs, thus expressed as:

$$L^t_{\boldsymbol{N}_t} = \min\left(C, N_{t,v} + N_{t,d}\alpha_d^{\mathrm{UL}} q^{\mathrm{UL}}\right)/C \tag{4.43}$$

On the other hand, the load in a WCDMA-based system may be calculated by means of the uplink load factor defined previously in (4.20), thus

$$L^w_{\boldsymbol{N}_w} = \eta^{\mathrm{UL}}_{\boldsymbol{N}_w} \tag{4.44}$$

4.6.1 Parameter Settings

Parameter settings for TDMA/WCDMA are those given in Table 4.1, Table 4.2 and Table 4.3. In addition, it is considered that the transmission bit rate of a single voice TSL is $\kappa_v = 12.2$ kbps and for a single data TSL is $\kappa_d = 29.6$ kbps. Numerical results are obtained for a fixed total offered voice traffic of $T_v = \lambda_v/\mu_v = 10$ Erlangs and a total offered data traffic, $T_d = \lambda_d/\mu_d$, varying between 0 and 100 Erlangs.

4.6.2 Performance Metrics

Once the probabilities $P_{\boldsymbol{N}}$ for each state $S_{\boldsymbol{N}} \in \mathcal{S}$ are computed, a number of relevant performance metrics can be defined. In particular, as mentioned in section 4.3, it is of special interest to measure the congestion probability in each RAT for a specific RAT selection policy as in (4.8), along with the overall congestion probability as in (4.10). Other relevant performance metrics are detailed in the following subsections.

4.6.2.1 Blocking Probability

A given state $S_{\boldsymbol{N}} \in \mathcal{S}$ is said to be a *blocking state*, if the addition of any service type user into any RAT forces the system to move to a non-feasible state $S'_{\boldsymbol{N}}$, meaning that $S'_{\boldsymbol{N}} \notin \mathcal{S}$. Let the set of all blocking states be represented by $\mathcal{S}_b$. Then, the total blocking probability yields

$$P_b = \sum\nolimits_{S_{\boldsymbol{N}} \in \mathcal{S}_b} P_{\boldsymbol{N}}. \tag{4.45}$$

4.6.2.2 Throughput

Under congestion situations, users may undergo different levels of QoS degradation over a range of key performance indicators (e.g. delay, throughput, etc.). It has been shown that, in TDMA, congestion arises due to excessive TSL sharing, which in turn causes a degradation of the perceived throughput in data users. This degradation is quantified by means of factor ξ introduced in (4.37), being the throughput per data user in TDMA $q\kappa_d\xi$. In WCDMA, and for the sake of comparison, it is assumed that the throughput per user in WCDMA is also degraded due to congestion in such a way that the perceived throughput for voice and data users is $R_{b,v}\cdot[1-P_c^w(\boldsymbol{N}_w)]$ and $R_{b,d}\cdot[1-P_c^w(\boldsymbol{N}_w)]$ respectively. Thus reflecting that congestion in WCDMA turns into an excess of interference so that power control cannot ensure the target requirements. Then, the overall throughput per data user in a given state $S_{\boldsymbol{N}}\in\mathcal{S}$ can be defined as:

$$\Gamma_{d,u}(\boldsymbol{N})=\frac{N_{t,d}q\kappa_d\xi_{(N_{t,v},N_{t,d}q\alpha_d)}+N_{w,d}R_{b,d}[1-P_c^w(\boldsymbol{N}_w)]}{N_{t,d}+N_{w,d}}, \tag{4.46}$$

with $\xi_{(\tau_v,\tau_d)}$ defined in (4.37) and where q, $R_{b,d}$ and P_c^w should be particularized for the UL and DL accordingly. Then, the average throughput per data user can be obtained as:

$$\Gamma_{d,u}=\sum\nolimits_{S_{\boldsymbol{N}}\in\mathcal{S}}\Gamma_{d,u}(\boldsymbol{N})\cdot P_{\boldsymbol{N}}. \tag{4.47}$$

Finally, it is also interesting to define the total aggregate throughput contributed by all services and RATs. For a given state $S_{\boldsymbol{N}}\in\mathcal{S}$ the aggregate throughput may be expressed as

$$\begin{aligned}\Gamma_a(\boldsymbol{N})=&N_{t,v}\kappa_v+N_{t,d}\alpha_d q\kappa_d\xi_{(N_{t,v},N_{t,d}\alpha_d q)}\\&+N_{w,v}\alpha_v R_{b,v}[1-P_c^w(\boldsymbol{N}_w)]+N_{w,d}\alpha_d R_{b,d}[1-P_c^w(\boldsymbol{N}_w)],\end{aligned} \tag{4.48}$$

where q, $R_{b,v}$, $R_{b,d}$ and P_c^w should be particularized for the UL and DL accordingly. As a result, the average aggregate throughput may be written as

$$\Gamma_a=\sum\nolimits_{S_{\boldsymbol{N}}\in\mathcal{S}}\Gamma_a(\boldsymbol{N})\cdot P_{\boldsymbol{N}}. \tag{4.49}$$

4.6.3 Numerical Results

Fig. 4.7 shows the overall congestion probability, as a result of applying (4.10), when RAT selection policies SB, LB and CA are used. The CA policy was designed so as to balance the congestion probabilities in both RATs, which therefore

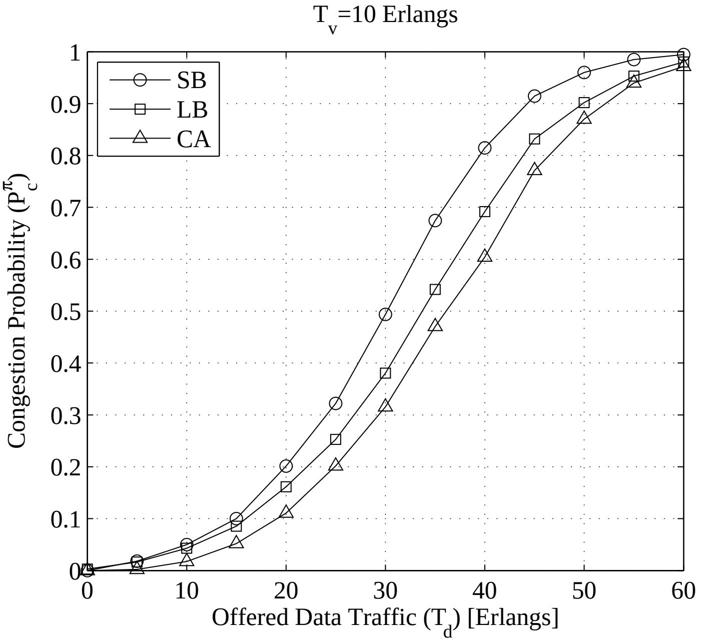


Figure 4.7: Overall congestion probability for SB, LB and CA.

translates into the lowest overall congestion probability in the considered scenario. Since LB provides similar congestion probability in TDMA with respect to SB (see Fig. 4.8 left), while improved congestion probability in WCDMA (see Fig. 4.8 right), the overall congestion probability is better for the LB policy than for the SB policy.

Fig. 4.8 shows the congestion probability for the different considered policies in each RAT as formulated in (4.8). Accordingly, since SB policy mainly directs voice users to TDMA, and congestion in TDMA is exclusively caused by data TSL reuse, hardly any congestion is detected in this RAT. Only for high offered data traffic, when WCDMA is unable to handle all the data traffic load, congestion in TDMA starts to rise (see for $T_d = 60$ Erlangs). On the other hand, SB mainly directs data users to WCDMA which causes congestion in this RAT to rise as the data traffic load is increased. It should be noted, see for instance Fig. 4.5, that although excessive voice users in WCDMA may cause congestion, an excess of data users is somewhat more problematic. LB policy, on the other hand, intends to distribute users so that loads in each RAT (conveniently defined in (4.43) and (4.44)) are balanced. Accordingly, LB operates in such a way that data users are not forced to share TSLs in TDMA unless no capacity is left in WCDMA [22]. Then, for the case under study, LB prevents from TSL reuse in TDMA thus exhibiting a low congestion profile in this RAT. In WCDMA, LB will allocate both voice

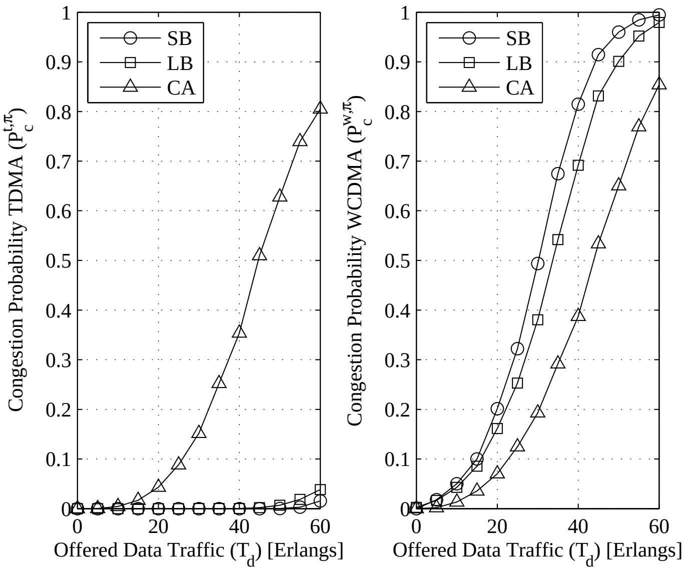


Figure 4.8: Congestion probability in TDMA (left) and WCDMA (right) for SB, LB and CA.

and data users, therefore congestion will rise whenever data traffic load increases. Nevertheless, given LB allocates voice and data resources in WCDMA, as opposed to SB that allocates only data users, the congestion probability in WCDMA is somewhat better for LB than for SB. Finally, as expected, it can be seen how CA policy balances the congestion probability in each RAT.

Fig. 4.9 shows the blocking probability as defined in (4.45) for the different policies under study. It follows from (4.18) that the allocation of a single voice user in TDMA implies a TSL consumption of 1 TSL over a total of C available TSLs, i.e. $1/C$. For a single data user in TDMA this consumption is given by $\alpha_d q/n_C C$. Based on parameters given in Table 4.3, the TSL consumption of both voice and data users in TDMA is equal to 1/23. This means that, from a resource consumption point of view, it is equally suitable to allocate voice or data users to TDMA. Moreover, the resource consumption in WCDMA may be quantified in both the UL and DL: in the UL, by means of the load factor definition, given in (4.20), with $\alpha_v^{\mathrm{UL}}/[A_v^{\mathrm{UL}}\eta_{max}^{\mathrm{UL}}]$ and $\alpha_d^{\mathrm{UL}}/[A_d^{\mathrm{UL}}\eta_{max}^{\mathrm{UL}}]$ the load fractions consumed by voice and data users respectively; and in the DL, by means of the code condition provided in (4.22), with $1/[SF_v^{\mathrm{DL}}\mathcal{C}_{max}]$ and $\alpha_d^{\mathrm{DL}}/[SF_d^{\mathrm{DL}}\mathcal{C}_{max}]$ the code consumption fractions for voice and data services respectively. With the WCDMA parameters provided in Table 4.1 and Table 4.2, it can be shown that in WCDMA a data user demands more resources than a voice user. In this sense, it is much more suitable, in the considered scenario, to allocate voice users in WCDMA and data users in

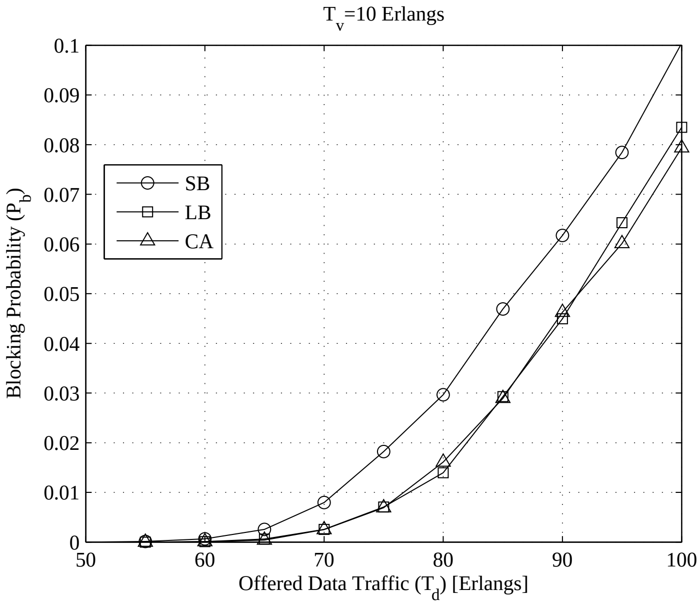


Figure 4.9: Total blocking probability for SB, LB and CA.

TDMA, thus explaining the worse behavior in terms of blocking probability of SB, which allocates data users to WCDMA and voice users to TDMA. This is in line with conclusions raised in [6], where the allocation suitability of two services onto two RATs with linear admission limits depends on the slope of these limits, provided by (4.18), (4.20) and (4.22), and illustrated in Fig. 4.1 and Fig. 4.2. On the other hand, policies LB and CA are more flexible in allocating voice and data users in TDMA and WCDMA than SB policy, thus achieving an improved blocking probability with respect to SB.

Fig. 4.10 and Fig. 4.11 illustrate the average downlink data throughput per user and the average downlink throughput as defined in (4.47) and (4.49) respectively. For SB policy, given it mainly allocates data users to WCDMA, throughput per data user will be approximately 64 kbps in the absence of congestion, which is the considered allocated DL bit-rate for data users in WCDMA (see Table 4.2). This happens for an offered data traffic load of $T_d = 10$ Erlangs. As the offered data traffic increases, so does the congestion probability in WCDMA (see Fig. 4.8 right) and thus the throughput per data user is severely degraded up to $T_d = 55$ Erlangs. This effect is also noted in the average downlink throughput in Fig. 4.11, where for SB, and after an initial throughput rise, the impact of congestion over data users in WCDMA causes a severe throughput degradation. From $T_d = 60$ Erlangs and onwards, data users start getting allocated to TDMA since WCDMA is at full capacity. Then, throughput per user is slightly increased given that the throughput

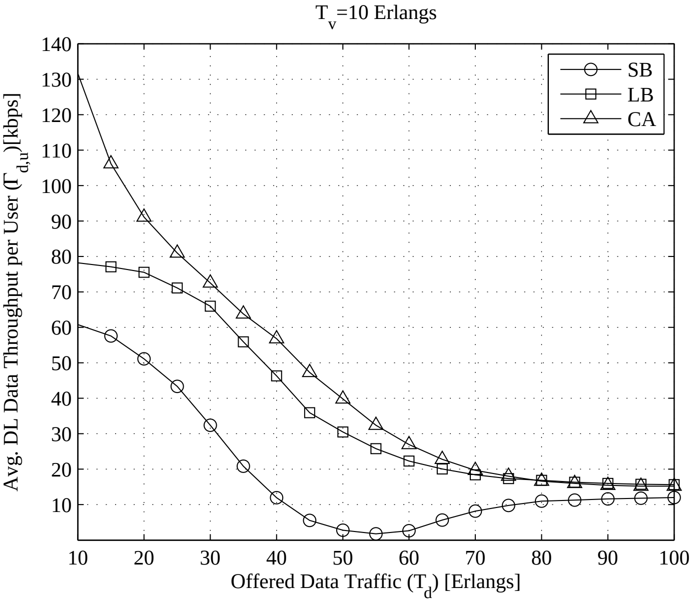


Figure 4.10: DL data throughput per user for SB, LB and CA.

degradation in TDMA is less harsh than in WCDMA. This throughput increase is also noted in the average downlink throughput given in Fig. 4.11. On the contrary, both LB and CA may eventually allocate data users to TDMA where the achieved throughput per data user depends on the TSL reuse. If such TSL reuse is low, with current parameter settings, the achieved throughput per data user in TDMA can be improved with respect to WCDMA. Nevertheless, as the number of data users increases in TDMA, the higher the TSL sharing and thus the lower the data throughput per user. It is worthwhile noticing that LB policy disregards the effect of throughput degradation in WCDMA due to congestion while only capturing the effect of TSL reuse in TDMA (since it prevents from TSL reuse unless it is strictly necessary). This explains the improved performance of CA with respect to LB, given CA captures both throughput degradation impacts in TDMA and WCDMA. This effect can be observed both in Fig. 4.10 and in Fig. 4.11. As for the average downlink throughput, Fig. 4.11, in order to achieve high throughput values, a RAT selection policy must achieve both low blocking probability and high throughput per user. Then, from Fig. 4.9 and Fig. 4.10 it can be expected that policies CA and LB present an overall best performance with respect to SB, and that CA outperforms LB for the case under study.

Despite having an efficient initial RAT selection policy that minimizes the occurrence of congestion events, further actions may be required in the case that such

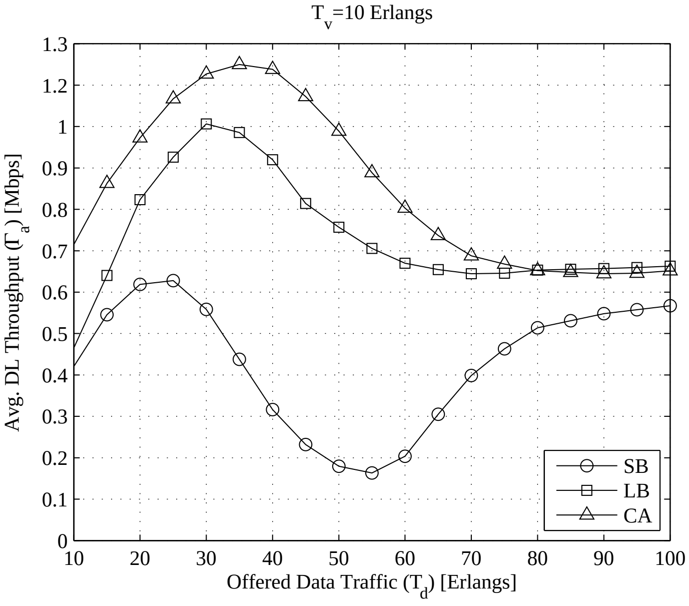


Figure 4.11: DL aggregate throughput for SB,LB and CA.

events happen. These mechanisms, referred to as congestion resolution mechanisms, provide the means to alleviate congestion situations by actuating over some specific network parameters. One possible mechanism to lessen the congestion status of a given RAT consists in actuating over the Admission Control (AC) such that fewer users are allowed in the system [2, 3]. In practice, this implies a reduction of the feasible limits, depicted in Figs. 4.1 and 4.2 for TDMA and WCDMA respectively. In WCDMA this can be achieved by conveniently setting the maximum load factor (η_{max}^{UL}) and the maximum OVSF code condition limit ($\mathcal{C}_{max}$) for the UL and DL respectively. In the same way, the feasible limit in TDMA can be managed by setting the maximum number of data users per TSL, n_C^{UL} and n_C^{DL}, for the uplink and downlink respectively.

In order to show the benefits of considering congestion information when performing initial RAT selection procedures, some additional results are presented. In this respect, the numerical study considers a range of values (see Table 4.4) for n_C^{UL} and n_C^{DL}, which determines the AC for TDMA, along with η_{max}^{UL} and $\mathcal{C}_{max}$ which control the AC procedure in WCDMA. A decrease in the value of these parameters indicates higher constraints in terms of user admission. Consequently, Fig. 4.12 shows the congestion probability and the blocking probability for policies SB and CA under the AC scenarios defined in Table 4.4. Offered voice and data traffic values are T_v=10 and T_d=40 Erlangs respectively. It can be observed for both policies, that by restricting the admission, i.e. moving from case (a) towards case (f), the congestion probability is reduced. On the other hand, imposing tighter admission conditions comes at the cost of increased blocking probabilities, as also reflected in Fig. 4.12, which also degrades the users' perceived QoS. Note also that, by considering the congestion probability as an input criterion for initial RAT selection, as in policy CA, lower congestion probability situations can be achieved at the cost of improved blocking probabilities. Whereas for the SB policy, in order to reduce the experienced congestion probability, solely acting over the AC will severely penalize users in terms of blocking. For example, CA policy achieves a congestion probability just over 0.2 when the AC setup is (e), which in turn gives a blocking probability below 0.02. As for the SB policy, in order to achieve a similar congestion probability value, AC needs to be more stringent than for the CA policy case, which in turn results in a blocking probability over 0.15. This suggests that, although congestion resolution mechanisms can be applied to solve congestion situations, such as actuating over the AC, an appropriate election of the initial RAT selection policy is of great importance.

Table 4.4: Admission Control (AC) parameters in TDMA and WCDMA.

AC Setup	n_C^{UL}	n_C^{DL}	η_{max}^{UL}	$\mathcal{C}_{max}$
(a)	3	4	0.9	63/64
(b)	2	3	0.8	53/64
(c)	1	2	0.7	43/64
(d)	1	1	0.6	33/64
(e)	1	1	0.5	23/64
(f)	1	1	0.4	13/64

4.6.4 Computational Considerations

While the values of congestion probabilities $P_c^t(\boldsymbol{N}_t)$ and $P_c^w(\boldsymbol{N}_w)$ are easily computed, determining probabilities $P_{\boldsymbol{N}}$ for all states $S_{\boldsymbol{N}} \in \mathcal{S}$ by solving the SSBEs may seem computationally complex. In general, computational complexity for solving the system of equations given by the SSBEs increases with the state dimension M and, consequently, with the number of states in the state space N [21]. It was shown (see section 4.2.1) that the higher the number of services J and/or RATs K, the higher the dimensionality of our model M. Therefore, for large values of K and J, computational cost may increase dramatically. Nevertheless, operators may typically manage no more than 3 or 4 RATs in a given area and, although

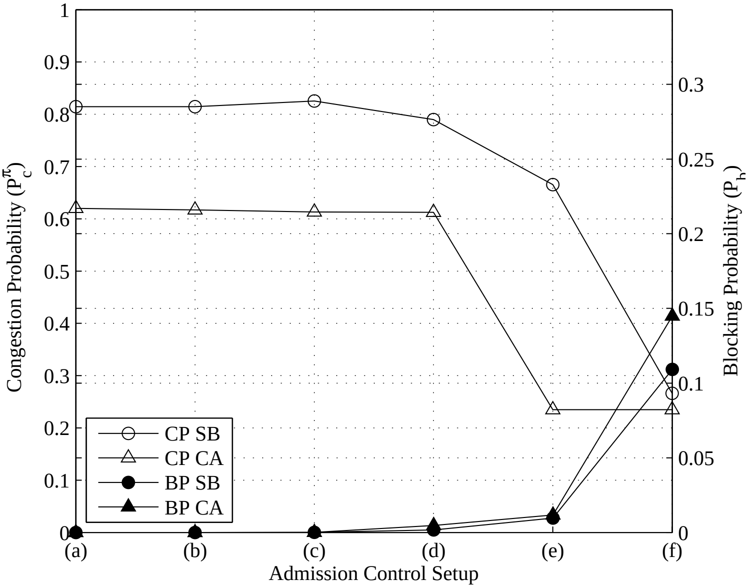


Figure 4.12: Blocking probability (BP) and Congestion Probability (CP) against several Admission Control setups provided in Table 4.4.

offered services are high in number, they may be typically grouped into 4 different QoS traffic classes according to how delay sensitive they are [23], namely: conversational, streaming, interactive and background. In addition, not all RATs support all traffic classes, thus diminishing the impact on the state dimensionality. Bearing this in mind, the solution of the SSBEs may be carried out using well-known efficient numerical methods. In particular, a so-called *iterative power* procedure will be utilized for such task [20, 21]. Other methods such as Gauss-Seidel have also proved to be effective, see e.g. [24]. The iterative power operation is based on iteratively performing the product of a probability vector (of dimension $N{\times}1$) with the $N{\times}N$ transition probability matrix ($\boldsymbol{P}$). If i iterations are needed for convergence then a total number of $i{\times}N^2$ multiplications are needed. Fortunately, matrix $\boldsymbol{P}$ is usually sparse, i.e. it contains a large amount of zero entries. Then, if N_z is the total number of non-zero entries in matrix $\boldsymbol{P}$, a total of $i{\times}N_z$ multiplications are now required [21]. In addition, the involved memory storage requirements of such procedure can be easily handled by off-the-shelf computer equipments.

4.7 Chapter Summary

In this chapter, a complete, detailed and generalized framework for the evaluation of multi-service allocation in multi-access systems by means of policies has been provided. In this sense, a generalized policy definition framework has been introduced capable to respond to different allocation principles. In addition, an analytical statistical characterization for radio access congestion in multi-RAT environments has also been presented. The evaluation of several RAT selection policies has been carried out in a combined TDMA/WCDMA scenario with voice and data services. In this case, specific analytical expressions for the congestion probability have been provided for both TDMA and WCDMA.

In particular, three RAT selection policies have been appropriately defined and evaluated: Load Balancing (LB), Service-Based (SB) and Congestion-Aware (CA). In the case under study, results revealed that SB prevents from data resource reuse in TDMA by allocating voice users in this RAT, but at the cost of increased congestion probability in WCDMA due to data users. Furthermore, SB provides the worst blocking probability behavior among the considered policies. As for LB policy, it also prevents from data TSL reuse in TDMA but exhibits higher flexibility in allocating voice and data users in TDMA and WCDMA. As a result, the congestion probability and throughput per user is improved with respect to SB. However, LB disregards the impact of congestion probability in the throughput perceived in WCDMA. Then, using congestion information, as with CA policy, can lead to a better performance in terms of both congestion probability and throughput. In addition, the use of congestion information as a guiding principle for initial RAT selection can also prevent from high blocking situations which result from applying tighter AC mechanisms in order to reduce such congestion.

With the proposed framework, the impact of different RAT selection policies on the congestion probability can be measured and QoS degradation is assessed. The presented framework enables to extend the current results in other scenarios with other technologies and QoS requirements.

Bibliography

[1] 3GPP, "Radio resource management strategies (rel. 6)," TR 25.922 v6.0.1, 3rd Generation Partnership Project (3GPP), 2004.

[2] X. Gelabert, J. Pérez-Romero, O. Sallent, and R. Agustí, "Congestion control strategies in multi-access networks," in *Wireless Communication Systems, 2006. ISWCS '06. 3rd International Symposium on*, 2006, pp. 579–583.

[3] X. Gelabert, J. Perez-Romero, O. Sallent, and R. Agusti, "On managing multiple radio access congestion events in b3g scenarios," in *Proc. 65th Semi-annual IEEE Vehicular Technology Conference Spring (VTC-Spring'07)*, Dublin, Ireland, April 22-25, 2007.

[4] 3GPP, "Improvement of RRM across RNS and RNS/BSS," TR 25.881 v5.0.0, 3rd Generation Partnership Project (3GPP), 2001.

[5] 3GPP, "Improvement of RRM across RNS and RNS/BSS (post rel-5) (release 6)," TR 25.891 v0.3.0, 3rd Generation Partnership Project (3GPP), 2003.

[6] A. Furuskar and J. Zander, "Multiservice allocation for multiaccess wireless systems," vol. 4, no. 1, pp. 174–184, 2005.

[7] Gábor Fodor, Anders Furuskär, and Johan Lundsjö, "On access selection techniques in always best connected networks," in *In Proc. ITC Specialist Seminar on Performance Evaluation of Wireless and Mobile Systems*, 2004.

[8] R. Chakravorty, P. Vidales, K. Subramanian, I. Pratt, and J. Crowcroft, "Performance issues with vertical handovers - experiences from gprs cellular and wlan hot-spots integration," in *Second IEEE Annual Conference on Pervasive Computing and Communications, 2004. Proceedings of the.* 2004, pp. 155–164, IEEE Comput. Soc.

[9] J. Perez-Romero, O. Sallent, and R. Agusti, "A generalized framework for multi-rat scenarios characterisation," in *2007 IEEE 65th Vehicular Technology Conference - VTC2007-Spring.* April 2007, pp. 980–984, IEEE.

[10] Lorenza Giupponi, Ramon Agusti, Jordi Perez-Romero, and Oriol Sallent Roig, "A novel approach for joint radio resource management based on fuzzy neural methodology," *IEEE Transactions on Vehicular Technology*, vol. 57, no. 3, pp. 1789–1805, 2008.

[11] F. Malavasi, M. Breveglieri, L. Vignali, P. Leaves, and J. Huschke, "Traffic control algorithms for a multi access network scenario comprising gprs and umts," in *The 57th IEEE Semiannual Vehicular Technology Conference, 2003. VTC 2003-Spring.* 2003, pp. 145–149, IEEE.

[12] T. Halonen, J. Romero, and J. Melero, *GSM, GPRS and EDGE Performance: Evolution Towards 3G/UMTS*, John Wiley & Sons, 2003.

[13] Jordi Pérez-Romero, Oriol Sallent, Ramon Agustí, and Miguel Ángel Díaz-Guerra, *Radio Resource Management Strategies in UMTS*, John Wiley & Sons, 2005.

[14] S. Nanda, "Teletraffic models for urban and suburban microcells: cell sizes and handoff rates," vol. 42, no. 4, pp. 673–682, 1993.

[15] Daehyoung Hong and S. S. Rappaport, "Traffic model and performance analysis for cellular mobile radio telephone systems with prioritized and nonprioritized handoff procedures," vol. 35, no. 3, pp. 77–92, 1986.

[16] Jingao Wang, Qing-An Zeng, and D. P. Agrawal, "Performance analysis of a preemptive and priority reservation handoff scheme for integrated service-based wireless mobile networks," vol. 2, no. 1, pp. 65–75, 2003.

[17] Christian Hartmann and Nidal Nasser, "Modeling and performance analysis of multi-service wireless cdma cellular networks using smart antennas," *Wireless Communications and Mobile Computing*, vol. 9, no. 1, pp. 117–129, 2009.

[18] Yi-Bing Lin, Wei-Ru Lai, and Rong-Jaye Chen, "Performance analysis for dual band pcs networks," *Transactions on Computers*, vol. 49, no. 2, pp. 148–159, 2000.

[19] Leonard Kleinrock, *Queueing Systems. Volume 1: Theory*, Wiley-Interscience, 1 edition, January 1975.

[20] Gunter Bolch, Stefan Greiner, Hermann de Meer, and Kishor S. Trivedi, *Queueing Networks and Markov Chains: Modeling and Performance Evaluation with Computer Science Applications*, WileyBlackwell, 2nd edition edition, May 2006.

[21] William J. Stewart, *Introduction to the numerical solution of Markov chains*, Princeton University Press, 1994.

[22] X. Gelabert, J. Perez-Romero, O. Sallent, and R. Agusti, "A 4-Dimensional Markov Model for the Evaluation of Radio Access Technology Selection Strategies in Multiservice Scenarios," in *Proc. 64th Semi-annual IEEE Vehicular Technology Conference Fall (VTC-Fall'06)*, Montreal, Canada, September 25-28, 2006.

[23] 3GPP, "Quality of Service (QoS) concept and architecture (rel. 5)," TS 23.107 v5.7.0, 3rd Generation Partnership Project (3GPP), 2002.

[24] S. J. Lincke, "Validation of a load shared integrated network with heterogeneous services," in *40th Annual Simulation Symposium (ANSS'07)*, Los Alamitos, CA, USA, 2007, vol. 0, pp. 49–59, IEEE Computer Society.

Part II

Spectrum Management (SM) in Cognitive Radio Networks

Chapter 5

Spectrum Sharing in Cognitive Radio Networks with Imperfect Sensing: A Discrete-Time Markov Model

An efficient and utmost utilization of currently scarce and underutilized radio spectrum resources has stimulated the introduction of what has been coined Cognitive Radio (CR) access methodologies and implementations. While the long-established approach has been based on licensed (or primary) spectrum access, this new communication paradigm enables an opportunistic secondary access to shared spectrum resources provided mutual interference is kept below acceptable levels. In this chapter we address the problem of primary-secondary spectrum sharing in cognitive radio access networks using a framework based on a Discrete Time Markov Chain (DTMC) model. Its applicability and advantages with respect to other approaches is explained and further justified. Spectrum awareness of primary activity by the secondary users is based on spectrum sensing techniques, which are modeled in order to capture sensing errors in the form of false-alarm and missed-detection. Model validation is successfully achieved by means of a system-level simulator which is able to capture the system behavior with high degree of accuracy. Parameter dependencies and potential tradeoffs are identified enabling an enhanced operation for both primary and secondary users. The suitability of the specified model is justified while allowing a wide range of extended implementations and enhanced capabilities to be considered.

5.1 Spectrum Sharing: Motivation and Problem Statement

In this chapter we tackle the problem of dynamic spectrum access considering the Hierarchical Access Model [1], where the licensed (or primary) spectrum is opened to secondary users (SUs) provided the interference over the primary users (PUs, or licensees) is kept under acceptable limits. In addition, two approaches for spectrum sharing have been devised: Spectrum Underlay and Spectrum Overlay. Spectrum underlay aims at operating below the floor noise of primary users by using ultra-wideband (UWB) techniques which, on the other hand, limits the transmitted power by secondary users. As for spectrum overlay, it targets at spatio-temporal spectrum holes by allowing secondary users to identify and exploit them in a non-intrusive manner. In the remainder of the chapter, it will be assumed that spectrum overlay is used as a basis of our model. From a regulatory perspective, the Federal Communications Commission (FCC) in the U.S. and Ofcom in the U.K. are currently considering the use of cognitive radio technologies [2]. Accordingly, the unlicensed use of VHF and UHF TV bands, provided no harmful interference is caused, was targeted by the FCC in [3]. This was a first milestone in the development of the IEEE 802.22 standard, proposing a cognitive radio-based physical and medium access control (MAC) layer for use of TV spectrum bands by license-exempt devices on a non-interfering basis [4]. Furthermore, the IEEE activities in developing architectural concepts and specifications for network management interoperability, including CR and dynamic spectrum access, are addressed by SCC41/P1900 standardization groups [5]. Finally, many operative standards such as WiFi (IEEE 802.11), Zigbee (IEEE 802.15.4), and WiMAX (IEEE 802.16) already include some degree of CR technology today [2], in the form of coexistence among radios, Dynamic Frequency Selection (DFS) and Power Control (PC). The primary-secondary (P-S) spectrum sharing operation can take the form of cooperation or coexistence. Cooperation means there is explicit communication and coordination between primary and secondary systems, and coexistence means there is none [6]. When sharing is based on coexistence, secondary devices are essentially invisible to the primary. Thus, all of the complexity of sharing is borne by the secondary and no changes to the primary system are needed. Among the different forms of coexistence, we adopt the opportunistic exploitation of white spaces in spatial-temporal domain sustained on spectrum sensing, coordination with peers and fast spectrum handover, i.e. the spectrum overlay case. As for cooperation, again different forms of P-S interactions are possible. For example, spatial-temporal white spaces can be signaled through a common control channel from the primary network side, such as the Cognitive Pilot Channel (CPC) or the CSCC (Common Spectrum Coordination Channel) [7–12], which would provide primary spectrum usage information to SUs. In addition, the interaction between PUs and SUs provides an opportunity for the license-holder to demand payment according to the different quality of service grades offered to SUs. In the abovementioned context, the use of Markov models becomes an important aid in modeling problems dealing with the dynamic access to shared spectrum resources. In this sense, a significant number of papers in the literature have been devoted to the characterization of such scenarios using Markov models as, e.g., in [13–18]. In [13] (along with amendments in [19]), a Continuous Time Markov Chain (CTMC) model is presented to model spectrum access of primary (wideband) and secondary (narrowband) users over a partitioned spectrum bandwidth. In [14] and [15] a CTMC model is also provided for the opportunistic access of wideband and narrowband users. However, as a difference from [13], a finite population traffic model is used for the characterization of secondary users. It is worthwhile noting that work in [13–15] disregards the effect of erroneous sensing on the secondary network side, i.e. a perfect knowledge on the activity of primary users is assumed. An attempt to introduce the impact of sensing errors is provided in [16], where a CTMC model is also considered and sensing information is available upon secondary user arrival. Despite the fact that some considerations about sensing errors are introduced in [16], these are not related to any particular spectrum sensing mechanism (i.e., energy detection, pilot detection, etc. [20]). Conversely, in this chapter, missed-detection and false-alarm values are obtained according to the well-known expressions regarding the energy detection of signals in Rayleigh fading as in [21] and [22], accordingly achieving higher modeling accuracy. In [17, 18], CTMC models are used to characterize the interactions between primary and secondary users and random spectrum access protocols are proposed and evaluated. In this chapter, a Markovian framework based on Discrete Time Markov Chains (DTMC) to evaluate the opportunistic spectrum access in a P-S spectrum sharing scenario is proposed. The rationale behind using DTMCs instead of CTMCs is based on the fact that sensing mechanisms operate on a periodic time basis, and where the sensing periodicity is an important design parameter. Therefore, the DTMC models, which observe the state of the system at discrete time instants, can accurately model the proposed scenarios by considering the observation instants of the DTMC as the sensing instants. Model validation and evaluation studies considering several parameter dependency issues and tradeoffs are addressed in this chapter revealing the usefulness of the proposed model for cognitive radio networks system design, realization and operation. In particular, relevant parameters are identified that influence the performance of the spectrum sharing model. Among these, sensing periodicity (*how often do we sense?*) and sensing accuracy (*how well do we sense?*) are shown to be key parameters that greatly affect the behavior of the system. In addition, this chapter reflects the importance of time-sharing between spectrum sensing (*for how long do we sense?*) and data transmission (*for how long do we transmit?*), which tradeoffs the sens-

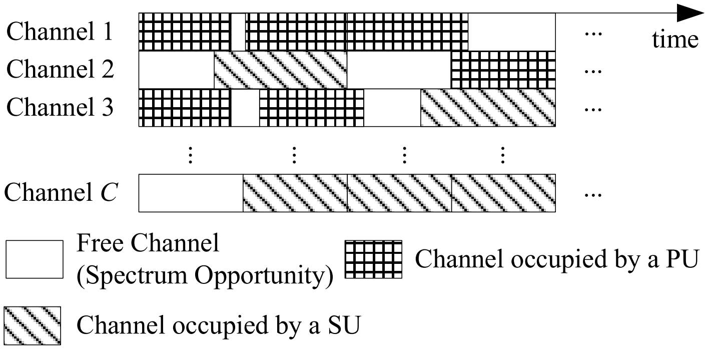


Figure 5.1: Bandwidth partitioning model for spectrum sharing between PUs and SUs.

ing accuracy with the obtained throughput, thus leading to possible parameter optimization which will be also addressed in this chapter. Finally, the awareness of both primary and secondary traffic load distributions also enables to identify optimized parameter values for an overall enhanced network operation as will be shown in the following. The remainder of the chapter is organized as follows. In Section 5.2 the system model is described along with the considered procedures and the implementation approach. Subsequently, in Section 5.3, the DTMC model is formulated along with the main hypothesis and considerations. A number of relevant performance metrics are derived in Section 5.4 which will be evaluated numerically in Section 5.5. Finally, Section 5.6 concludes the chapter with some final remarks and future considerations.

5.2 System Model

The considered system involves a Primary Network (PN), serving PUs, and a Secondary Network (SN), serving SUs. Both the PN and the SN operate autonomously and each network implements efficient protocols for the correct and coordinated operation among their own users (i.e. PUs and SUs respectively). Thus, the PN is aware of the spectrum occupancy by PUs and, correspondingly, the SN is aware of the spectrum occupancy of SUs. The PN has been assigned a total number of C channels, partitioning a certain frequency bandwidth. SUs can make use of free channels; though PUs have strict priority over SUs (i.e. if a SU is using a given channel and this channel is required by a PU, then the SU must release it). Fig. 5.1 shows an example of the considered bandwidth sharing model.

5.2.1 Procedure

The procedure to be followed for the operation in the SN and the corresponding required functionalities is presented in the following. In the first stage, a frequency band and a specific available channel where a secondary communication can be established have to be identified. Then, both secondary communication ends have to be configured to be able to transmit and receive over the identified channel. While maintaining the secondary communication, it is required that the presence of a primary communication is detected, so that if a PU arrives the secondary communication must evacuate the channel. Spectrum handover (SpHO) procedures will intend to find an appropriate alternative channel where the secondary communication can be continued in order to avoid its interruption.

5.2.2 Spectrum Awareness Implementation Approach

Among the two aforementioned implementation approaches (i.e. coordination vs. coexistence) the coexistence case will be adopted in the sense that the SN implements spectrum discovery mechanisms in order to exploit unused spectrum in an opportunistic fashion. In this approach, the identification of a candidate frequency band (or channel) for the secondary communication as well as the monitoring of primary's presence is performed within the SN based on sensing mechanisms without any direct interaction with PN. Depending on the secondary network architecture, whether it is infrastructure-based (centralized) or infrastructure-less (i.e. decentralized or *ad-hoc*), sensing information may be gathered in different forms. In the centralized case, SUs equipped with sensors may sense the whole spectrum and report to a centralized entity (e.g. located at the Base Station, BS) which may in turn schedule or re-schedule SU transmissions accordingly. Alternatively, a centralized entity at the BS may be responsible for sensing tasks, thus alleviating SUs from sensing capabilities. For the decentralized approach, spectrum sensing and use is entirely borne by SUs, thus information exchange mechanisms among them should be implemented. For the sake of simplicity, and to avoid a further increase in the model complexity, we will assume the centralized case where a centralized entity is responsible for spectrum sensing tasks. Nevertheless, the presented model allows further implementation alternatives to be considered which is left for future work.

Channel occupancy detection performed at the SU's terminal side through sensing mechanisms is affected by a number of aspects (e.g. adverse channel conditions, hidden terminal problem, limited sensitivity on the sensing equipment, etc.) that may limit the reliability of sensing results [21]. Typically, spectrum detection through sensing in the presence of errors performs a binary hypotheses test over a given band (or channel), that is: $\mathcal{H}_0$ if the channel is available and $\mathcal{H}_1$ if the channel is occupied. Accordingly, the miss-detection and false-alarm probabilities, δ and ε, can be defined as:

$$\delta = \Pr\left[\mathcal{H}_0|\mathcal{H}_1 \text{ is true}\right] \tag{5.1}$$

$$\varepsilon = \Pr\left[\mathcal{H}_1|\mathcal{H}_0 \text{ is true}\right] . \tag{5.2}$$

An appropriate selection of the so-called *time-bandwidth product*, defined as

$$m = T \cdot W \, , \tag{5.3}$$

where T is the time devoted to sense bandwidth W [21], is of great relevance. In general, the longer we sense the bandwidth W seeking for spectrum opportunities the more reliable are our sensing measures (i.e. lower δ and ε values), however, high T values, as shown further on, will trade-off the achievable throughput experienced by SUs [23]. Moreover, spectrum errors can be improved by means of the cooperation of sensing entities as suggested in [21].

With respect to the availability of updated primary spectrum occupancy information based on sensing, ΔT would represent the time between two consecutive sensing information updates. Considering that a generic underlying MAC level time-frame structure enabling sensing would devote some time, T_{sens}, for sensing purposes, i.e. the *sensing time*, the *sensing efficiency* can be defined as

$$\eta_{sens} = 1 - T_{\text{sens}}/\Delta T \, . \tag{5.4}$$

We assume that time T_{sens} will be the time T, devoted to sense a single channel, multiplied by the number of channels with bandwidth W that should be sensed, i.e.

$$T_{sens} = T \cdot C \, , \tag{5.5}$$

To account for possible detection errors during spectrum sensing procedures, a probabilistic model will be developed in order to compute the number of detected (or sensed) PUs in the system. This model will consider specific missed-detection and false-alarm probabilities values, δ and ε, obtained from well-known expressions in the literature, see e.g. [21, 22].

5.3 DTMC Model Formulation

The proposed DTMC model is devoted to determine the statistical occupancy of the shared spectrum by PUs and SUs. It is mainly fed by traffic-related input parameters, such as arrival and departure rates (λ_p and λ_s along with μ_p and μ_s for PUs and SUs correspondingly), and also the number of channels to be shared, C. Following the same assumptions than in [13–18] and as it is classically considered in most of the existing literature, it is assumed that arrivals follow a Poisson distribution and that service times are exponentially distributed.

The sensing periodicity, ΔT, which denotes the periodic time instants in which updated spectrum occupancy information is made available for secondary communication, is, on the other hand, the operating time-basis of the DTMC.

The proposed DTMC model accounts for the spectrum usage of PUs and SUs in a shared spectrum scenario. For simplicity reasons, it is supposed that the whole spectrum bandwidth is partitioned into a total number of C channels (bands) to be shared among both PUs and SUs. It is further considered that both PUs and SUs demand a single channel for transmission purposes (recall Fig. 5.1). These assumptions, although simplifying, will keep the algebra at an understandable and tractable level while still capturing the essence of the problem under study. If desirable, more elaborate shared bandwidth models can be easily considered and adapted to the model here presented (e.g, considering different bandwidth requirements for PUs and SUs). In this respect, Chapter 6 will deal with an enhanced bandwidth model.

In a DTMC, [24], we observe the system state at discrete time instants

$$\{t_0, t_1, t_2, ..., t_n, ...\},$$

with $t_n = t_0 + n \cdot \Delta T$ and periodicity ΔT , which is, on the other hand, assumed to specify the time instants where primary spectrum usage information is made available for secondary communication use. In addition, let $I_n = (t_n, t_{n+1}]$ define the *n-th* time interval between two successive observation times. The DTMC model formulation involves a number of steps which are presented in the following subsections.

5.3.1 State Space Definition

Let $N_p(t_n)$ and $N_s(t_n)$ be stochastic processes indicative of the number of PUs and SUs in the system at time t_n. Accordingly, allow

$$\mathbf{X}_n = S_{(i,j)} = \{N_p(t_n) = i, N_s(t_n) = j\}$$

to represent a state of the DTMC at time t_n. Thus, if C channels are available, the considered state space $\mathcal{S}$ must contain all possible sates $S_{(i,j)}$ which fulfill both $i \leq C$ and $j \leq C$, formally:

$$\mathcal{S} = \left\{ S_{(i,j)} : i \leq C, j \leq C \right\} . \tag{5.6}$$

Nevertheless, for a correct spectrum use (i.e. with no spectrum collisions), the number of PUs (i) plus the number of SUs (j) must not exceed the total number of available channels (C). In addition, due to spectrum detection errors, a SU might be erroneously assigned to a band already in use by a PU. Then, for convenience, we define the following three subsets of accounting for those states that necessarily imply spectrum collision, i.e.

$$\mathcal{S}_c = \left\{ S_{(i,j)} : i + j > C \right\} \subset \mathcal{S} \,, \tag{5.7}$$

those states which possibly imply a spectrum collision, i.e.

$$\mathcal{S}_{pc} = \left\{ S_{(i,j)} : i + j \leq C, j > 0, i > 0 \right\} \subset \mathcal{S} \,, \tag{5.8}$$

and those states that are collision-free, i.e.

$$\mathcal{S}_{nc} = \left\{ \left(S_{(i,j)} : i = 0 \right) \cup \left(S_{(i,j)} : j = 0 \right) \right\} \subset \mathcal{S} \,. \tag{5.9}$$

Fig. 5.2 shows an example of the considered state space for $C = 3$ channels, where it must be regarded that $\mathcal{S} = \mathcal{S}_c \cup \mathcal{S}_{pc} \cup \mathcal{S}_{nc}$ and $\mathcal{S}_c \cap \mathcal{S}_{pc} \cap \mathcal{S}_{nc} = \emptyset$. In addition, the total number of states is given by $N_{states} = (C + 1)^2$.

A clear advantage in considering the total number of PUs (i) and SUs (j) in the system as indices of the Markov state space (as, e.g. in [13]) is that system dimensionality (and consequently complexity) of the model can be reduced. Contrarily, if the status of each channel (that is, free, occupied by PU or occupied by SU) was to be captured by the Markov model, the number of states would dramatically increase thus limiting the applicability of the model. On the other hand, the main drawback of adopted approach is that it remains uncertain how different users are distributed over the channels. Fortunately, this can be solved by considering probabilistic models for the occupation of different channels by PUs and SUs.

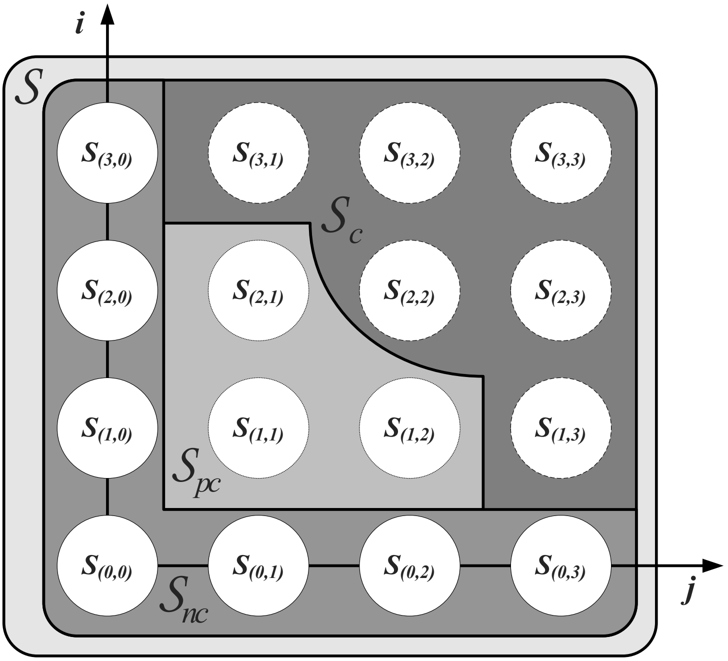


Figure 5.2: The state space for $C = 3$, where different shaded regions determine the spectrum collision states $[S_{(i,j)} \in \mathcal{S}_c]$, possible collision states $[S_{(i,j)} \in \mathcal{S}_{pc}]$ and collision-free states $[S_{(i,j)} \in \mathcal{S}_{nc}]$

5.3.2 Detection of Primary Spectrum Occupancy

At a particular time t_n, let the state of the DTMC be $\mathbf{X}_n = S_{(i,j)} \in \mathcal{S}$. At this same time instant, spectrum occupancy information is made available to the SN side (either to some centralized infrastructure-based entity or to a specific SU). Due to spectrum detection errors, the observed state at time t_n using such erroneous information may be $\mathbf{Y}_n = S_{(k,j)} \in \mathcal{S}$, i.e. $\mathbf{Y}_n \neq \mathbf{X}_n$, with k denoting the number of detected PUs (note the number of SUs at time t_n, j, is known by the SN, so it is not subject to errors). Consequently, we are interested in determining the conditional probability of detecting k PUs when there are in fact i PUs in the system at time t_n, which may formally be expressed as:

$$b_{(k,i)} = \Pr[\mathbf{Y}_n = S_{(k,j)} | \mathbf{X}_n = S_{(i,j)}] \tag{5.10}$$

Theorem 1. *The conditional primary user detection probability, $b_{(k,i)}$, subject to false alarm and missed detection probabilities, ε and δ, is given by*

$$b_{(k,i)} = \sum_{m=\max(0,i-k)}^{\min(i,C-k)} \binom{C-i}{m+k-i} \cdot \varepsilon^{m+k-i} \cdot \bar{\varepsilon}^{C-m-k} \cdot \binom{i}{m} \cdot \delta^{m} \cdot \bar{\delta}^{i-m}\,, \tag{5.11}$$

with $\bar{\varepsilon} = 1 - \varepsilon$ and $\bar{\delta} = 1 - \delta$.

Proof. See Appendix 5.A. □

Then, function $b_{(k,i)}$ provides the application function between the so-called *true state space* given by states $\mathbf{X}_n = S_{(i,j)} \in \mathcal{S}$ and the *detected state space* given by states $\mathbf{Y}_n = S_{(k,j)} \in \mathcal{S}$. Since the SN operation will be based on the knowledge of $\mathbf{Y}_n$ as opposed to $\mathbf{X}_n$, the values of ε and δ will considerably affect the performance of such system and lead, in the worst case, to an ineffective operation.

5.3.3 Arrival and Departure Processes

Let $N^A \in \{N^{PA}, N^{SA}\}$ along with $N^D \in \{N^{PD}, N^{SD}\}$ denote the number of arrivals and departures of PUs and SUs respectively in I_n (i.e. in a time interval of duration ΔT).

Given PUs and SUs arrive at the system according to a Poisson distribution with rates λ_p and λ_s respectively, the probability that k arrivals occur in I_n, P_k^A, is given by [24]:

$$P_k^A = \Pr[N^A = k] = \left[(\lambda \Delta T)^k / k!\right] e^{-\lambda \Delta T}\,, \tag{5.12}$$

where for $\lambda \in \{\lambda_p, \lambda_s\}$ we will refer to $P_k^A \in \{P_k^{PA}, P_k^{SA}\}$ correspondingly.

If the session duration is exponentially distributed with mean $1/\mu$ (i.e., rate μ), the probability of a session departure in I_n is [24]:

$$P^D = 1 - e^{-\mu \Delta T}\,. \tag{5.13}$$

Then, the probability of having k-out-of-m departures in I_n, P_k^D, is given by the binomial distribution [24]:

$$P_k^D = \Pr[N^D = k] = \binom{m}{k} (1 - e^{-\mu \Delta T})^k (e^{-\mu \Delta T})^{m-k}\,, \tag{5.14}$$

where for $\mu \in \{\mu_p, \mu_s\}$ we will refer to $P_k^D \in \{P_k^{PD}, P_k^{SD}\}$ respectively.

Note that enabling multiple arrivals and departures in one ΔT period will affect the decision process on whether a SU can be assigned or not given that detection information is retrieved only at times t_n. This also constitutes a differentiating aspect with respect to other approaches to the same problem such as in [13–18].

5.3.4 Transition Probabilities

The transition probabilities between each pair of states $S_{(k,l)} \rightarrow S_{(i,j)}$ in the DTMC model can be expressed as [24]:

$$\begin{aligned} P_{(i,j|k,l)} &= \Pr[X_{n+1}=S_{(i,j)}|X_n=S_{(k,l)}] \\ &= \Pr[N_p(t_{n+1})=i, N_s(t_{n+1})=j|N_p(t_n)=k, N_s(t_n)=l] \\ &= \Pr[N_p(t_{n+1})=i|N_p(t_n)=k, N_s(t_n)=l] \\ &\quad \times \Pr[N_s(t_{n+1})=j|N_p(t_n)=k, N_s(t_n)=l]\,, \end{aligned} \tag{5.15}$$

where the conditional independence of processes $N_p(t_n)$ and $N_s(t_n)$ has been considered (since primary and secondary arrival/departure processes are also assumed independent). Probabilities $P_{(i,j|k,l)}$ constitute the elements of the transition probability matrix $\mathbf{P}$, from which the steady state probabilities, $P_{(i,j)}$, of the DTMC will be determined [25]. For the sake of algebra tractability, the following assumptions are considered in the presented expressions:

Hypothesis 1. *A primary or secondary session arriving in I_n will not depart in the same I_n. This implies that $\Delta T << 1/\mu$ with $\mu \in \{\mu_p,\mu_s\}$ and where $1/\mu$ is the average session duration.*

Hypothesis 2. *We disregard the order in which session arrivals and departures occur in a given I_n by considering the resulting net number of users, i.e. those obtained after subtracting the departures and adding the new arrivals.*

Note that both previous hypotheses rely on the relation between ΔT and traffic-related parameters λ and μ, which can be adjusted to fit our needs. The validation and justification of these hypotheses will be addressed in Section 5.5 when we deal with the model validation.

For convenience we present the following set of lemmas:

Lemma 1. *The probability of assigning k PUs when we have also l PU de-assignments in state $S_{(i,j)}$, $a^P_{(i,j,k,l)}$, is given by:*

$$a^P_{(i,j,k,l)} = \begin{cases} P^{PA}_k, & \textit{if } i-l+k<C \\ 1-\sum\limits_{m=0}^{k-1} P^{PA}_m, & \textit{if } i-l+k=C \end{cases}. \tag{5.16}$$

Proof. See Appendix 5.B.1. □

Lemma 2. *The probability of de-assigning k PUs in state $S_{(i,j)}$, $d^P_{(i,j,k,l)}$, is given by:*

$$d^P_{(i,j,k)} = P^{PD}_k \, . \tag{5.17}$$

Proof. See Appendix 5.B.2. □

Lemma 3. *The probability of assigning k SUs when we have also l SU de-assignments in state $S_{(i,j)}$, $a^S_{(i,j,k,l)}$, is given by:*

$$a^S_{(i,j,k,l)} = \begin{cases} \sum\limits_{m=0}^{C-k-j+l} \bar{a}^S_{(m,j,k,l)} \cdot b_{(m,i)} & \textit{for } k>0 \\ \sum\limits_{m=0}^{C-j+l} \bar{a}^S_{(m,j,0,l)} \cdot b_{(m,i)} + \sum\limits_{m=C-j+l}^{C} b_{(m,i)} & \textit{for } k=0 \end{cases} , \tag{5.18}$$

with

$$\bar{a}^S_{(m,j,k,l)} = \begin{cases} P^{SA}_k, & \textit{if } m+j-l+k<C \\ 1-\sum\limits_{r=0}^{k-1} P^{SA}_r, & \textit{if } m+j-l+k=C \end{cases} . \tag{5.19}$$

Proof. See Appendix 5.B.3. □

Lemma 4. *The probability of de-assigning k SUs in state $S_{(i,j)}$, $d^S_{(i,j,k,l)}$, is given by:*

$$d^S_{(i,j,k)} = \sum_{r=0}^{k} d^{S,S}_{(i,j,k-r,r)} \cdot d^{S,SC}_{(i,j,r)} \, , \tag{5.20}$$

with

$$d^{S,S}_{(i,j,k,l)} = \begin{cases} b_{(C+k-j+l,i)} & \textit{if } 0<k\leq j-l \\ 1-\sum\limits_{r=1}^{j-l} b_{(C+r-j+l,i)} & \textit{if } k=0 \end{cases} , \tag{5.21}$$

and

$$d^{S,SC}_{(i,j,k)} = P^{SD}_k \, . \tag{5.22}$$

Proof. See Appendix 5.B.4. □

Then, it follows that:

Theorem 2. *The general transition probability between states $S_{(i,j)} \to S_{(i+N,j+M)}$, $P_{(i+N,j+M|i,j)}$, with $-i \le N \le C-i$ and $-j \le M \le C-j$, is given by:*

$$P_{(i+N,j+M|i,j)}{=}\left(\sum_{k=\max(-N,0)}^{i} a^{P}_{(i,j,N+k,k)}{\cdot}d^{P}_{(i,j,k)}\right)\cdot\left(\sum_{k=\max(-M,0)}^{j} a^{S}_{(i,j,M+k,k)}{\cdot}d^{S}_{(i,j,k)}\right), \tag{5.23}$$

Proof. See Appendix 5.C. □

5.4 Performance Metrics

From the resulting transition probability matrix $\mathbf{P}$ defined through (5.23), we obtain the true steady state probabilities, $P_{(i,j)} = \lim_{n\to\infty} \Pr\left[\mathbf{X}_n = S_{(i,j)}\right]$, for each true state $S_{(i,j)}$ in the state space $\mathcal{S}$. The knowledge of such statistical distribution enables the definition of several performance metrics which are addressed in the following.

On the other hand, it is also relevant to determine the steady state probabilities of the detected states (i.e. including possible sensing errors):

$$P'_{(i,j)} = \lim_{n\to\infty} \Pr\left[\mathbf{Y}_n = S_{(i,j)}\right],$$

which are computed as:

$$P'_{(i,j)} = \sum\nolimits_{n=0}^{C} b_{(i,n)} \cdot P_{(n,j)}\,. \tag{5.24}$$

In this case, $P'_{(i,j)}$is the steady state probability observed by the SN, i.e. without true knowledge of PU activity and, therefore, sensible to sensing errors. Then, by considering $P'_{(i,j)}$ instead of $P_{(i,j)}$, metrics computed from the SN side, which account for possible sensing errors, can be obtained.

In order to obtain probabilities $P_{(i,j)}$, we apply numerical methods [25] so as to solve the matrix equation given by $\nu = \nu \cdot \mathbf{P}$, where

$$\nu = \left[P_{(0,0)}, \ldots, P_{(0,C)}, P_{(1,0)}, \ldots, P_{(1,C)}, \ldots, P_{(C,0)}, \ldots, P_{(C,C)}\right]$$

is the steady-state probability vector.

5.4.1 Average Number of Users

The average number of PUs and SUs (i.e. average served traffic) is computed as:

$$N_p = \sum\nolimits_{S_{(i,j)} \in \mathcal{S}} i \cdot P_{(i,j)} \,, \tag{5.25}$$

and

$$N_s = \sum\nolimits_{S_{(i,j)} \in \mathcal{S}} j \cdot P_{(i,j)} \,. \tag{5.26}$$

In addition, we can compute the average number of sensed PUs, N_p', by considering $P_{(i,j)}'$, thus leading to:

$$N_p' = \sum\nolimits_{S_{(i,j)} \in \mathcal{S}} i \cdot P_{(i,j)}' \,. \tag{5.27}$$

5.4.2 Blocking Probability

Blocking occurs whenever a new user cannot be assigned a channel given all channels are occupied or thought to be occupied. Subsequently, blocking probability for PUs can be computed from the true steady state probabilities as:

$$P_B^P = \sum\nolimits_{j=0}^{C} P_{(C,j)} \,. \tag{5.28}$$

On the other hand, the blocking probability for SUs can be computed as:

$$P_B^S = \sum\nolimits_{i=0}^{C} \sum\nolimits_{j=C-i}^{C} P_{(i,j)}' \,, \tag{5.29}$$

with $P_{(i,j)}'$ defined in (5.24).

5.4.3 Interruption Probability

Interruption of secondary service occurs whenever a SU is forced to release a channel before its session has ended due to the appearance of a PU and no other channel is sensed to be free. To compute the interruption probability the average number of secondary users defined in (5.26) can be regarded as the average served SU traffic $T_s^{\text{served}} = N_s$. Then, we can express

$$T_s^{\text{served}} = T_s \cdot \left(1 - P_B^S\right) \cdot \left(1 - P_D\right) \,, \tag{5.30}$$

indicating that the served traffic is the offered traffic which is not blocked nor interrupted. From (5.30), we express the interruption probability as:

$$P_D = 1 - \frac{T_s^{\text{served}}}{T_s \left(1 - P_B^S\right)} = 1 - \frac{N_s}{\frac{\lambda_s}{\mu_s} \left(1 - P_B^S\right)} \,. \tag{5.31}$$

5.4.4 Interference Probability: An Upper-Bound

At a given sensing instant in state $S_{(i,j)} \in \mathcal{S}$, the probability of miss-detecting n PUs out of i PUs in the system can be expressed as the binomial distribution with δ the probability of miss-detection (refer to Appendix 5.A). Consequently, the average number of missed detections in state $S_{(i,j)} \in \mathcal{S}$ is given by

$$N_{MD}(i,j) = i \cdot \delta \,. \tag{5.32}$$

The average number of collisions in state $S_{(i,j)} \in \mathcal{S}_{pc}$ (recall from section 5.3.1) may be initially upper-bounded by

$$N_c(i,j) \leq N_{MD}(i,j) \,, \tag{5.33}$$

indicating that at most (i.e. in the worst case) we will have a collision for each missed detection. In addition, this upper-bound can be tightened by considering that the average number of collisions will be also less than the average number of SUs in state $S_{(i,j)}$, i.e. $N_c(i,j) \leq j$. Hence, we have

$$N_c(i,j) \leq \min\left[i \cdot \delta, j\right] \triangleq N_c^*(i,j) \,. \tag{5.34}$$

The number of collisions in state $S_{(i,j)} \in \mathcal{S}_c$ (i.e with $i + j > C$) can be initially lower-bounded as

$$N_c(i,j) \geq (i + j - C) \triangleq \kappa_c \,, \tag{5.35}$$

where we know that at least κ_c channels are being simultaneously shared by both a PU and a SU.

In addition, remaining $(i - \kappa_c)$ PUs and $(j - \kappa_c)$ SUs may be also in a collision situation, hence the number of collisions can be upper-bounded, similar to (5.34), as

$$N_c(i,j) \leq \kappa_c + \min\left[(i - \kappa_c) \cdot \delta, (j - \kappa_c)\right] \triangleq N_c^*(i,j) \,. \tag{5.36}$$

As for the case where $S_{(i,j)} \in \mathcal{S}_{nc}$, the number of collisions is zero, i.e. $N_c(i,j) = 0$.

Given the maximum possible number of collisions would be C, we define the collision probability ratio, hereon the *interference probability*, in state $S_{(i,j)}$ as

$$P_c(i,j) = \frac{N_c^*(i,j)}{C} \,, \tag{5.37}$$

where we have used collision upper-bounds $N_c^*(i,j)$ provided in (5.34) and (5.36) as worst cases, along with $N_c^*(i,j) = 0$ for $S_{(i,j)} \in \mathcal{S}_{nc}$.

Finally, the *average interference probability* can be then expressed as

$$P_c = \sum\nolimits_{S_{(i,j)} \in \mathcal{S}} P_c(i,j) \cdot P_{(i,j)} \,. \tag{5.38}$$

5.4.5 Throughput

In a given state $S_{(i,j)} \in \mathcal{S}$, the number of channels that are being simultaneously used by both a PU and a SU, $N_c^*(i,j)$, has been computed in the previous subsection.

It is considered that a channel being shared by both a PU and a SU does not contribute to throughput; consequently we can define the throughput of PUs or SUs in state $S_{(i,j)} \in \mathcal{S}$ as:

$$\Gamma^p_{(i,j)} = [i - N_c^*(i,j)] \cdot R_p \, , \tag{5.39}$$

along with

$$\Gamma^s_{(i,j)} = [i - N_c^*(i,j)] \cdot R_s \, , \tag{5.40}$$

measured in bits per second (bps), where R_p and R_s are the average bit rate per channel for PUs and SUs respectively. Given that SUs need some time to sense the medium (or to remain silent so that the sensing entity can sense the medium) the net throughput per channel achieved by SUs, R_s, will be lower than the net throughput per channel achieved by PUs, R_p, under our assumption that the same amount of bandwidth is devoted to both users. Then, we can write, $R_s = R_p \cdot \eta_{sens}$ where $\eta_{sens} \in [0,1]$ is the sensing efficiency defined in (5.4). The granted bit rate for a single PU will be simply computed using the Shannon bound as

$$R_p = W \cdot \log_2(1 + \bar{\gamma}) \, , \tag{5.41}$$

where W (Hz) is the bandwidth of a single channel and $\bar{\gamma}$ (dB) is the average signal-to-noise ratio (SNR).

Finally, the average throughput for both PUs and SUs is computed as

$$\Gamma^p = \sum\nolimits_{S_{(i,j)} \in \mathcal{S}} \Gamma^p_{(i,j)} \cdot P_{(i,j)} \, , \tag{5.42}$$

along with

$$\Gamma^s = \sum\nolimits_{S_{(i,j)} \in \mathcal{S}} \Gamma^s_{(i,j)} \cdot P_{(i,j)} \, . \tag{5.43}$$

5.5 Performance Evaluation

In the following, the considered parameter setup for the numerical evaluation and also some implementation aspects for the uncoordinated spectrum awareness case are introduced. Subsequently, numerical results address, in the first place, the model validation by means of a system-level simulator. Secondly, numerical results will be given so as to capture the tradeoff between the time devoted to sensing and the throughput, the impact on the number of considered channels (C) and, finally, the effect on the spectrum awareness periodicity ΔT.

5.5.1 Parameter Setup

As an initial reference, it is considered a total bandwidth partitioned into $C = 16$ channels to be shared amongst PUs and SUs. The offered primary traffic load is $T_p = \lambda_p/\mu_p = \{5, 10\}$ Erlangs (E), while secondary offered traffic, defined as $T_s = \lambda_s/\mu_s$, will span over a specified range of values. Note that for $T_p = 10$E and $C = 16$ channels the system is significantly loaded by PUs, so that the system is evaluated under rather unfavorable conditions for SUs. Average session duration is assumed to be equal for both PUs and SUs with $1/\mu = 1/\mu_p = 1/\mu_s = 120$ seconds (i.e. assuming a time-based service). Sensing periodicity is, unless otherwise stated, $\Delta T = 100$ms. The characterization of sensing errors will be provided by means of missed-detection and false-alarm probabilities, i.e. δ and ε respectively (the following subsection addresses such implementation in detail). In addition, for comparative purposes, we will also consider the case where perfect sensing (error-free) is available. For this case, not only δ and ε are zero but also the time devoted to sensing purposes is considered to be zero.

5.5.2 Spectrum Detection in Rayleigh Fading

The detection of unknown signals by means of sensing has captured a lot of attention in the past, and the advent of cognitive radio has indeed contributed to increase the work devoted to this matter [21, 22].

In general, sensing in the presence of errors due to imperfect channel conditions can be characterized through miss-detection and false-alarm probabilities (i.e. δ and ε respectively) extracted from so-called Receiver Operating Curves (ROC) which relate both magnitudes. These curves mainly depend on the considered channel model (Rayleigh, Rician, etc.), the average SNR ($\bar{\gamma}$) at the sensor's end and the time-bandwidth product (m) previously introduced in (5.3). Fig. 5.3 shows the ROC curves for the energy detector under Rayleigh fading considering a number of time-bandwidth products (m) which have been computed according to analytical expressions in [21].

Note that, for a fixed observed bandwidth W, the increase of the time devoted to sensing, T, i.e. also meaning an increase in the time-bandwidth product (recall that $m = T \cdot W$), will mean that a more accurate sensing information is retrieved. In this sense, setting different target values for the miss-detection probability δ^* we obtain the corresponding values for the false alarm probability ε^* for the different time-bandwidth products m which define the working point of the receiver. In

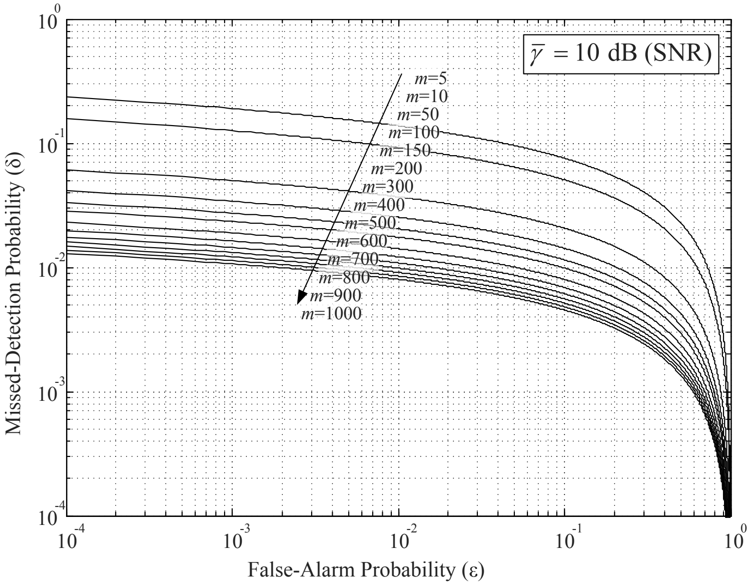


Figure 5.3: Receiver Operating Characteristic (ROC) curves for the energy detector in Rayleigh fading considering $\bar{\gamma} = 10$dB and time-bandwidth products m between 5 and 1000.

Table 5.1: ROC Values for $\bar{\gamma} = 10$dB and $W = 200$kHz extracted from Fig. 5.3

.

m	δ^*	ε^*	$T = m/W$
5	0.01	0.7277	0.000025
10	0.01	0.6278	0.00005
50	0.01	0.3378	0.00025
100	0.01	0.2073	0.0005
150	0.01	0.1391	0.00075
200	0.01	0.0974	0.001
300	0.01	0.0511	0.0015
400	0.01	0.0281	0.002

practice, this is achieved by setting appropriate values for the decision threshold at the sensor's end [21]. Without loss of generality we assume a fixed value of $W = 200$kHz and a target miss-detection probability value of $\delta^* = 0.01$. Unless otherwise stated the average SNR is chosen to be $\bar{\gamma} = 10$dB. Then, and for the sake of representation, performance results will refer to particular values of m (which in turn refer to specific values of δ^* and ε^* as shown in Table 5.1).

5.5.3 Numerical Results

5.5.3.1 Validation and Preliminary Model Assessment

One main objective of analytical system modeling is to retain the basic interactions between the main parameters affecting the performance of a real system. How close is the performance behavior of the model with respect to the real system will largely depend on the considered model assumptions and hypothesis. In this sense, in this section we intend to identify those parameters that may limit the applicability of the model, which is, on the other hand, inherent to all analytical models.

For the sake of a complete model validation, a system-level simulator has been developed so as to extract relevant metrics and compare them to those defined for the model in Section 5.4. Specifically, this simulator operates on a discrete-time basis and is somewhat more realistic than the model in a way that it is not constrained by hypothesis 1 and 2 given in Section 5.3.4.

In order to assess the validation of the proposed model we compare the steady state probabilities obtained through the Markov chain with the computed values through simulation. Since the desired metrics (see Section 5.4) are directly computed from the steady state probabilities, $P_{(i,j)}$ and $P'_{(i,j)}$, a positive validation at this point results in a good indicator about the validity of the proposed model. In this respect, Fig. 5.4 plots the steady state probabilities resulting from the Markov model against the same probabilities computed by means of simulation. In addition it is also plotted the curve $y = x$ for reference purposes (note that for a perfect match, plotted data in Fig. 5.4 should lay over the curve $y = x$). In addition, log-log scaling enables a detailed comparison between both magnitudes. It can be observed that a good match between the model and the simulation is attained, in particular for higher probability values (i.e. in the range between 10^{-2} and 10^{-1}). For probabilities below this range, differences become more evident indicating that a larger number of samples (i.e. larger simulation times) are needed in order to obtain a good statistics of simulated data.

Fig. 5.5 and Fig. 5.6 show the average number of users and the interruption probability, as defined in (5.25), (5.26) and (5.31) respectively, where simulation results appear as circles and the DTMC model performance appear as lines. At a first glance, note how the system-level simulator values (i.e. circles) closely match those obtained by means of the DTMC model.

Regarding Fig. 5.5, when spectrum sensing is subject to errors, performance evaluation is shown for different considered values of δ and ε through specified values

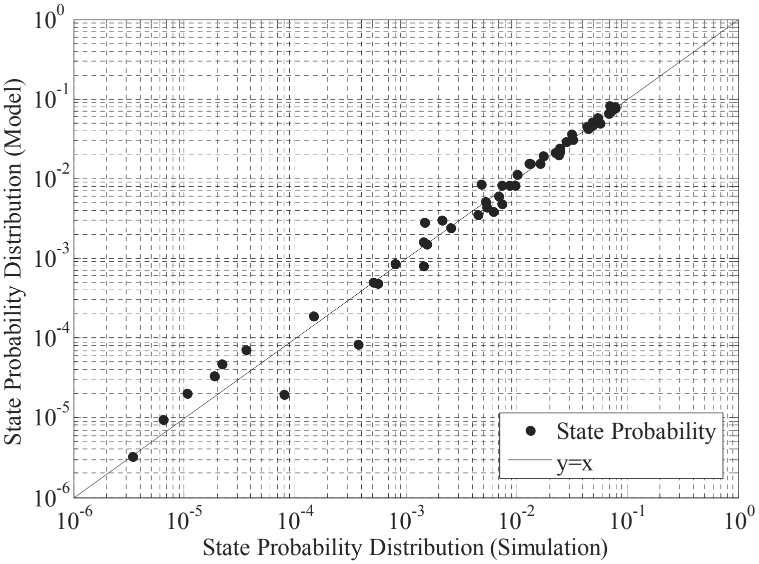


Figure 5.4: Model validation of the steady state probabilities by plotting the obtained simulation value against the theoretical (model) value. Perfect match implies data laying on y=x curve.

of m (regard Table 5.1). Provided that PUs have spectrum access priority over SUs, the average number of served PUs remains constant (note that we have fixed $T_p = 10$E). As for the average number of SUs, as expected, the better the spectrum sensing information (i.e. higher m values) the better spectrum opportunities can be exploited. For the case of perfect sensing the highest utilization of free bands is attained by SUs.

Fig. 5.6 shows the interruption probability (i.e. the probability that a SU is forced to suspend its session due to primary activity) defined in (5.31) for the different values of m. While the perfect-sensing case exhibits the most favorable behavior, a decrease in m results into higher false-alarm probabilities which in turn triggers SUs to release occupied channels. In addition, the interruption probability increases with the offered secondary load, which is, on the other hand, somewhat expected.

Additionally, it is necessary to determine when hypotheses 1 and 2 (provided in Section 5.3.4) pose some risk on the model's validity. Indeed, such hypotheses rely in one way or another on the relation of arrival and departure process parameters with the DTMC observation period ΔT. As for hypothesis 1, the relation between $1/\mu$ and ΔT will determine if a call/session arriving within a period ΔT will depart in the same ΔT. On the other hand, the longer the period ΔT the more important becomes the order in which arrival and departures occur, thus hypothesis 2 will be less true as ΔT increases. To that end, the behavior of the average number of

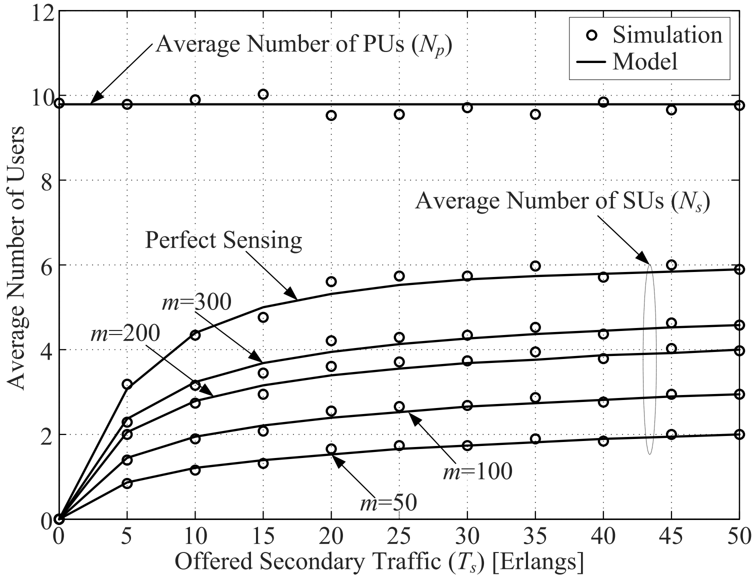


Figure 5.5: Average number of users against offered secondary traffic load (T_s). The offered PU load is $T_p = 10$E.

users and the interruption probability against several $\Delta T \cdot \mu$ values is presented in Fig. 5.7(a) and Fig. 5.7(b) respectively. The offered traffic is fixed to $T_p = 5$E and $T_s = 20$E. It can be observed [see Fig. 5.7(b)] that the interruption probability decreases when ΔT values increase. Not surprisingly, very frequent tests on the occupancy of a channel (i.e. low ΔT values) will translate into higher chances that a SU is refrained from transmitting specially for the cases that the false-alarm probability is significant (i.e. low m values). Consequently, see Fig. 5.7(a), the number of secondary users increases with ΔT.

Note that for $\Delta T \cdot \mu$ values over $2 \cdot 10^{-2}$, the model and the simulated data start to drift apart. This drift is more noticeable in the interruption probability case, see shaded region in Fig. 5.7(b), than for the average number of users. Nevertheless, for values of $\Delta T \cdot \mu$ which remain in the range 10^{-4} to 10^{-2}, the model is able to capture the behavior exhibited by the system-level simulator, thus validating the proposed hypothesis in these cases. Note, that this dependency between the model validation and ΔT was somewhat expected when expressing hypotheses 1 and 2. In terms of model applicability, practical system parameter values can be evaluated bearing in mind the relation between ΔT and μ.

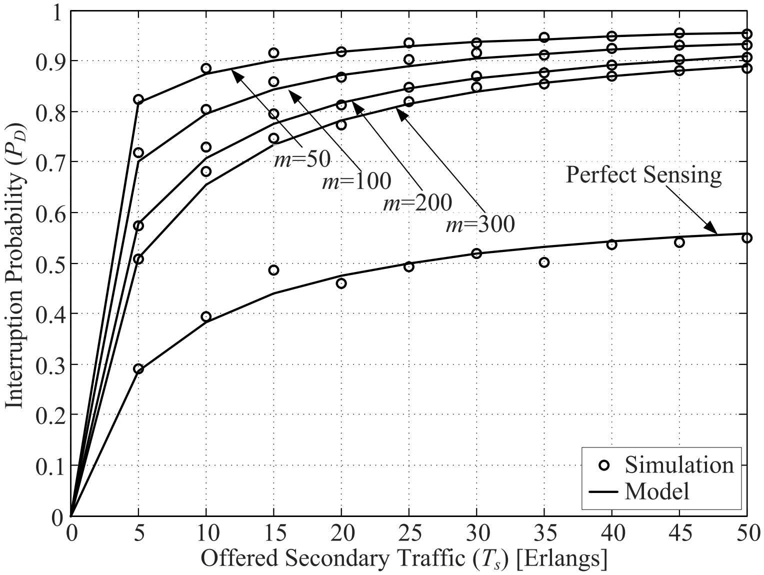


Figure 5.6: Average number of users against offered secondary traffic load (T_s). The offered PU load is $T_p = 10$E.

5.5.3.2 Sensing Time / Throughput Trade-Off

Fig. 5.8 shows the existing trade-off between the time-bandwidth product (m) and the throughput experienced by SUs for several offered secondary traffic loads (T_s) and average SNRs ($\bar{\gamma}$). It can be seen that following an initial throughput increase due to accurate sensing information, throughput degradation starts to rise when the sensing time increases to values that are high compared to the time devoted to data transmission (i.e. the sensing efficiency $\eta_{sens} \to 0$). The extreme case is when all transmission time is devoted to sensing purposes, in this case when $T_{sens} = \Delta T$ then $\eta_{sens} = 0$ which occurs for $m = 1250$ considering $\Delta T = 0.1$s. For the particular study case in Fig. 5.8, several optimum values for the time-bandwidth product can be selected in order to maximize the average secondary throughput. These values are represented by m^* in Fig. 5.8. Note that increasing the secondary offered load requires a better knowledge of spectrum occupancy, thus increasing the time-bandwidth product is convenient. In the same way, higher SNR values require less sensing time, i.e. lower m values, to attain acceptable accuracy in terms of missed-detection and false-alarm probabilities. In addition, a decrease in the average SNR results in a dramatic decrease of the average throughput as could be expected by regarding (5.41).

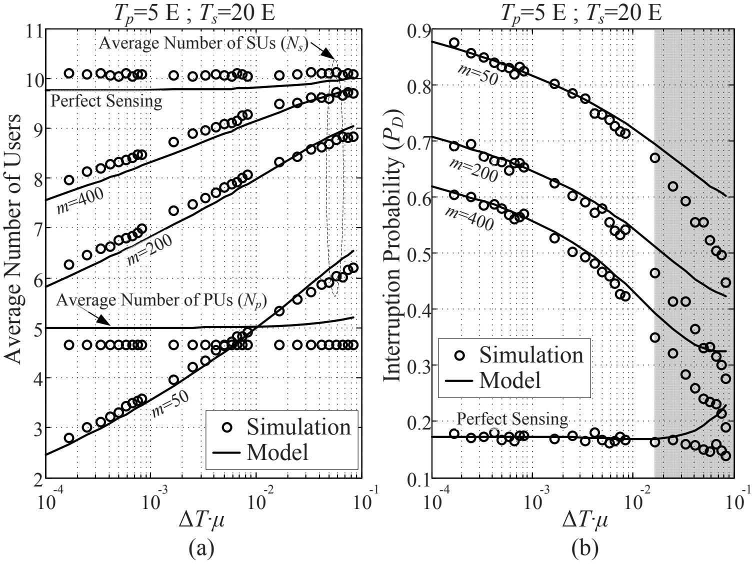


Figure 5.7: The impact of $\Delta T \cdot \mu$ in model validation. (a) Average number of users and (b) Interruption probability.

5.5.3.3 Channel number impact

Fig. 5.9 shows the impact of the number of total channels (C) on the experienced secondary throughput for two different values of ΔT and constant offered traffic load of $T_p = T_s = 5$E along with average SNR of $\bar{\gamma} = 10$dB. In Fig. 5.9(a), for $\Delta T = 0.1$s, note that between 2 and 10 channels, the better the sensing accuracy (i.e. higher m) the higher the number of SUs are getting assigned, thus higher throughput is attained. However, given that sensing requires a time $T_{sens} = C \cdot T$ [see (5.5)]; as the number of channels increases, high sensing accuracy (i.e. high m which implies also high T values) does not payoff the increased time devoted to sensing and its consequences on throughput [as already observed in subsection 5.5.3.2)]. Then, it is observed in Fig. 5.9(a) that at some value of C it is better to reduce the sensing accuracy, i.e. decrease m, which in turn produces higher false alarm. The grey shadowed zones indicate regions where a suitable time-bandwidth product (denoted as m^*) maximizes average secondary throughput. Given that the offered load is constant and more channels are available by rising C, the increased false-alarm as a result of lowering m remains bearable. On the other hand, in Fig. 5.9(b), for $\Delta T = 5$s we have that the time devoted to sensing can be larger in this case with no observed throughput degradation. This can be followed by observing the sensing efficiency in (5.4), where if ΔT increases, we may tolerate higher T_{sens} values so that the sensing efficiency is still acceptable. Then, for

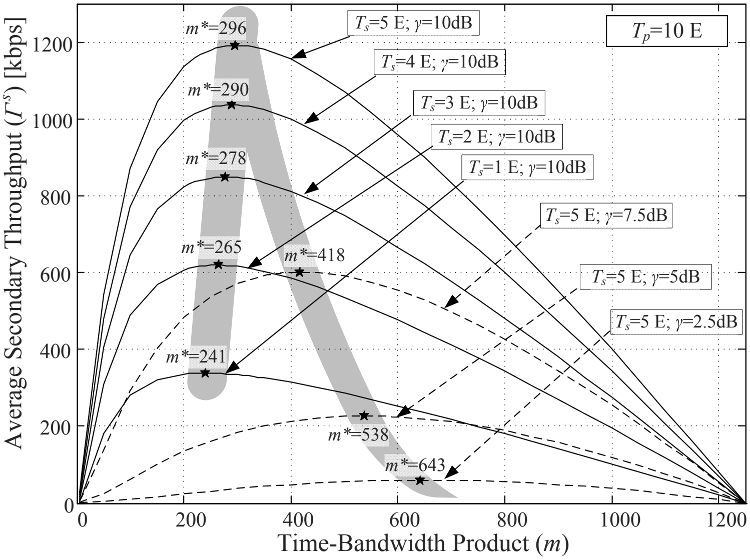


Figure 5.8: Secondary throughput against the time-bandwidth product (m) for different traffic (solid) and SNR (dashed) conditions. Shaded regions emphasize optimum m values (m^*) for which secondary throughput is maximum.

$\Delta T = 5$s, accurate sensing does payoff the time devoted to sensing procedures and, thus, better throughput is observed for the case of higher values of m.

5.5.3.4 Spectrum Awareness Periodicity

Fig. 5.10 shows the impact of spectrum occupancy information periodicity ΔT in terms of interference probability (P_c) as defined in Section 5.4.4. It is again assumed that $C = 16$ channels are available. Results indicate that high values of ΔT cause the secondary system to take decisions with out-of-date primary spectrum occupancy information which translates into higher interference probabilities, i.e. higher chances that a PU and a SU are assigned the same channel. In this sense, given high time-bandwidth products maximize spectrum utilization (see Fig. 5.5), higher chances that SUs interfere with PUs arise. On the contrary, for low m values the higher false-alarm probability prevents from assigning SUs thus less interference is observed as opposed to higher m values.

The impact of interference on primary user throughput, as revealed by (5.39), is given in Fig. 5.11(a), where primary throughput benefits from high false-alarm (i.e. low m values) since interference is lowered as already shown in Fig. 5.5.

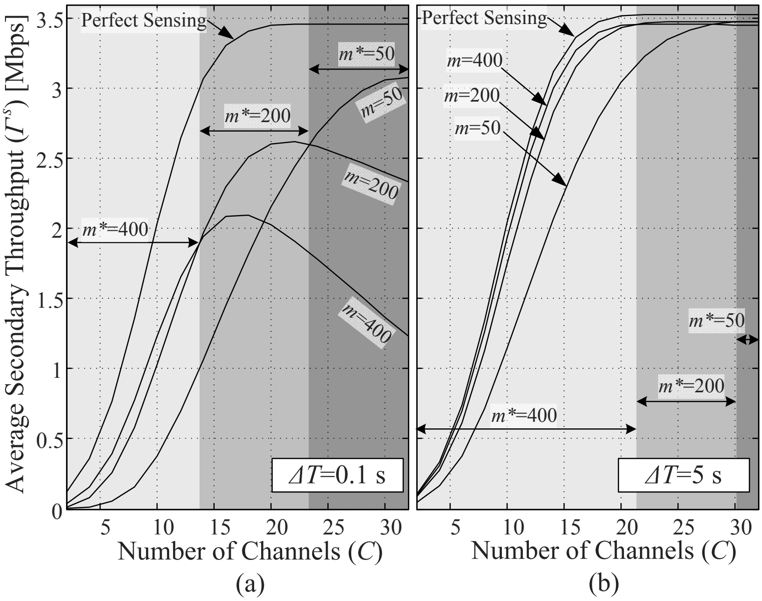


Figure 5.9: Secondary throughput against number of channels (C) for (a) $\Delta T = 0.1$s and (b) $\Delta T = 5$s. Shadowed regions indicate the suitability of particular time-bandwidth products (m^*) such that throughput is maximized.

As for the secondary throughput, Fig. 5.11(b), the sensing periodicity ΔT conditions the time devoted to sensing purposes (T_{sens}) and, hence, the time-bandwidth product. Then, for low ΔT values, large time-bandwidth products causes sensing efficiency to decrease and thus reduced throughput is attained. On the contrary, increased ΔT values allow a longer sensing period and consequently better spectrum awareness which in turn improves secondary spectrum usage and throughput. Shaded regions in Fig. 5.11(b) reflect the suitable time-bandwidth values (m^*), among those considered, indicating the existing trade-off between sensing efficiency and spectrum awareness quality.

5.6 Chapter Summary

In this chapter a generalized and flexible framework for the definition and evaluation of opportunistic shared spectrum scenarios has been presented. This framework is capable of supporting a wide range of implementation possibilities and functionalities. In this sense, the suitability of a DTMC model as the core of the framework has been suggested and further justified. The DTMC model has been formulated with a high degree of generality and some performance metrics

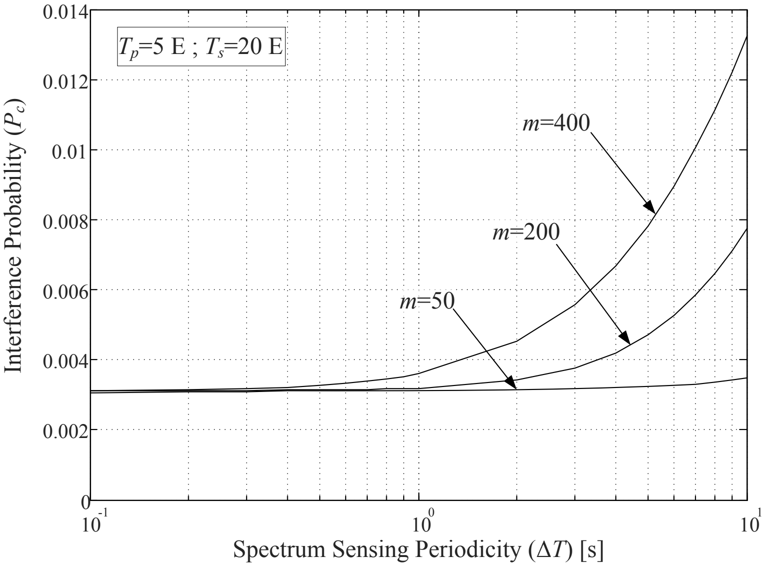


Figure 5.10: Interference probability against sensing periodicity (ΔT) when $T_p = 5E$ and $T_s = 20E$ for several time-bandwidth products (m).

extracted. An uncoordinated operation between primary and secondary networks has been assumed where primary spectrum occupancy information is retrieved through sensing mechanisms. A first goal was to determine the validity of the proposed model which has been assessed by means of its comparison with a system-level simulator. In addition, the limitations of the model have also been determined and the parameters influencing such limitation have been identified. It was shown that, for a correct model operation, the value $\Delta T \cdot \mu$ should be appropriately chosen. Consequently, practical system parameters values can be evaluated bearing in mind such limitations. In this case, in our particular scenario, values of $\Delta T \cdot \mu$ below $2 \cdot 10^{-2}$ are suggested in order to achieve a good match between the model and the simulations.

The existing tradeoff between the sensing accuracy and the exhibited secondary throughput has also been studied. As expected, an increased sensing accuracy through longer sensing periods will, at a given point, not payoff the degradation obtained in terms of throughput since less time is then devoted to the actual data transmission. Results revealed that the sensing time (equivalently, the time-bandwidth product) can be conveniently adjusted in order to maximize throughput. In our numerical analysis we observed how for decreasing offered secondary traffic loads (from T_s = 5E to T_s = 1E) the sensing accuracy can be slightly decreased (from $m^* = 296$ to $m^* = 241$) in order to favor secondary throughput. In addition, an improved SNR condition allows a decrease in sensing time since energy detection

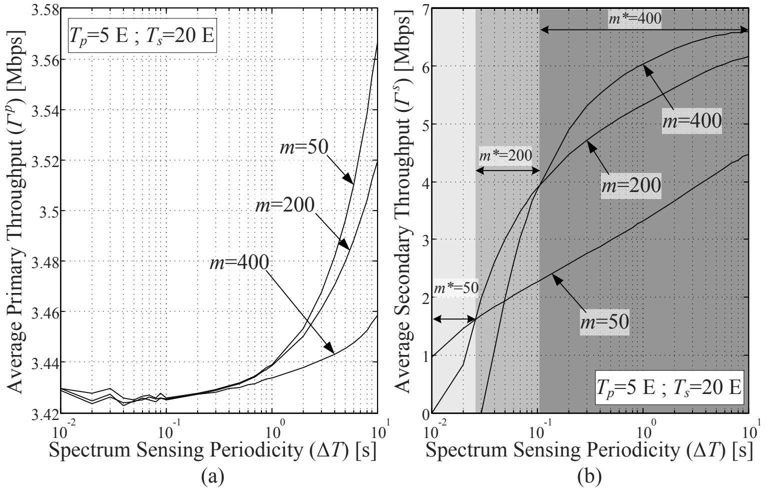


Figure 5.11: Primary (a) and secondary (b) throughput against ΔT when $T_p = 5\text{E}$ and $T_s = 20\text{E}$ for several time-bandwidth products (m).

is better. In this case, suitable time-bandwidth values span from $m^* = 296$ (for $\bar{\gamma} = 10\text{dB}$) to $m^* = 643$ (for $\bar{\gamma} = 2.5\text{dB}$). In addition, if the number of channels to be sensed is large, sensing procedures will take longer to determine the spectrum occupancy of the whole band, consequently reducing the sensing efficiency. Then, the time-bandwidth product should be reduced when increasing the bandwidth on which PUs and SUs operate. The impact of the spectrum awareness periodicity ΔT has also been evaluated. As expected, the longer the time between sensing instants (i.e. ΔT) the higher the chances of collision events between PUs and SUs happen. In addition, improved sensing accuracy degrades the experienced interference by allowing an increased number of SUs in the system. As shown, secondary operation can be optimized by choosing adequate values for the time-bandwidth product (i.e. the time devoted to sensing) in such way that the throughput is maximized.

Future work, addressed in Chapter 6, will be devoted to extend the presented model to include a flexible bandwidth partition scheme in which PUs and SUs do not necessarily occupy the same amount of spectrum. In addition, bandwidth requirements could be dynamically adjusted according to specific demands which may vary over time. Furthermore, besides the considered time-based service type of SUs in this chapter, a volume-based service type which intends to transmit a certain volume of data will be also considered.

5.A Proof of Theorem 1

Given a total number of C channels we aim to compute the probability of having n detected PUs when actually i PUs are assigned. Let N_{FA} and N_{MD} represent the total number of false alarms and missed detections after scanning the C channels. The number of possible miss detections, N_{MD}, will be consequently in the range $0 \leq N_{MD} \leq i$, i.e. we can not miss-detect more than the number of "true" assigned PUs. In the same way, the number of false alarms is in the range $0 \leq N_{FA} \leq C - i$.

On the other hand, we can express the number of detected PUs, n, as

$$n = i + N_{FA} - N_{MD}\,, \tag{5.44}$$

indicating that the total number of detected PUs are those actually assigned (i) plus those we believe are assigned (N_{FA}) minus those we have missed (N_{MD}).

From (5.44), we may re-write

$$N_{FA} = N_{MD} + n - i\,, \tag{5.45}$$

where if $N_{FA} \geq 0$ it follows that $N_{MD} \geq i - n$. Then, given that also $N_{MD} \geq 0$, we have that $N_{MD} \geq \max(0, i - n)$. On the other hand, from $0 \leq N_{FA} \leq C - i$ and (5.45), we have that $0 \leq N_{MD} + n - i \leq C - i$, thus $N_{MD} \leq C - n$. Provided $0 \leq N_{MD} \leq i$ must be also satisfied we finally obtain $N_{MD} \leq \min(i, C - n)$.

The probability of detecting n PUs given i PUs are assigned can be expressed using the total probability definition for conditional probabilities as

$$b_{(n,i)} = \sum_{N_{MD}} \Pr(N_{FA} = N_{MD} + n - i | N_{MD}) \Pr(N_{MD})\,, \tag{5.46}$$

where, by considering binomial distributions for both false-alarm and miss-detection, we have

$$\begin{array}{rcl} \Pr(N_{FA} & = & N_{MD} + n - i | N_{MD}) \\ & = & \left(\begin{array}{c} C - i \\ N_{MD} + n - i \end{array} \right) \varepsilon^{N_{MD}+n-i} (1-\varepsilon)^{C-i-N_{MD}-n+i} \\ & = & \left(\begin{array}{c} C - i \\ N_{MD} + n - i \end{array} \right) \varepsilon^{N_{MD}+n-i} (1-\varepsilon)^{C-N_{MD}-n} \end{array}, \tag{5.47}$$

along with

$$\Pr(N_{MD}) = \left(\begin{array}{c} i \\ N_{MD} \end{array} \right) \delta^{N_{MD}} (1-\delta)^{i-N_{MD}}\,, \tag{5.48}$$

from which expression (5.11) can be obtained.

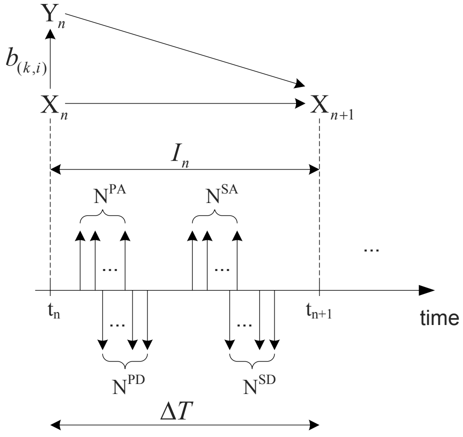


Figure 5.12: Spectrum assignments/de-assignments due to arrival/departure of PUs and SUs.

5.B Proof of Lemmas 1, 2, 3, and 4

The number of PU/SU spectrum assignments and de-assignments in I_n, $N_a \in \left\{N_a^P, N_a^S\right\}$ and $N_d \in \left\{N_d^P, N_d^S\right\}$, will depend on the spectrum occupancy given by the true or detected states at time t_n, i.e. $\mathbf{X}_n$ or $\mathbf{Y}_n$, and on the number of arrivals N^A and departures N^D in time interval I_n. These number of arrivals and departures will eventually lead to a number of channel assignments and de-assignments depending on the true or detected spectrum occupancy at time t_n. Fig. 5.12 illustrates the arrival and departure process which conditions the state transition probabilities.

In the following, expressions for lemmas 1 to 4 are derived.

5.B.1 Proof of Lemma 1

Let the true state be $\mathbf{X}_n = S_{(i,j)}$; we intend to compute the probability of assigning $N_a^P = k$ PUs in I_n given we have $N_d^P = l \leq i$ PU de-assignments in I_n, $a^P_{(i,j,k,l)}$. In words, it is the probability of assigning exactly k PUs in the case that least k channels are available (i.e. having exactly k PU arrivals), and the probability of having at least k PU arrivals if exactly k channels are available. Thus, we may write:

$$\begin{aligned} a^P_{(i,j,k,l)} &= \Pr[N_a^P = k|\mathbf{X}_n = S_{(i,j)}, N_d^P = l] \\ &= \begin{cases} \Pr[N^{PA} = k] = P_k^{PA}, & \text{if } i-l+k < C \\ \Pr[N^{PA} \geq k] = 1 - \sum\limits_{m=0}^{k-1} P_m^{PA}, & \text{if } i-l+k = C \end{cases}, \end{aligned} \tag{5.49}$$

with P_k^{PA} given in (5.12). For the case of assigning more PUs than available channels the probability in (5.49) is zero. Implicitly in (5.49) we consider that k PU assignments are made upon l PU de-assignments, thus using the assumption of disregarding the order in which arrivals and departures occur in I_n which will be an assumed hypothesis in our model (see Hypothesis 2 in Section 5.3.4).

5.B.2 Proof of Lemma 2

Again, let the true state be $\mathbf{X}_n = S_{(i,j)}$; the probability of de-assigning $N_d^P = k$ PUs in I_n, where $0 \leq k \leq i$ (i.e. we can only de-assign those already assigned prior to t_n), depends on the number of PU departures in I_n:

$$d^P_{(i,j,k)} = \Pr[N_d^P = k | \mathbf{X}_n = S_{(i,j)}] = \Pr[N^{PD} = k] = P_k^{PD}\,, \tag{5.50}$$

with P_k^{PD} given in (5.14). Note that we have made use of the assumption that a new arrival in I_n will not depart in I_n by specifying that the number of de-assignments is bounded as $0 \leq k \leq i$ in I_n. For any other value of k, the probability in (5.50) is zero.

5.B.3 Proof of Lemma 3

Let the true state be $\mathbf{X}_n = S_{(i,j)}$; the probability of assigning $N_a^S = k > 0$ SUs in I_n given we have $N_d^S = l \leq j$ SU de-assignments in I_n, $a^S_{(i,j,k,l)}$, will depend on the detected state at t_n, $\mathbf{Y}_n = S_{(m,j)}$ and on the number of SU arrivals as:

$$\begin{aligned} a^S_{(i,j,k,l)} &= \Pr[N_a^S = k | \mathbf{X}_n = S_{(i,j)}, N_d^S = l] \\ &= \sum_{m=0}^{C-k-j+l} \Pr[N_a^S = k | \mathbf{Y}_n = S_{(m,j)}, N_d^S = l] \cdot b_{(m,i)} \\ &= \sum_{m=0}^{C-k-j+l} \bar{a}^S_{(m,j,k,l)} \cdot b_{(m,i)}\,, \end{aligned} \tag{5.51}$$

where the total probability formula has been used to relate the true state $\mathbf{X}_n = S_{(i,j)}$ with the detected state $\mathbf{Y}_n = S_{(m,j)}$ through probabilities $b_{(k,i)}$. In particular, (5.51) states that $N_a^S = k$ SUs will be assigned provided the detected number of PUs, m, fulfills $m + j + k - l \leq C$, i.e., there are at least k detected free channels for secondary use provided that we also have l SU de-assignments. In addition, the number of k SU assignments in state $\mathbf{X}_n = S_{(i,j)}$ is bounded by $0 < k \leq C-i-j+l$, omitting the case $k = 0$ which will be treated separately. Finally, $\bar{a}^S_{(m,j,k,l)}$ in (5.51) is obtained similar to (5.49) as:

$$\bar{a}^S_{(m,j,k,l)} = \begin{cases} P^{SA}_k, & \text{if } m + j - l + k < C \\ 1 - \sum\limits_{r=0}^{k-1} P^{SA}_r, & \text{if } m + j - l + k = C \end{cases} . \tag{5.52}$$

For the specific case of no SU assignments (i.e. $k = 0$), the probability of assigning $k = 0$ SUs is the probability of assigning $k = 0$ SUs when there is at least one free detected channel or the probability that there are no detected free channels, i.e.:

$$\begin{aligned} a^S_{(i,j,0,l)} &= \Pr[N^S_a = 0|\mathbf{X}_n = S_{(i,j)}, N^S_d = l] = \sum_{m=0}^{C-j+l} \bar{a}^S_{(m,j,0,l)} \cdot b_{(m,i)} \\ &+ \sum_{m=C-j+l}^{C} b_{(m,i)} \, . \end{aligned} \tag{5.53}$$

Combining (5.51) with (5.53), along with the definition in (5.52), expression (5.18) is obtained.

5.B.4 Proof of Lemma 4

As for the de-assignment processes of SUs, there are mainly three independent events which imply an SU de-assignment: in the first place, a number of $N^{S,S}_d$ SUs may be de-assigned provided detection at time t_n determines that there are $N^{S,S}_d$ SUs sharing the same channel with PUs. Secondly, a number of $N^{S,SC}_d$ SUs may be de-assigned in I_n simply because their sessions have ended (here, SC stands for Service Completion).

Then, let the true state be $\mathbf{X}_n = S_{(i,j)}$; the probability of de-assigning $N^{S,S}_d = k$ SUs in I_n due to detection of state $\mathbf{Y}_n = S_{(m,j)}$ at time t_n, given the number of de-assignments due to service completion is $N^{S,SC}_d = l$, is given by:

$$\Pr[N^{S,S}_d = k|\mathbf{X}_n = S_{(i,j)}, N^{S,SC}_d = l] = \Pr[m+j-l = C+k] = b_{(C+k-j+l,i)} \, , \tag{5.54}$$

provided that $0 < k \leq j - l$. Accordingly, the probability of no SU de-assignments due to detection of state $\mathbf{Y}_n = S_{(m,j)}$ is:

$$\Pr[N^{S,S}_d = 0|\mathbf{X}_n = S_{(i,j)}, N^{S,SC}_d = l] = 1 - \sum_{k=1}^{j-l} b_{(C+k-j+l,i)} \, . \tag{5.55}$$

Then, from (5.54) and (5.55), we can write:

$$\begin{aligned} d^{S,S}_{(i,j,k,l)} &= \Pr[N^{S,S}_d = k|\mathbf{X}_n = S_{(i,j)}, N^{S,SC}_d = l] \\ &= \begin{cases} b_{(C+k-j+l,i)} & \text{if } 0 < k \leq j-l \\ 1 - \sum_{r=1}^{j-l} b_{(C+r-j+l,i)} & \text{if } k = 0 \end{cases} . \end{aligned} \tag{5.56}$$

On the other hand, the probability of de-assigning k SUs in I_n due service completions is given by [similar to (5.50)]:

$$d^{S,SC}_{(i,j,k)} = \Pr[N^{S,SC}_d = k|\mathbf{X}_n = S_{(i,j)}] = \Pr[N^{SD} = k] = P^{SD}_k \, . \tag{5.57}$$

Finally, we can express the global probability of de-assigning k SUs in I_n (i.e. without specifying if the de-assignment is due to detection or due to session completion) as:

$$d^{S}_{(i,j,k)} = \Pr[N^{S}_d = k|\mathbf{X}_n = S_{(i,j)}] = \sum_{r=0}^{k} d^{S,S}_{(i,j,k-r,r)} \cdot d^{S,SC}_{(i,j,r)} \, . \tag{5.58}$$

The above expressions proof Lemma 4.

5.C Proof of Theorem 2

In this appendix, a detailed derivation of all possible transition probabilities in the DTMC model is presented.

5.C.1 Transition $S_{(i,j)} \to S_{(i+N,j)}$ with $0 < N \leq C - i$

The probability associated to transition $S_{(i,j)} \to S_{(i+N,j)}$, with $0 < N \leq C - i$, is the probability to have N more PU assignments in than PU de-assignments in I_n; and equal number of SU assignments and de-assignments in I_n. Then, applying conditional independence between PU and SU processes, we can write:

$$P_{(i+N,j|i,j)} = \Pr[N^P_a - N^P_d = N|i,j] \cdot \Pr[N^S_a = N^S_d|i,j] \, , \tag{5.59}$$

where multiplicative probabilities in (5.59) are given by

$$\begin{aligned}
\Pr[N_a^P - N_d^P = N] &= \Pr[N_a^P = N | N_d^P = 0] \cdot \Pr[N_d^P = 0] \\
&+ \Pr[N_a^P = N+1 | N_d^P = 1] \cdot \Pr[N_d^P = 1] \\
&+ \Pr[N_a^P = N+2 | N_d^P = 2] \cdot \Pr[N_d^P = 2] \\
&+ \cdots + \Pr[N_a^P = N+i | N_d^P = i] \cdot \Pr[N_d^P = i] \\
&= a^P_{(i,j,N,0)} \cdot d^P_{(i,j,0)} + a^P_{(i,j,N+1,1)} \cdot d^P_{(i,j,1)} \\
&+ a^P_{(i,j,N+2,2)} \cdot d^P_{(i,j,2)} + \cdots + a^P_{(i,j,N+i,i)} \cdot d^P_{(i,j,i)} \\
&= \left(\sum_{k=0}^{i} a^P_{(i,j,N+k,k)} \cdot d^P_{(i,j,k)} \right),
\end{aligned} \tag{5.60}$$

along with

$$\begin{aligned}
\Pr[N_a^S = N_d^S] &= \Pr[N_a^S = 0 | N_d^S = 0] \Pr[N_d^S = 0] \\
&+ \Pr[N_a^S = 1 | N_d^S = 1] \Pr[N_d^S = 1] \\
&+ \Pr[N_a^S = 2 | N_d^S = 2] \Pr[N_d^S = 2] \\
&+ \cdots + \Pr[N_a^S = j | N_d^S = j] \Pr[N_d^S = j] = a^S_{(i,j,0,0)} \cdot d^S_{(i,j,0)} \\
&+ a^S_{(i,j,1,1)} \cdot d^S_{(i,j,1)} + a^S_{(i,j,2,2)} \cdot d^S_{(i,j,2)} + ... + a^S_{(i,j,j,j)} \cdot d^S_{(i,j,j)} \\
&= \left(\sum_{k=0}^{j} a^S_{(i,j,k,k)} \cdot d^S_{(i,j,k)} \right).
\end{aligned} \tag{5.61}$$

Note that, for the sake of notation relief, we have intentionally omitted the conditioning to i and j stated in (5.59).

Then, by replacing (5.60) and (5.61) in (5.59), we may write

$$P_{(i+N,j|i,j)} = \left(\sum_{k=0}^{i} a^P_{(i,j,N+k,k)} \cdot d^P_{(i,j,k)} \right) \cdot \left(\sum_{k=0}^{j} a^S_{(i,j,k,k)} \cdot d^S_{(i,j,k)} \right). \tag{5.62}$$

5.C.2 Transition $S_{(i,j)} \to S_{(i-N,j)}$ with $0 < N \leq i$

The probability associated to transition $S_{(i,j)} \to S_{(i-N,j)}$, with $0 < N \leq i$, is the probability to have N more PU de-assignments than PU assignments in I_n; and equal number of SU assignments and de-assignments in I_n. Then, we have

$$P_{(i-N,j|i,j)} = \Pr[N_d^P - N_a^P = N | i,j] \cdot \Pr[N_a^S = N_d^S | i,j], \tag{5.63}$$

where the first multiplicative term in (5.63) is given by[1]

[1] where conditioning to i and j has been intentionally dropped.

$$\begin{aligned}
\Pr[N_d^P - N_a^P = N] &= \Pr[N_d^P = N] \cdot \Pr[N_a^P = 0 | N_d^P = N] \\
&+ \Pr[N_d^P = N+1] \cdot \Pr[N_a^P = 1 | N_d^P = N+1] \\
&+ \Pr[N_d^P = N+2] \cdot \Pr[N_a^P = 2 | N_d^P = N+2] \\
&+ \cdots + \Pr[N_d^P = i] \cdot \Pr[N_a^P = i-N | N_d^P = i] \\
&= d^P_{(i,j,N)} \cdot a^P_{(i,j,0,N)} + d^P_{(i,j,N+1)} \cdot a^P_{(i,j,1,N+1)} \\
&+ d^P_{(i,j,N+2)} \cdot a^P_{(i,j,2,N+2)} + \cdots + d^P_{(i,j,i)} \cdot a^P_{(i,j,i-N,i)} \\
&= \left(\sum_{k=0}^{i-N} d^P_{(i,j,N+k)} \cdot a^P_{(i,j,k,N+k)} \right) ,
\end{aligned} \tag{5.64}$$

and, with $\Pr[N_a^S = N_d^S | i, j]$ given in (5.61), we have

$$P_{(i-N,j|i,j)} = \left(\sum_{k=0}^{i-N} d^P_{(i,j,N+k)} \cdot a^P_{(i,j,k,N+k)} \right) \cdot \left(\sum_{k=0}^{j} a^S_{(i,j,k,k)} \cdot d^S_{(i,j,k)} \right) . \tag{5.65}$$

5.C.3 Transition $S_{(i,j)} \rightarrow S_{(i+N,j)}$ with $-i < N \leq C - i$

In subsections 5.C.1 and 5.C.1 we have separately presented transitions involving an increase and decrease of (positive) PUs respectively. We may generalize the case of transition $S_{(i,j)} \rightarrow S_{(i+N,j)}$ when N can be either positive or negative for $-i \leq N \leq C - i$.

From (5.64), we have $P_{(i-N|i)}$, i.e. the transition probability from having i PUs to having $i - N$ PUs with $0 < N \leq i$. Note that this is equivalent to consider the probability $P_{(i+N|i)}$ with $-i < N \leq 0$, i.e. negative N. Then we can re-write (5.64) as:

$$P_{(i+N|i)} = \left(\sum_{k=0}^{i+N} d^S_{(i,j,k-N)} \cdot a^S_{(i,j,k,k-N)} \right) , \tag{5.66}$$

for $-i < N \leq 0$.

With the change of variable $t = k - N$ in (5.66), we have:

$$P_{(i+N|i)} = \left(\sum_{t=-N}^{i} d^P_{(i,j,t)} \cdot a^P_{(i,j,t+N,t)} \right) , \tag{5.67}$$

for $-i \leq N \leq 0$, which resembles transition probability $P_{(i+N|i)}$ for $0 < N \leq C - i$ given in (5.60), except for the summation index starting value. Nevertheless, we can state the general expression for the transition probability $S_{(i,j)} \rightarrow S_{(i+N,j)}$ for $-i \leq N \leq C - i$ as:

$$P_{(i+N,j|i,j)} = \left(\sum_{k=\max(-N,0)}^{i} a^{P}_{(i,j,N+k,k)} \cdot d^{P}_{(i,j,k)} \right) \cdot \left(\sum_{k=0}^{j} a^{S}_{(i,j,k,k)} \cdot d^{S}_{(i,j,k)} \right) . \quad (5.68)$$

For the case of transition probabilities due to the assignment or de-assignment of SUs, an analogous expression to (5.68) can be derived, yielding, for the transition $S_{(i,j)} \rightarrow S_{(i,j+M)}$ with $-j \leq M \leq C - j - i$:

$$P_{(i,j+M|i,j)} = \left(\sum_{k=\max(-M,0)}^{j} a^{S}_{(i,j,M+k,k)} \cdot d^{S}_{(i,j,k)} \right) \cdot \left(\sum_{k=0}^{i} a^{P}_{(i,j,k,k)} \cdot d^{P}_{(i,j,k)} \right) . \quad (5.69)$$

Finally, from (5.68) and (5.69), expression (5.23) is obtained.

Bibliography

[1] Qing Zhao and B. M. Sadler, "A survey of dynamic spectrum access," *Signal Processing Magazine, IEEE*, vol. 24, no. 3, pp. 79–89, 2007.

[2] M. Sherman, A. N. Mody, R. Martinez, C. Rodriguez, and R. Reddy, "IEEE standards supporting cognitive radio and networks, dynamic spectrum access, and coexistence," *Communications Magazine, IEEE*, vol. 46, no. 7, pp. 72–79, 2008.

[3] Federal Communications Commission (FCC), "Notice of Proposed Rule Making, ET Docket no. 04-113, May 25, 2004.," .

[4] C. Cordeiro, K. Challapali, D. Birru, and N. Sai Shankar, "IEEE 802.22: the first worldwide wireless standard based on cognitive radios," in *New Frontiers in Dynamic Spectrum Access Networks, 2005. DySPAN 2005. 2005 First IEEE International Symposium on*, December 2005, pp. 328–337.

[5] Standards Coordinating Committee (SCC) 41, "(http://www.scc41.org/).," .

[6] J. M. Peha, "Emerging Technology and Spectrum Policy Reform," in *International Telecommunications Union (ITU) Workshop on Market Mechanisms for Spectrum Management*, Jan. 2007.

[7] M. M. Buddhikot, P. Kolodzy, S. Miller, K. Ryan, and J. Evans, "DIMSUMnet: new directions in wireless networking using coordinated dynamic spectrum," in *World of Wireless Mobile and Multimedia Networks, 2005. WoWMoM 2005. Sixth IEEE International Symposium on a*, 2005, pp. 78–85.

[8] Jun Zhao, Haitao Zheng, and Guang-Hua Yang, "Distributed coordination in dynamic spectrum allocation networks," in *New Frontiers in Dynamic Spectrum Access Networks, 2005. DySPAN 2005. 2005 First IEEE International Symposium on*, 2005, pp. 259–268.

[9] D. Raychaudhuri and Xiangpeng Jing, "A spectrum etiquette protocol for efficient coordination of radio devices in unlicensed bands," in *Personal, Indoor and Mobile Radio Communications, 2003. PIMRC 2003. 14th IEEE Proceedings on*, February 2004, vol. 1, pp. 172–176 Vol.1.

[10] D. Bourse, P. Ballon, P. Cordier, S. Delaere, B. Deschamps, D. Grandblaise, K. Moessner, and O. Simon, "The E2R II Flexible Spectrum Management (FSM) - technical, business & regulatory perspectives," E2R II White Paper, July 2007.

[11] P. Cordier, P. Houze, S. B. Jemaa, O. Simon, D. Bourse, D. Grandblaise, K. Moessner, J. Luo, C. Kloeck, K. Tsagkaris, R. Agusti, N. Olaziregi, Z. Boufidis, E. Buracchini, P. Goria, and A. Trogolo, "E2R cognitive pilot channel concept," in *15th IST Mobile and Wireless Communications Summit*, June 2006.

[12] J. Perez-Romero, O. Sallent, R. Agusti, and L. Giupponi, "A novel on-demand cognitive pilot channel enabling dynamic spectrum allocation," in *New Frontiers in Dynamic Spectrum Access Networks, 2007. DySPAN 2007. 2nd IEEE International Symposium on*, June 2007, pp. 46–54.

[13] Xiaorong Zhu, Lianfeng Shen, and T. S. P. Yum, "Analysis of cognitive radio spectrum access with optimal channel reservation," *Communications Letters, IEEE*, vol. 11, no. 4, pp. 304–306, April 2007.

[14] M. Raspopovic and C. Thompson, "Finite population model for performance evaluation between narrowband and wideband users in the shared radio spectrum," in *New Frontiers in Dynamic Spectrum Access Networks, 2007. DySPAN 2007. 2nd IEEE International Symposium on*, June 2007, pp. 340–346.

[15] Shensheng Tang and B. L. Mark, "An analytical performance model of opportunistic spectrum access in a military environment," in *Wireless Communications and Networking Conference, 2008. WCNC 2008. IEEE*, April 2008, pp. 2681–2686.

[16] D. T. C. Wong, Anh T. Hoang, Ying-Chang Liang, and F. P. S. Chin, "Dynamic spectrum access with imperfect sensing in open spectrum wireless networks," in *Wireless Communications and Networking Conference, 2008. WCNC 2008. IEEE*, April 2008, pp. 2765–2770.

[17] Yiping Xing, R. Chandramouli, S. Mangold, and N, S.S., "Dynamic spectrum access in open spectrum wireless networks," *Selected Areas in Communications, IEEE Journal on*, vol. 24, no. 3, pp. 626–637, March 2006.

[18] Beibei Wang, Zhu Ji, and K. J. R. Liu, "Primary-prioritized markov approach for dynamic spectrum access," in *New Frontiers in Dynamic Spectrum Access Networks, 2007. DySPAN 2007. 2nd IEEE International Symposium on*, June 2007, pp. 507–515.

[19] W. Ahmed, J. Gao, H. A. Suraweera, and M. Faulkner, "Comments on "analysis of cognitive radio spectrum access with optimal channel reservation"," *Wireless Communications, IEEE Transactions on*, vol. 8, no. 9, pp. 4488–4491, October 2009.

[20] D. Cabric, A. Tkachenko, and R. W. Brodersen, "Spectrum sensing measurements of pilot, energy, and collaborative detection," in *Military Communications Conference, 2006. MILCOM 2006. IEEE*, February 2007, pp. 1–7.

[21] A. Ghasemi and E. S. Sousa, "Collaborative spectrum sensing for opportunistic access in fading environments," in *New Frontiers in Dynamic Spectrum Access Networks, 2005. DySPAN 2005. 2005 First IEEE International Symposium on*, December 2005, pp. 131–136.

[22] F. F. Digham, M. S. Alouini, and M. K. Simon, "On the energy detection of unknown signals over fading channels," in *Communications, 2003. ICC '03. IEEE International Conference on*, June 2003, vol. 5, pp. 3575–3579 vol.5.

[23] Yiyang Pei, Anh T. Hoang, and Ying-Chang Liang, "Sensing-throughput tradeoff in cognitive radio networks: How frequently should spectrum sensing be carried out?," in *Personal, Indoor and Mobile Radio Communications, 2007. PIMRC 2007. IEEE 18th International Symposium on*, December 2007, pp. 1–5.

[24] Dimitri Bertsekas and Robert Gallager, *Data Networks*, Prentice Hall, second edition, 1992.

[25] William J. Stewart, *Introduction to the numerical solution of Markov chains*, Princeton University Press, 1994.

Chapter 6

Flexible Spectrum Access for Opportunistic Secondary Operation in Cognitive Radio Networks

In this chapter we adopt a Discrete-Time Markov Chain (DTMC) framework presented in Chapter 5 in order to capture the effect of flexible spectrum channelization for the opportunistic access of secondary users (SUs) in a primary-secondary shared spectrum scenario. In this sense, two implementation alternatives are proposed: a fixed channelization scheme (FCS), which partitions the whole spectrum bandwidth in a fixed number of channels, and an adaptive channelization scheme (ACS), which dynamically adjusts the spectrum channelization in order to maximize the use of spectrum without interfering with primary spectrum use. Additionally, the service-type characterization of SUs is also explored by defining two different types of secondary traffic behaviors, time vs. volume based, which are evaluated in the presented framework for the FCS and the ACS. Results indicate the suitability of the ACS over the FCS exhibiting a higher flexibility and efficiency in allocating spectrum resources.

6.1 Motivation and Problem Statement

A common abstraction of spectrum resources in a shared spectrum environment is that of a given frequency band (Hz) partitioned into a number of channels, see e.g. [1–3]. In this case, it is necessary to ensure that a particular channel is not accessed by both a PU and a SU at the same time thus causing mutual interference.

According to the spectrum partition sizes (i.e. channel sizes) devoted to PUs and SUs in a spectrum sharing system, users may be categorized in Wide Band (WB) access users as opposed to Narrow Band (NB) access users. For example, in [1] it is assumed that PUs have WB access as opposed to SUs which have NB access. Oppositely, [2] considers PUs have NB access while SUs have WB access. As for [3], both PUs and SUs access the same amount of spectrum. It is worthwhile noting that the spectrum access models in [1–3] solely apply for the specific partition cases explained above, thus exhibiting limited applicability. To this respect, this chapter explores several alternatives for the partition, or channelization, of the available spectrum in order to provide an efficient access and spectrum utilization for both PUs and SUs. Specifically, a Fixed Channelization Scheme (FCS) and an Adaptive Channelization Scheme (ACS) are proposed as two possible spectrum access mechanisms for SUs. As may be intuited, the FCS partitions the whole spectrum bandwidth in a number of fixed channels for both PUs and SUs. As opposed to [1–3], where specific models accounted for particular WB-NB relations between PUs and SUs, the FCS presented in this chapter allows to define the entire set of WB vs. NB cases between PUs and SUs, thus exhibiting an improved applicability and scope. To the best of the author's knowledge, only [4] describes a similar approach where the channelization values for PUs and SUs can be adjusted. Moreover, the ACS proposed in this chapter is able to adapt the channelization value of SUs according to current traffic demands of both PUs and SUs, thus exhibiting a higher flexibility as compared to both the FCS and the model in [4].

In addition, it is also considered the characterization of demanding SU services by defining two types of service requests. Firstly, Time-Based Services (TBSs) aim for the use of a particular amount of bandwidth for a given time. Secondly, Volume-Based Services (VBSs) aim at transmitting a given amount of data so that the service duration depends on the achievable bit-rate, i.e. on the amount of spectrum bandwidth assigned to each user. The aforementioned service characterization for SUs also constitutes a major contribution with respect to existing works which usually assume the TBS case (see, e.g., [1–4]). In this chapter, we explore the use of the aforementioned channelization schemes when different SU service characterizations apply. To accomplish this task, we rely on a Markov model previously introduced in Chapter 5 and adapt it to include the aforementioned channelization and service characterizations.

In the abovementioned context, the use of Markov models becomes an important aid in modeling problems dealing with the dynamic access to shared spectrum resources. In this sense, a significant number of papers in the literature have been devoted to the characterization of such scenarios using Markov models as, e.g., in [1–4]. In [1] (along with amendments in [5]), a Continuous Time Markov Chain

(CTMC) model is presented to model spectrum access of primary (wideband) and secondary (narrowband) users over a partitioned spectrum bandwidth. In [2, 3] a CTMC model is also provided for the opportunistic access of wideband SUs and narrowband PUs in [2], and equal-band PUs and SUs in [3]. However, as a difference from [1], a finite population traffic model is used for the characterization of secondary users. It is worthwhile noting that work in [1–3] disregards the effect of erroneous sensing on the secondary network side, i.e. a perfect knowledge on the activity of primary users is assumed. An attempt to introduce the impact of sensing errors is provided in [4], where a CTMC model is also considered and sensing information is available upon secondary user arrival. Despite the fact that some considerations about sensing errors are introduced in [4], these are not related to any particular spectrum sensing mechanism (i.e., energy detection, pilot detection, etc. [6]). Conversely, in this chapter, missed-detection and false-alarm values are obtained according to the well-known expressions regarding the energy detection of signals in Rayleigh fading as in [7, 8], hence achieving higher modeling accuracy.

The remainder of the chapter is organized as follows. In Section 6.2 we describe the spectrum channelization schemes FCS and ACS along with service types TBS and VBS. Section 6.3 presents the DTMC model where the channelization and service type characterization are included thus extending the work in Chapter 5. Performance metrics extracted from the model are defined in Section 6.4. Numerical results are provided in Section 6.5 and conclusions are given in Section 6.6.

6.2 Spectrum Access Model

As in Chapter 5, the considered scenario is that of an infrastructure-based primary network (PN), where the primary Base Station (BS) provides connectivity to PUs. The PN has a prioritized use of the available spectrum band W_T where individual PUs are assigned a specific channel for use. Alongside, a secondary infrastructure-based network provides connectivity to SUs through the opportunistic use of unoccupied spectrum. We are concerned with the case where PUs and SUs coexist in an area with coverage of both the PN and the SN.

The following sections address the considered alternatives for the PN and the SN to partition the whole spectrum band for PU and SU access along with the characterization of demanding services.

6.2.1 Spectrum Channelization

Assume that the PN partitions the whole spectrum into M_p channels (or frequency bands). On the other hand, the SN may consider a spectrum channelization of M_s. The SN could decide the most appropriate value of M_s according to the network status or SU service characteristics at a given time. Fig. 6.1 illustrates the spectrum channelization concept through parameters M_p and M_s. It can be observed that, with one SU accessing the shared spectrum, by choosing $M_s = \{16, 12, 8\}$ [see Fig. 6.1(a), Fig. 6.1(b) and Fig. 6.1(c)] we do not maximize spectrum usage, whereas if $M_s = 4$ [Fig. 6.1(d)] this is accomplished. Furthermore, in the same example with one SU and three PUs, if $M_s = 2$ or $M_s = 1$ [Fig. 6.1(e) and Fig. 6.1(f)] note that non-harmful secondary access is not possible. For the sake of algebra tractability, the following assumption is considered:

(A1) $M_s/M_p \in \mathbb{Z}$ provided $M_s \geq M_p$ and $M_p/M_s \in \mathbb{Z}$ for the case of $M_p \geq M_s$,

meaning that subchannels are always an integer fraction of a channel. Furthermore, and without loss of generality, it is assumed that

(A2) primary channelization is given by $M_p = 2^n$, for $n > 0$;

(A3) secondary channelization is given by $M_s = 2^m$ for $M_p \geq M_s$ $(m = 0, 1, 2...)$, and $M_s = m \cdot M_p$ for $M_s \geq M_p$ $(m = 1, 2, ...)$ with $M_s \leq M_{s,max}$ accounting for minimum spectrum requirements.

Note that both (A2) and (A3) fulfill assumption (A1).

For convenience, we define the set of secondary channelization values, M_s, as $\mathcal{M}_s$. For example, considering the layout in Fig. 6.1, we would have that $\mathcal{M}_s = \{1, 2, 4, 8, 12, 16\}$ where it has been assumed that $M_{s,max} = 16$.

The secondary BS (SBS) is responsible to inform SUs about the considered value for M_s through appropriate signaling channels. Since this operation involves time and resource consumption, two different schemes are proposed according to the interaction cadence between the SBS and the SUs. First, a *fixed channelization scheme* (FCS) will be adopted. In this case, the value of M_s is updated and signaled to the SUs once or, alternatively, at very large time-scales. Secondly, we assume an *adaptive channelization scheme* (ACS) where the value of M_s is

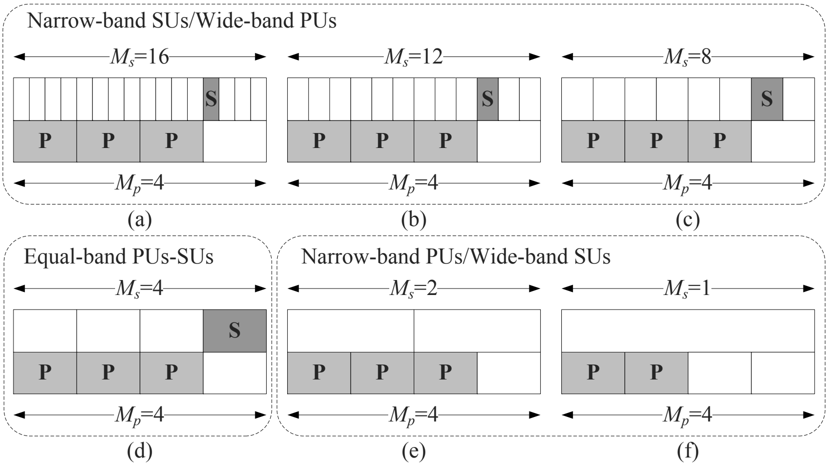


Figure 6.1: Spectrum channelization model given by M_p and M_s.

constantly adjusted so as to maximize spectrum utilization, as defined later on, among SUs.

Since secondary spectrum resources are varied according to the value of M_s, the SU throughput R_s (in bits-per-second, bps) will also vary accordingly. Indeed, the Shannon capacity expression which relates the achievable throughput per channel to the available bandwidth can be expressed as:

$$R_s = \frac{W_T}{M_s} \log_2 \left(1 + \gamma\right), \tag{6.1}$$

where γ is the Signal-to-Noise Ratio (SNR) affecting the SU. According to (6.1), a high spectrum partition (i.e. high M_s values) will allow a higher number of SUs in the system at the cost of reduced throughput performance. On the other hand, low M_s values imply increased throughput at the expense of low number of SUs accessing spectrum resources.

Finally, it is assumed that the SN knows the value of the primary channelization, M_p, which, moreover, will be assumed to have a fixed value.

6.2.2 Service Type Characterization

As in [9, 10], we consider two different service types for SUs according to the amount of time these users access the shared spectrum.

First, a *time-based service*[1] (TBS) is considered, where a specific user demands spectrum access during some given time regardless of the achieved throughput R_s. This could be the case of a service demanding a Constant Bit Rate (CBR) during its session length, or a Variable Bit Rate (VBR) service, such as e.g. a video streaming service, where an increased throughput would mean an improved perceived QoS without affecting the duration of the service.

Secondly, we consider a *volume-based service*[2] (VBS) where a specific SU intends to transmit an amount of data bits. Accordingly, the access time spent by this user will depend on the achievable throughput.

As for PUs, it is assumed that they solely demand TBS.

6.3 Discrete-Time Markov Model

The following model departs from the work in Chapter 5 which addressed the particular case where $M_p = M_s = C$. The interested reader is referred to the previous chapter for a detailed description of the expressions that, in this chapter, we extend to capture the FCS and ACS for TBS and VBS.

6.3.1 State Space Definition

In a DTMC we observe the system state at discrete time instants $\{t_0, t_1, t_2, ..., t_n, ...\}$, with $t_n = t_0 + n \cdot \Delta T$ and periodicity ΔT, which is, moreover, assumed to specify the sensing periodicity. In addition, let $I_n = (t_n, t_{n+1}]$ define the *n-th* time interval between two successive observation times.

If $N_p(t_n)$ and $N_s(t_n)$ are stochastic processes indicative of the number of PUs and SUs in the system at time t_n, then, let $\mathbf{X}_n = S_{(i,j)} = \{N_p(t_n) = i, N_s(t_n) = j\}$ represent a state of the DTMC at time t_n.

The *state space* will define what states in our DTMC model are feasible or not and will depend on the adopted channelization scheme.

[1] Sometimes named as inelastic or stream service in the literature [10].

[2] Sometimes named as elastic service in the literature [10].

6.3.1.1 Fixed Channelization Scheme (FCS)

In this case, the state space $\mathcal{S}$ is the set of states such that

$$\mathcal{S} = \{S_{(i,j)} : 0 \leq i \leq M_p, 0 \leq j \leq M_s\}, \tag{6.2}$$

which implies that at most M_p PUs and M_s SU are allowed in the system. However, whenever a PU and a SU occupy the same channel we assume that interference occurs such that the performance of both users is severely degraded. Then, for convenience, we define the following three subsets of $\mathcal{S}$ accounting for those states that necessarily imply spectrum collision, i.e.

$$\mathcal{S}_c = \{S_{(i,j)} : i/M_p + j/M_s > 1\} \subset \mathcal{S}, \tag{6.3}$$

those states which possibly imply a spectrum collision, i.e.

$$\mathcal{S}_{pc} = \{S_{(i,j)} : i/M_p + j/M_s \leq 1, j > 0, i > 0\} \subset \mathcal{S}, \tag{6.4}$$

and those states that are collision-free, i.e.

$$\mathcal{S}_{nc} = \big\{\big(S_{(i,j)} : i = 0\big) \cup \big(S_{(i,j)} : j = 0\big)\big\} \subset \mathcal{S}. \tag{6.5}$$

6.3.1.2 Adaptive Channelization Scheme (ACS)

For the ACS, the value of M_s will vary according to current spectrum occupation. It is further assumed that its value is bounded so that $0 \leq M_s \leq M_{s,max}$, where $M_{s,max}$ is a design parameter. Then, in this case, the state space $\mathcal{S}$ is the set of states such that

$$\mathcal{S} = \{S_{(i,j)} : 0 \leq i \leq M_p, 0 \leq j \leq M_{s,max}\}, \tag{6.6}$$

where, analogously to the FCS case in Section 6.3.1.1, we can define the set of states implying collision as

$$\mathcal{S}_c = \{S_{(i,j)} : i/M_p + j/M_{s,max} > 1\} \subset \mathcal{S}, \tag{6.7}$$

along with those states possibly implying a spectrum collision, i.e.

$$\mathcal{S}_{pc} = \{S_{(i,j)} : i/M_p + j/M_{s,max} \leq 1, j > 0, i > 0\} \subset \mathcal{S}. \tag{6.8}$$

As for the non-collision state space $\mathcal{S}_{nc}$, it is given by (6.5).

6.3.2 Spectrum Awareness Model

Due to spectrum sensing errors, the observed state at time t_n may be $\mathbf{Y}_n = S_{(k,j)} \in \mathcal{S}$, such that $\mathbf{Y}_n \neq \mathbf{X}_n$, with k denoting the number of sensed PUs. Consequently, it can be shown (refer to Appendix 5.A in Chapter 5), that the conditioned probability of sensing k PUs when there are actually i PUs at time t_n is:

$$b_{(k,i)} = \sum_{m=\max(0,i-k)}^{\min(i,M_p-k)} \binom{M_p - i}{m+k-i} \cdot \varepsilon^{m+k-i} \cdot \bar{\varepsilon}^{M_p-m-k} \cdot \binom{i}{m} \cdot \delta^m \cdot \bar{\delta}^{i-m}, \qquad (6.9)$$

with $\bar{\varepsilon} = 1 - \varepsilon$ and $\bar{\delta} = 1 - \delta$.

According to (6.9), false-alarm and misdetection, ε and δ respectively, will affect the sensed number of PUs by the SN, thus potentially causing erroneous decisions due to the inaccuracy of spectrum occupation information. In this sense, the DTMC model will capture such effects which results in a more realistic scenario characterization compared to known approaches so far.

6.3.3 Channelization Scheme

According to the chosen channelization scheme, i.e. FCS or ACS, the value of M_s will be different.

For the ACS the value of M_s will vary depending on the state $S_{(i,j)}$. We are then interested in finding the minimum value of M_s (which implies higher spectrum occupation and, thus higher throughput) such that the spectrum utilization is maximized. In particular, we focus on those values of M_s belonging to the set $\mathcal{M}_s$:

$$M_s(i,j) = \arg\max_{M_s \in \mathcal{M}_s} [U(i,j)], \qquad (6.10)$$

where

$$U(i,j) = i/M_p + j/M_s \qquad (6.11)$$

is the spectrum utilization.

Expression (6.10) guarantees that the utmost spectrum utilization is achieved fairly among all SUs. However, since the SN will select the channelization value and it is not aware of the ”true” number of PUs (i) but rather on the number of sensed PUs (m), it will be able to compute $M_s(m,j)$ instead of $M_s(i,j)$. Therefore, sensing errors will affect the channelization adjustment process.

For the FCS, SU channelization M_s is simply a constant value considered as an input parameter, thus $M_s(i,j) = M_s$.

6.3.4 Secondary Service Characterization

For TBS, the service-time distribution is given by the exponentially distributed service rate with average duration $1/\mu_s$. Consequently, the service departure rate in state $S_{(i,j)}$ is given by $\mu_s(i,j) = j \cdot \mu_s$ [11].

For the VBS, the average service-time of a SU will depend on the data volume to be transmitted (L), which is assumed to be exponentially distributed, as usually considered in different works, such as e.g. in [9, 10], and on the achieved data-rate (R_s) as:

$$1/\mu_s = \frac{E[L]}{R_s \cdot \eta_{sens}} = \frac{E[L]}{\eta_{sens} \cdot [W_T/M_s(i,j)] \cdot \log_2(1+\gamma)} \triangleq M_s(i,j)/\mu_T \tag{6.12}$$

where (6.1) has been used and $1/\mu_T$ is defined as the average service-time when a single SU is accessing the full spectrum W_T. In addition, the secondary bit-rate is affected by the sensing efficiency $\eta_{sens} \in [0,1]$, which will be defined in section 6.4. Consequently, the service departure rate in state $S_{(i,j)}$ is given by $\mu_s(i,j) = \mu_T/M_s(i,j)$, indicating that the service rate is diminished when increasing channelization M_s. Note that the value of M_s will depend on the adopted channelization scheme, thus yielding $M_s(i,j) = M_s$ for the FCS, and $M_s(i,j)$ defined in (6.10) for the ACS.

6.3.5 Arrival and Departure Processes

Assuming PUs and SUs arrive to the system according to a Poisson distribution with rates λ_p and λ_s respectively, the probability that k arrivals occur in I_n, P_k^A, is given by [see (5.12) in Section 5.3.3]:

$$P_k^A = [(\lambda \Delta T)^k / k!] \exp[-\lambda \Delta T], \tag{6.13}$$

where for $\lambda \in \{\lambda_p, \lambda_s\}$ then we will refer to $P_k^A \in \{P_k^{PA}, P_k^{SA}\}$.

If the session duration is exponentially distributed with rate μ, the probability of having k-out-of-m departures in I_n, P_k^D, is given by [see (5.14) in Section 5.3.3]

$$P_k^D = \binom{m}{k} (1 - \exp[-\mu \Delta T])^k (\exp[-\mu \Delta T])^{m-k}, \tag{6.14}$$

where for $\mu \in \{\mu_p, \mu_s\}$ then we will refer to $P_k^D \in \{P_k^{PD}, P_k^{SD}\}$. Specifically for the secondary departure rates, and according to the defined services TBS and VBS, the value of μ_s will be different in each case. For the TBS, μ_s is a constant value regarded as an input parameter. On the other hand, for the VBS, $\mu_s = \mu_s(i, j) = \mu_T / M_s(i, j)$ where if the FCS is applied then $M_s(i, j) = M_s$, and if the ACS is adopted then $M_s(i, j)$ is given in (6.10).

For the sake of algebra tractability, it will be assumed in the remainder of the chapter, as also considered in Chapter 5 (see hypothesis 1 and 2 in Section 5.3.4), that

(A4) a session arriving in I_n will not depart in the same I_n.

Note that this implies that both ΔT and the duration of a session must be carefully chosen such that $\Delta T << 1/\mu$ with $\mu \in \{\mu_p, \mu_s\}$ and where $1/\mu$ is the average session duration.

In addition,

(A5) we disregard the order in which session arrivals and departures occur in a given I_n by considering the resulting net number of users, i.e. those obtained after subtracting the departures and adding the new arrivals.

Note that enabling multiple arrivals in one ΔT will affect the decision process on whether a SU can be assigned or not given that detection information is retrieved only at times t_n.

The applicability range of both (A4) and (A5) was assessed and validated against a system-level simulator in Chapter 5, to which in the interested reader is referred.

6.3.6 State Transition Probabilities

In this section we provide the expressions for the transition probabilities in the presented DTMC model. Given that this model was detailed in Chapter 5 for the simple case of $M_p = M_s \triangleq C$, we just outline the new probability expressions considering the channelization model and service type differentiation. For a deeper understanding of such expressions the interested reader is particularly referred to the aforementioned Chapter 5.

According to Theorem 2 in Chapter 5 expressed by (5.23), the general transition probability from $S_{(k,l)} \rightarrow S_{(i+N,j+M)}$ is given by

$$\begin{aligned} P_{(i+N,j+M|i,j)} &= \left(\sum_{k=\max(-N,0)}^{i} a^{P}_{(i,j,N+k,k)} \cdot d^{P}_{(i,j,k)} \right) \\ &\quad \times \left(\sum_{k=\max(-M,0)}^{j} a^{S}_{(i,j,M+k,k)} \cdot d^{S}_{(i,j,k)} \right), \end{aligned} \tag{6.15}$$

where $-i \leq N \leq M_p - i$ along with $-j \leq M \leq M_s - j$ for the FCS and $-j \leq M \leq M_{s,max} - j$ for the ACS. Parameters, $a^{P}_{(i,j,k,l)}$, $d^{P}_{(i,j,k,l)}$, $a^{S}_{(i,j,k,l)}$ and $d^{S}_{(i,j,k,l)}$ in (6.15) account for assignment and de-assignment probabilities of primary and secondary users respectively (refer to Chapter 5 for a detailed explanation of these probabilities in the case of $M_p = M_s \triangleq C$).

Remark In the forthcoming expressions, secondary channelization is simply referred to as $M_s(i,j)$, where this value should be conveniently replaced according to the channelization schemes FCS and ACS. Specifically, when dealing with the assignment process of a SU for the ACS, we have $M_s(i,j) = M_{s,max}$ indicating that the assignment of a SU in this case depends on the maximum channelization value. Similarly, secondary service rate will be referred to as $\mu_s(i,j)$, where it should be also particularized for expressions of each adopted service type, i.e. TBS and VBS.

In state $S_{(i,j)}$, the probability of assigning k PUs when also l PU de-assignments occur in I_n, $a^{P}_{(i,j,k,l)}$, is given by

$$a^{P}_{(i,j,k,l)} = \begin{cases} P^{PA}_{k} & \text{if } i - l + k < M_p \qquad (6.16a) \\ 1 - \displaystyle\sum_{m=0}^{k-1} P^{PA}_{m} & \text{if } i - l + k = M_p \qquad (6.16b) \end{cases}$$

with P^{PA}_{k} given in (6.13) for $\lambda = \lambda_p$. Expression (6.16), analogue to (5.16), accounts for the arrival of exactly k PUs if more than k channels are available [subcase (6.16a)], and the arrival of at least k PUs if only k channels are disposable [subcase (6.16b)].

In state $S_{(i,j)}$, the probability of de-assigning k PUs in I_n, $d^{P}_{(i,j,k)}$, is, similar to (5.17),

$$d^{P}_{(i,j,k)} = P^{PD}_{k}, \tag{6.17}$$

where P^{PD}_{k} is given by (6.14) with $\mu = \mu_p$.

In state $S_{(i,j)}$, the probability of assigning $k > 0$ (the case $k = 0$ is treated separately) SUs when also l SU de-assignments occur in I_n, $a^S_{(i,j,k,l)}$, is given by

$$a^S_{(i,j,k,l)} = \sum_{m=0}^{\lfloor \psi_{(i,j,k,l)} \rfloor} \bar{a}^S_{(m,j,k,l)} \cdot b_{(m,i)}, \tag{6.18}$$

where, $\bar{a}^S_{(m,j,k,l)}$ is the SU assignment probability conditioned to the sensing of m PUs, thus (6.18) expresses the de-conditioning by considering probability $b_{(m,i)}$ given in (6.9). The upper summation index in (6.18), $\psi_{(i,j,k,l)}$, gives the maximum number of sensed PUs without incurring in spectrum collision, as defined in Section 6.3.1, given by

$$\psi_{(i,j,k,l)} = [M_s(i,j) - k - j + l] \cdot [M_p/M_s(i,j)]. \tag{6.19}$$

This value, given it depends on the ratio $M_p/M_s(i,j)$, may yield fractional values, thus the floor function $\lfloor x \rfloor$ (denoting the highest integer less than or equal to x) should be applied as reflected in (6.18). Note in (6.19) that for the case of FCS we will consider $M_s(i,j) = M_s$, whereas for ACS $M_s(i,j) = M_{s,max}$. Function $\bar{a}^S_{(m,j,k,l)}$ in (6.18) is given by

$$\bar{a}^S_{(m,j,k,l)} = \begin{cases} P_k^{SA} & \text{if } m < \psi_{(i,j,k,l)} & (6.20a) \\ 1 - \sum_{r=0}^{k-1} P_r^{SA} & \text{if } m = \psi_{(i,j,k,l)} & (6.20b) \end{cases}$$

where cases (6.20a) and (6.20b) are analogous to the PU assignment cases in (6.16a) and (6.16b) respectively.

For the specific case of no SU assignments (i.e. $k = 0$) in $S_{(i,j,k,l)}$ during I_n, the probability of assigning $k = 0$ SUs when also l SU de-assignments occur, $a^S_{(i,j,0,l)}$, is the probability of assigning $k = 0$ SUs when there is at least one free detected channel [first bracketed summand in (6.21)] or the probability that there are no detected free channels [second bracketed summand in (6.21)]:

$$a^S_{(i,j,0,l)} = \left(\sum_{m=0}^{\xi_{(i,j,l)}} \bar{a}^S_{(m,j,0,l)} \cdot b_{(m,i)} \right) + \left(\sum_{m=\xi_{(i,j,l)}+1}^{M_p} b_{(m,i)} \right), \tag{6.21}$$

where $\xi_{(i,j,l)} = \lfloor [M_s(i,j) - 1 - j + l] \cdot [M_p/M_s(i,j)] \rfloor$ gives the maximum number of sensed PUs yielding one single free detected channel.

Finally, the probability of de-assigning k SUs in state $S_{(i,j)}$ during I_n, $d^S_{(i,j,k)}$, is given by

$$d^S_{(i,j,k)} = \sum_{r=0}^{k} d^{S,S}_{(i,j,k-r,r)} \cdot d^{S,SC}_{(i,j,r)}, \tag{6.22}$$

where, similar to (5.20), the SU de-assignment process may be caused by two different events: 1) spectrum sensing determines that a number of SUs are sharing the same channels with PUs, and hence need to be interrupted, which is captured by probability $d^{S,S}_{(i,j,k,l)}$; and 2), SU service completion causes the de-assignment of SUs that have ended their transmission, which is captured by probability $d^{S,SC}_{(i,j,k)}$.

Accordingly, in state $S_{(i,j)}$, the probability of de-assigning k SUs due to sensing when l SUs are de-assigned due to service completion, $d^{S,S}_{(i,j,k,l)}$, is given, for $0 < k \leq j-l$, by

$$d^{S,S}_{(i,j,k,l)} = \begin{cases} b_{(\chi_{(i,j,k,l)},i)} & \text{if } \chi_{(i,j,k,l)} \in \mathbb{Z} \\ 0 & \text{if } \chi_{(i,j,k,l)} \in \mathbb{R} \end{cases}, \tag{6.23}$$

with $\chi_{(i,j,k,l)} = [M_s(i,j)+k-j+l] \cdot [M_p/M_s(i,j)]$ the number of sensed PUs in $S_{i,j}$ leading to the de-assignment of k PUs when l SUs are de-assigned due to service completion. According to (6.23), it can be shown that rational number of sensed PUs values, i.e. $\chi_{(i,j,k,l)} \in \mathbb{R}$, lead to infeasible de-assignments which are therefore disregarded in the model. For $k = 0$ we have

$$d^{S,S}_{(i,j,0,l)} = 1 - \sum_{r=1}^{j-l} d^{S,S}_{(i,j,r,l)}. \tag{6.24}$$

The de-assignment of k SUs due to service completion is, similar to (5.22),

$$d^{S,SC}_{(i,j,k)} = P^{SD}_{k}, \tag{6.25}$$

with P^{SD}_k given by (6.14) for $\mu = \mu_s$.

6.3.7 Steady State Distribution

From the resulting transition probability matrix $\mathbf{P}$, with elements given by (6.15), we obtain the true steady state probabilities, $P_{(i,j)} = \lim_{n\to\infty} \Pr\left[\mathbf{X}_n = S_{(i,j)}\right]$, for each true state $S_{(i,j)}$ in the state space $\mathcal{S}$. The knowledge of such statistical distribution enables the definition of several performance metrics which are addressed in the following.

On the other hand, it is also relevant to determine the steady state probabilities of the detected states (i.e. including possible sensing errors):

$$P'_{(i,j)} = \lim_{n\to\infty} \Pr\left[\mathbf{Y}_n = S_{(i,j)}\right],$$

which are computed as:

$$P'_{(i,j)} = \sum\nolimits_{n=0}^{M_p} b_{(i,n)} \cdot P_{(n,j)}. \tag{6.26}$$

In this case, $P'_{(i,j)}$ is the steady state probability observed by the SN, i.e. without true knowledge of PU activity and, therefore, sensible to sensing errors. Then, by considering $P'_{(i,j)}$ instead of $P_{(i,j)}$ metrics computed from the SN side, which account for possible sensing errors, can be computed.

So as to obtain probabilities $P_{(i,j)}$, numerical methods are used [12] to solve the matrix equation given by $\nu = \nu\mathbf{P}$, where

$$\nu = \left[P_{(0,0)}, P_{(0,1)}, \ldots, P_{(0,M_s)}, P_{(1,0)}, \ldots, P_{(1,M_s)}, \ldots, P_{(M_p,0)}, \ldots, P_{(M_p,M_s)}\right]$$

is the steady-state probability vector.

6.4 Performance Metrics

The definition of some relevant performance metrics is described in the following subsections.

6.4.1 Offered Traffic Load and Expected Service Times

Whereas the TBS enables to define the offered traffic load as an input parameter, given by the ratio $T_p = \lambda_p/\mu_p$ and $T_s = \lambda_s/\mu_s$ for PUs and SUs respectively, the offered traffic load for the VBS case of SUs is affected by the state-dependent service rate, i.e. $\mu_s = \mu_s(i,j)$. Hence, for the VBS case the offered secondary traffic has to be computed a posteriori. The average residence time (or average transfer delay) for a SU $E[t_s]$ can be computed as

$$E[t_s] = \left(\sum_{S_{(i,j)} \in \mathcal{S}, j>0} \frac{1}{\mu_s(i,j)} \cdot P_{(i,j)}\right) \Bigg/ \left(\sum_{S_{(i,j)} \in \mathcal{S}, j>0} P_{(i,j)}\right), \tag{6.27}$$

where the denominator accounts for the conditioning over $j > 0$.

Consequently, using Little's law, the offered secondary load for the VBS case can be computed as $T_s = \lambda_s \cdot E[t_s]$. Note that for the case of TBS we would have $E[t_s] = 1/\mu_s$.

6.4.2 Blocking Probability

Blocking of PUs happens whenever the admission of a new PU forces a transition to a non-feasible state. Thus, it is the probability of having $i{=}M_p$ PUs regardless of the number of SUs in the system, i.e.

$$P_B^P = \sum\nolimits_{j=0}^{M_s} P_{(M_p,j)}. \tag{6.28}$$

The blocking probability for SUs is defined as

$$P_B^S = \sum\nolimits_{S_{(i,j)} \in \mathcal{S}_b} P'_{(i,j)}, \tag{6.29}$$

where $\mathcal{S}_b$ is the set of blocking states defined as

$$\mathcal{S}_b = \left\{ S_{(i,j)} \; : \; S_{(i,j)} \notin \mathcal{S}_c, S_{(i,j+1)} \in \mathcal{S}_c \right\}, \tag{6.30}$$

which indicates that a blocking state $S_{(i,j)}$ is a non-collision state $[S_{(i,j)} \notin \mathcal{S}_c]$ in which the addition of a single SU forces the state to become a collision state, i.e. $S_{(i,j+1)} \in \mathcal{S}_c$. For FCS and ACS implementations, definitions of the collision set, $\mathcal{S}_c$, should be those provided in section 6.3.1.1 and 6.3.1.2 for FCS and ACS respectively.

6.4.3 Spectrum Collisions: An Upper-Bound

At a given sensing instant in state $S_{(i,j)} \in \mathcal{S}$, the probability of miss-detecting n-out-of-i PUs in the system can be expressed as the binomial distribution with δ the probability of miss-detection [refer to (6.9)]. Consequently, the average number of missed detections in state $S_{(i,j)} \in \mathcal{S}$ is given by

$$N_{MD}(i,j) = i \cdot \delta. \tag{6.31}$$

The average number of collisions in state $S_{(i,j)} \in \mathcal{S}_{pc}$ [recall from section 6.3.1] may be initially upper-bounded by

$$N_c(i,j) \leq N_{MD}(i,j), \tag{6.32}$$

indicating that at most (i.e. in the worst case) we will have a collision for each missed detected channel. In addition, this upper-bound can be tightened by considering that the average number of collisions will be also less than the average number of SUs in state $S_{(i,j)}$, i.e. $N_c(i,j) \leq j$. Hence, we have

$$N_c(i,j) \leq \min\left[i \cdot \delta, j\right] \triangleq N_c^*(i,j). \tag{6.33}$$

The number of collisions in state $S_{(i,j)} \in \mathcal{S}_c$ [see section 6.3.1] can be initially lower-bounded as

$$N_c(i,j) \geq \left\lceil \frac{M_s}{M_p} \cdot i \right\rceil + j - M_s \triangleq \kappa_c, \tag{6.34}$$

where we know that at least κ_c channels are being simultaneously shared by both a PU and a SU.

In addition, remaining $(i - \kappa_c)$ PUs and $(j - \kappa_c)$ SUs may be also in a collision situation, hence the number of collisions can be upper-bounded, similar to (6.33), as:

$$N_c(i,j) \leq \kappa_c + \min\left[(i - \kappa_c) \cdot \delta, (j - \kappa_c)\right] \triangleq N_c^*(i,j). \tag{6.35}$$

As for the case where $S_{(i,j)} \in \mathcal{S}_{nc}$, the number of collisions is zero, i.e. $N_c(i,j) = 0$.

6.4.4 Average Throughput

It is considered that a channel being shared by both a PU and a SU does not contribute to throughput; consequently we can define the throughput of PUs in state $S_{(i,j)}$ as

$$\Gamma^p_{(i,j)} = \left[i - N_c^*(i,j)\right] \cdot R_p, \tag{6.36}$$

measured in bits-per-second (bps) and where R_p is the granted bit-rate per channel for a PU and $N_c^*(i,j)$ is given by (6.33), (6.35) or $N_c^*(i,j) = 0$ for $S_{(i,j)} \in \mathcal{S}_{pc}$, $S_{(i,j)} \in \mathcal{S}_c$ and $S_{(i,j)} \in \mathcal{S}_{nc}$ respectively. In analogy to (6.1), we define $R_p = (W_T/M_p) \cdot \log_2(1+\gamma)$.

As for the throughput of SUs in state $S_{(i,j)}$ we have

$$\Gamma^s_{(i,j)} = \left[j - N_c^*(i,j)\right] \cdot \eta_{sens} \cdot R_s, \tag{6.37}$$

measured in bits per second (bps), where R_s is the granted bit rate per channel for SUs given by (6.1) and $\eta_{sens} \in [0,1]$ is the sensing efficiency which reflects the fraction of time devoted to data transmission (thus disregarding the time devoted to sensing purposes). The sensing efficiency is defined as

$$\eta_{sens} = 1 - T_{sens}/\Delta T, \tag{6.38}$$

where T_{sens} is the time devoted to sense all M_p channels, i.e. $T_{sens} = T \cdot M_p$ with T the time devoted to sense a single channel.

In addition, note that for ACS the secondary channelization depends on each state, i.e. $M_s = M_s(i,j)$; therefore, the granted bit rate R_s in (6.37) will be in this case $R_s = R_s(i,j)$.

Finally, based on expressions (6.36) and (6.37), the average throughput is given by

$$\Gamma^k = \sum_{S_{(i,j)} \in \mathcal{S}} \Gamma^k_{(i,j)} \cdot P_{(i,j)}, \tag{6.39}$$

with $k \in \{p, s\}$ accounting for PUs and SUs respectively. In addition, the average aggregate throughput is simply $\Gamma = \Gamma^p + \Gamma^s$.

6.4.5 Throughput-per-Secondary User

The throughput-per-SU in state $S_{(i,j)}$ is given by

$$\Gamma^s_u(i,j) = \Gamma^s_{(i,j)}/j, \tag{6.40}$$

for $j > 0$ and $\Gamma^s_{(i,j)}$ defined in (6.37). Then, the average throughput-per-SU can be obtained as

$$\Gamma^s_u = \left(\sum_{S_{(i,j)} \in \mathcal{S}, j>0} \Gamma^s_{(i,j)} \cdot P_{(i,j)} \right) \Bigg/ \left(\sum_{S_{(i,j)} \in \mathcal{S}, j>0} P_{(i,j)} \right). \tag{6.41}$$

6.4.6 Average SU Channelization

The average SU channelization value when employing the ACS is given by

$$\bar{M}_s = \left(\sum_{S_{(i,j)} \in \mathcal{S}, j>0} M_s(i,j) \cdot P_{(i,j)} \right) \Bigg/ \left(\sum_{S_{(i,j)} \in \mathcal{S}, j>0} P_{(i,j)} \right). \tag{6.42}$$

6.4.7 Arrival Rate Capacity

The arrival rate capacity limit is defined as the boundary, in offered arrival rates, of the feasible arrival rate capacity set given by

$$\alpha = \{ \boldsymbol{\lambda} = (\lambda_p, \lambda_s) \mid P^p_B(\boldsymbol{\lambda}) \le \hat{P}^p_B, P^s_B(\boldsymbol{\lambda}) \le \hat{P}^s_B \} \tag{6.43}$$

where $P^p_B(\boldsymbol{\lambda})$ and $P^s_B(\boldsymbol{\lambda})$ indicate the PU and SU blocking probabilities when the arrival rate is $\boldsymbol{\lambda} = (\lambda_p, \lambda_s)$, and $\hat{P}^p_B$ along with $\hat{P}^s_B$ are the maximum blocking probability requirements in order to ensure some QoS to both PUs and SUs.

6.5 Performance Evaluation

In the remainder, we provide numerical results for the evaluation of the proposed channelization mechanisms for the considered service types.

6.5.1 Parameter Setup

The primary channelization is fixed to M_p=8 channels and it is assumed the PUs request for TBSs. The secondary channelization set is $\mathcal{M}_s$=$\{1, 2, 4, 8, 16, 24, 32\}$, i.e. the maximum secondary channelization is $M_{s,max} = 32$. The total bandwidth to be partitioned is $W_T = 1.6$MHz and the average data size for VBS is, unless otherwise stated, $E[L]$=2Mbytes. The average service time of a TBS is 120s.

6.5.2 Numerical Results

In the following we explore the suitability of the proposed FCS and ACS for the cases when TBS and VBS characterizations apply.

6.5.2.1 TBS with FCS vs. ACS

Fig. 6.2 shows the arrival rate capacity limits [as defined in (6.43)] for the FCS and the ACS when SUs demand TBSs. The requirements in terms of maximum blocking probabilities are $\hat{P}_B^p = \hat{P}_B^s = 0.05$. At this point, perfect sensing conditions (i.e. $\delta = \varepsilon = 0.0$) are considered. As expected, increasing the channelization in FCS results in diminished blocking probabilities and, thus, larger arrival rate capacity limits. For the ACS case, since the maximum channelization is set to $M_{s,max} = 32$, the arrival rate capacity limit is the same than for FCS with $M_s = 32$.

Based on Fig. 6.2, the maximum offered PU arrival rate (for λ_s=0) is around 0.037 arrivals/s. We consider the case where the offered PU load is approximately 50% of the maximum by setting λ_p=0.02 in the following numerical evaluations. Fig. 6.3 shows the aggregate throughput (Γ) as defined in Section 6.4.4 for the case of the FCS and ACS when SU require TBSs. Results are only presented for arrival rates ensuring a maximum blocking probability of 0.05 for both PUs and SUs,

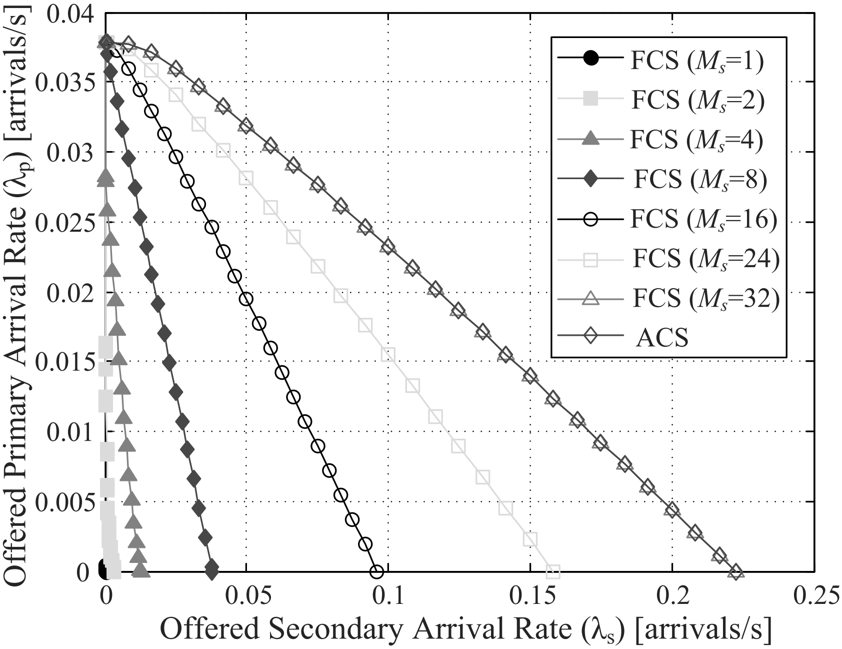


Figure 6.2: Arrival rate capacity regions for FCS and ACS with TBS.

which explains the disruption in the curves. In this way we ensure that throughput comparison is fair between the adopted schemes by ensuring the same QoS criteria reflected by (6.43). By observing Fig. 6.3 we note that, for the FCS cases, increasing the offered secondary arrival rate (λ_s) requires a corresponding increase in the channelization in order to satisfy the QoS requirements in terms of blocking. Nonetheless, this is at the cost of reducing the achievable throughput, since higher channelization decreases the granted bit-rate as reflected in (6.1). Moreover, it must be noted that the presence of PUs (for λ_s=0.02 the average number of PUs is around 2.4) makes infeasible for the FCS to use channelizations $M_s = 1$ along with $M_s = 2$ and, consequently, largely influences the achievable throughput by this scheme. On the contrary, the ACS achieves and overall improved throughput given its flexibility in allocating the maximum available spectrum, i.e. minimum channelization, and therefore higher bit-rates can be delivered by SUs in this case. Similar behaviour was obtained when the offered PU arrival rate (λ_p) was 25% and 75% of the maximum.

In terms of throughput-per-user experienced by SUs [Γ_u^s, defined in (6.41)], Fig. 6.4 shows the increased performance of the ACS with respect to the FCS in each operating range such that blocking probabilities are at most 5%. In particular, when the offered rate is low, the ACS is able to adjust the channelization such

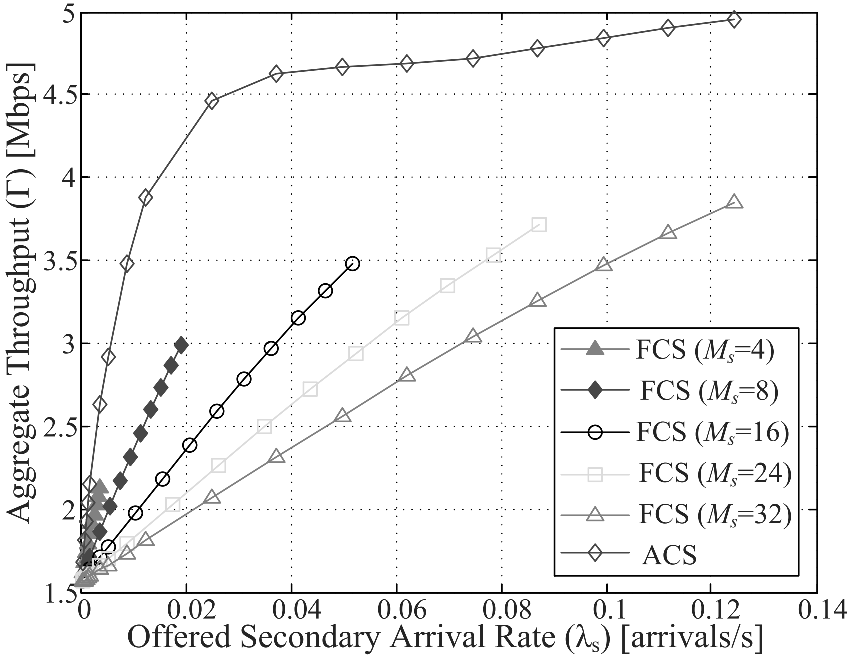


Figure 6.3: Aggregate throughput comparison between the FCS with $M_s = \{4, 8, 16, 24, 32\}$ and the ACS for the case of TBS. The offered PU arrival rate is $\lambda_p = 0.02$ arrivals/s.

that spectrum is not underutilized. When the offered arrival rate is increased, the performance of the ACS converges towards the FCS case with $M_s = 32$ since the maximum channelization for the ACS is also 32 channels. As expected, the FCS exhibits constant throughput-per-SUs values, insensitive to the offered SU arrival rate.

In order to understand the operation of the ACS, Fig. 6.5 shows the probability distributions of the SU channelization values for the ACS when increasing the offered SU arrival rate.. Then, for low SU arrival rate, the ACS primarily selects $M_s = 2$ until further channelization is needed to accommodate an increased number of SUs. Subsequently, the channelization value is increased to $M_s = 4, 8, 16, 24$. The average channelization [given by (6.42)] is also plotted in Fig. 6.5 reflecting the increase in channelization as the offered arrival rate increases. For comparison, we have also included the average channelization values for the FCS which are, as expected, constant.

Once assessed the large impact that PU spectrum usage has on the operation of SUs in the FCS, Fig. 6.6 plots, as a function of the offered PU arrival rate, the (optimum) SU channelization (M_s) that maximizes the average SU throughput, as

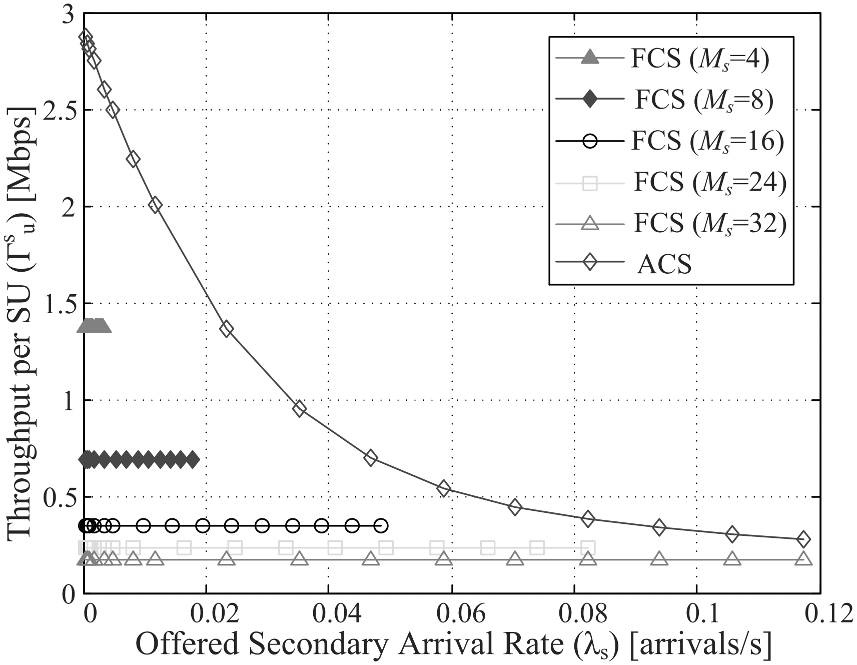


Figure 6.4: Average throughput-per-SU comparison between the FCS with $M_s = \{4, 8, 16, 24, 32\}$ and the ACS for the case of TBS. The offered PU arrival rate is $\lambda_p = 0.02$ arrivals/s.

defined in (6.37) for $k = s$. In addition, the corresponding average SU throughput is also plotted and compared to the case of the ACS. Results indicate the need to increase the channelization value as the number of PUs accessing the spectrum increases. In addition, this increase in M_s must carefully controlled in order to avoid excessive channelization which, in turn, would imply spectrum underutilization and, thus, reduced throughput. The ACS is able to provide an improved performance to that of the FCS with optimum channelization due to its inherent flexibility in allocating spectrum resources. It must be noted that blocking limitations (i.e. ensuring minimum blocking) do not apply in this particular case. In this sense, the ACS achieves improved throughput performance along with reduced blocking probabilities as compared to the FCS.

6.5.2.2 VBS with FCS vs. ACS

For the case of VBSs, the arrival rate capacity region is plotted in Fig. 6.7 for the case of $E[L]$=2Mbytes. A similar behaviour of the FCS is noted with respect to the TBS case in Fig. 6.2. In this sense, increasing the channelization value M_s

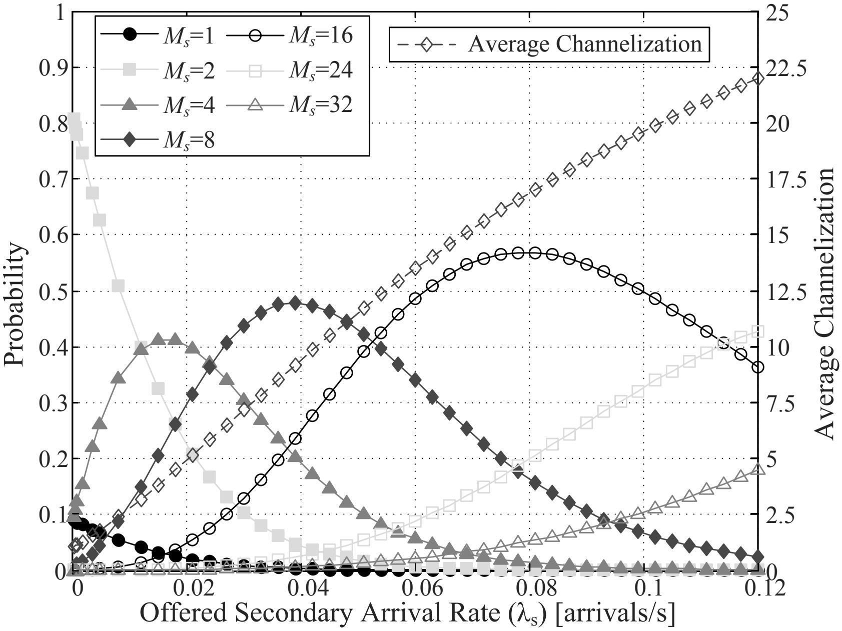


Figure 6.5: Probability distribution of channelization for the ACS for TBS. The offered PU arrival rate is $\lambda_p = 0.02$ arrivals/s.

allows higher secondary traffic to be supported while still guaranteeing the blocking probability constraints. As for the ACS, and different to the TBS case in Fig. 6.2, the arrival capacity region surpasses the FCS with M_s=32. This is due to the fact that SUs remain in the system until the data bulk $E[L]$ is transmitted and, given ACS is able to adapt bandwidth requirements (and therefore bit-rate) to current demands, it handles the transmission of VBSs more efficiently than the FCS case allowing more SUs to access the spectrum (i.e. increased λ_s values).

Fig. 6.8 shows the average transfer delay corresponding to the transmission of an average of $E[L]$ = 2Mbytes. The interpretation of the average transfer delay is analogous to the throughput-per-user, where the ACS is able to deliver the data load faster than the FCS given its ability for rate adaptation according to the current spectrum use.

6.5.2.3 Impact of average file size in VBS for FCS and ACS

In Fig. 6.9, the blocking probability of SUs is plotted against the average data length that a VBS should deliver. According to (6.12) and (6.27), an increase in

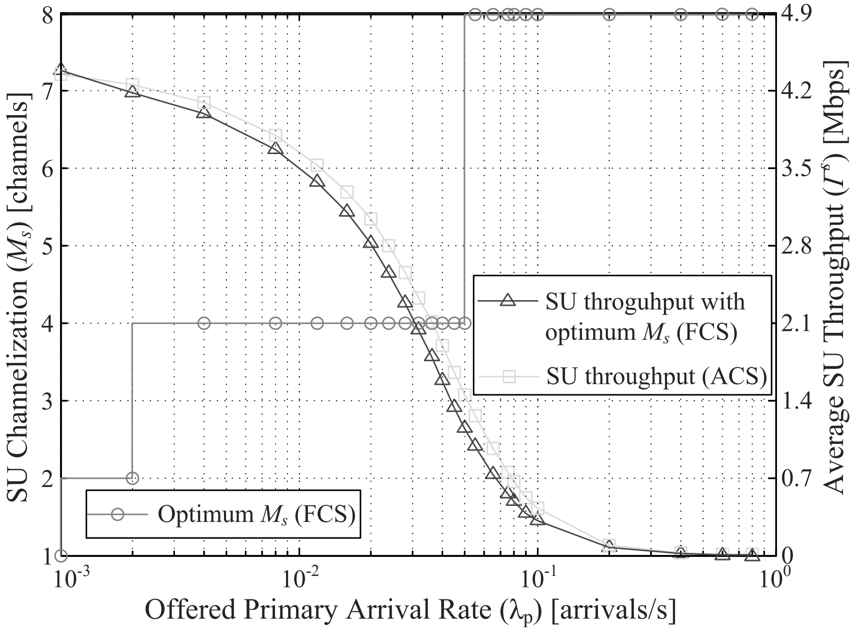


Figure 6.6: Optimum SU channelization and corresponding SU throughput for FCS along with the SU throughput for the ACS. The offered SU arrival rate is $\lambda_s = 0.05$ arrivals/s.

the average data length $E[L]$ is equivalent to increase in the offered secondary load, $T_S = \lambda_s \cdot E[t_s]$, since the average service time $E[t_s]$ increases with the data length. Then, as shown in Fig. 6.9, the ACS reflects an improved performance with respect to the FCS especially when the data bulk sizes are large.

6.5.2.4 Sensing error impact in FCS/ACS for TBS/VBS

In previous sections we have assumed perfect sensing and, thus, perfect knowledge of the spectrum occupation by PUs. We now consider the case where spectrum sensing is affected by errors in the form of miss-detection and false-alarm. In this sense, we compare the performance of the ACS and FCS when perfect sensing and erroneous sensing conditions apply. For the error case, assume the miss-detection probability is given by $\delta = 0.01$ and the false-alarm probability is given by $\varepsilon = 0.0974$ (see Chapter 5 for further details on these values). For these values, the sensing time is $T = 0.001$s/channel. This leads to a sensing efficiency of $\eta_{sens} = 0.93$ for the error case, whereas for the perfect sensing it will be assumed $\eta_{sens} = 1$.

Fig. 6.10 shows the aggregate throughput for the ACS and the FCS with $M_s = 32$ for TBS when perfect and erroneous sensing is carried out. As expected, the performance of the erroneous case is degraded with respect to the perfect sensing case for

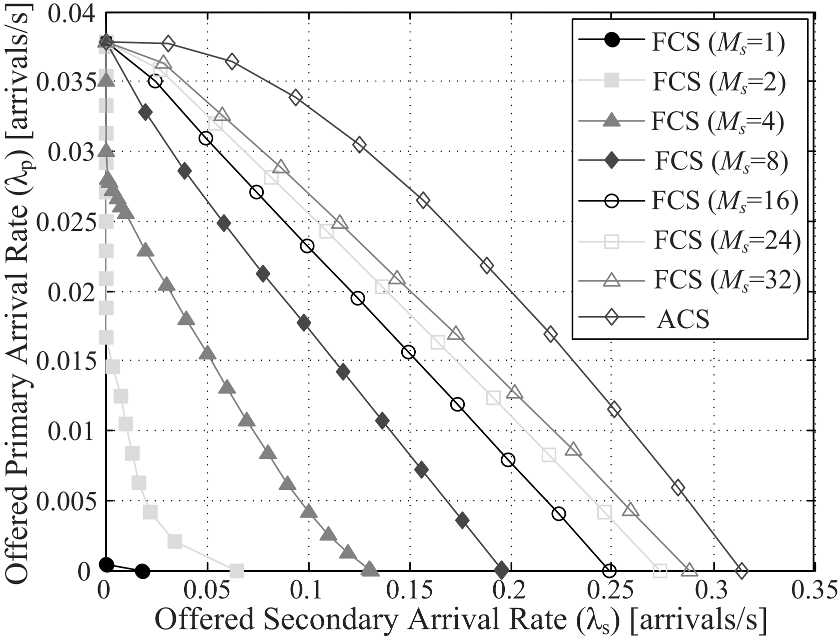


Figure 6.7: Arrival rate capacity regions for FCS and ACS with VBS.

both the ACS and FCS. However, note that for the FCS the degradation increases with the offered secondary traffic rate whereas for the ACS the degradation is fairly constant throughout the whole span of λ_s values. The false-alarm mainly affects the FCS by reducing the number of admitted SUs in the system given it detects PUs which are actually not occupying resources. This effect will be noticeable when the offered SU arrival rate increases since spectrum resources become scarce. As for the ACS, its operation is based on dynamically adjusting the channelization based on the detection of PUs. This may affect the performance of the ACS even if the offered SU traffic is low, thus explaining the similar degradation in the whole range of λ_s values.

Fig. 6.11 shows the average transfer delay for the case of the ACS and the FCS (with $M_s = 32$) under perfect and erroneous sensing conditions. Not surprisingly, the average transfer delay for the FCS remains insensitive to the offered SU arrival rate since the channelization is constant (i.e. $M_s = 32$). Moreover, sensing errors affect the average transfer delay in the case of the FCS mainly due to the sensing efficiency (η_{sens}) which reduces the throughput-per-SU and consequently the transfer delay. For the ACS, sensing errors in the form of false-alarm affects the adaptive channelization process by erroneously adjusting channelization values

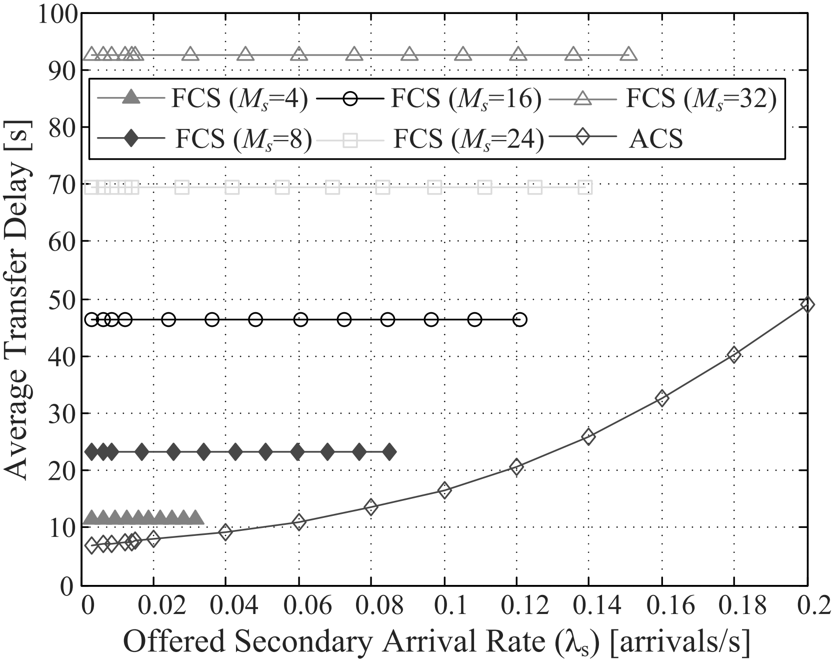


Figure 6.8: Average transfer delay for FCS and ACS with VBS. The offered PU arrival rate is $\lambda_p = 0.02$ arrivals/s.

M_s higher than those required, thus underutilizing the spectrum with a consequent reduction in throughput-per-SU and, hence, increased transfer delay. In addition, the sensing efficiency also contributes to an increase in the transfer delay, as noted also for the FCS.

6.6 Chapter Summary

In this chapter we have addressed the impact of channelization schemes in a primary-secondary opportunistic spectrum sharing system. Two channelization alternatives, namely FCS and ACS, have been proposed, modeled and evaluated in a Markovian framework. In addition, the service characterization of SUs has also been addressed considering two different types of services: a time-based service (TBS) and a volume-based service (VBS). Both service characterizations have been evaluated in the context of the abovementioned framework with the FCS and ACS. Numerical results indicate the suitability of the ACS with respect to the FCS given its ability to provide increased bit-rates at lower blocking probability for both

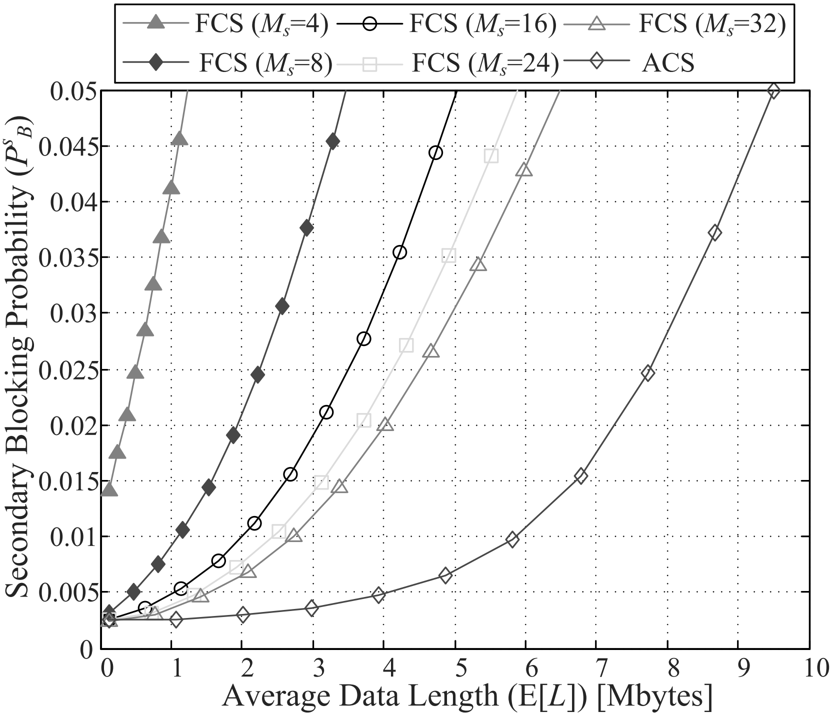


Figure 6.9: Secondary blocking probability for the FCS and the ACS against file size when VBS. The offered PU arrival rate is $\lambda_p = 0.02$ arrivals/s and the offered SU arrival rate is $\lambda_s = 0.05$ arrivals/s.

time-based and volume-based services. In addition, ACS has proven to be more resilient to sensing errors. Despite this, it should be noted that the operation of the ACS involves a higher complexity in terms of signaling which is necessary to update and inform all SUs about the channelization value M_s.

Bibliography

[1] Lianfeng Shen Xiaorong Zhu and Tak-Shing Peter Yum, “Analysis of Cognitive Radio Spectrum Access with Optimal Channel Reservation,” *IEEE Communications Letters*, vol. 11, no. 4, pp. 304–306, Apr. 2007.

[2] M. Raspopovic and C. Thompson, “Finite Population Model for Performance Evaluation Between Narrowband and Wideband Users in the Shared Radio Spectrum,” in *DySPAN’07*, Apr. 2007, pp. 340–346.

[3] Shensheng Tang and B.L. Mark, “An Analytical Performance Model of Opportunistic Spectrum Access in a Military Environment,” in *IEEE Wireless Communications and Networking Conference, 2008. (WCNC’08)*, Apr. 2008, pp. 2681–2686.

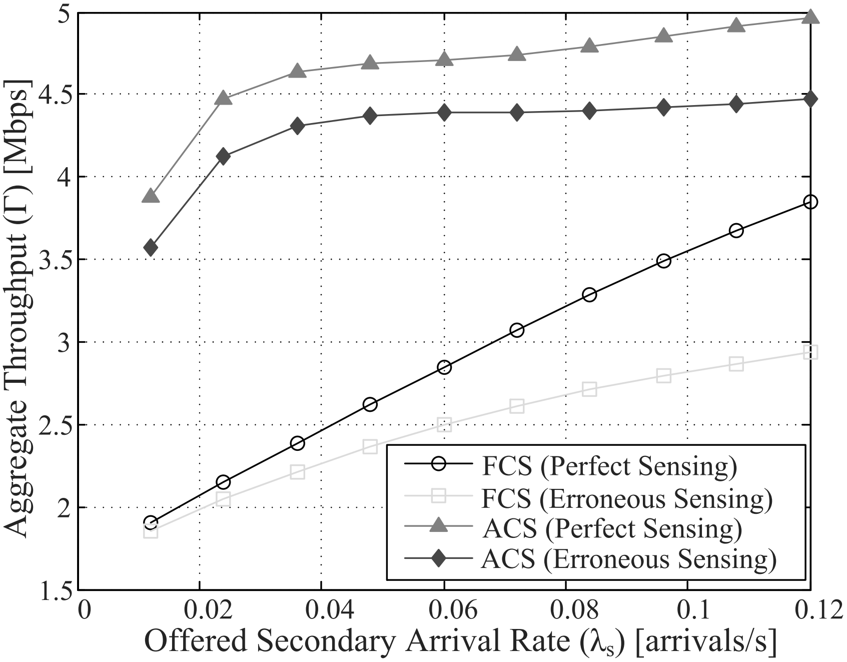


Figure 6.10: Aggregate throughput comparison between ACS and FCS (with $M_s = 32$) for TBS when perfect sensing and erroneous sensing applies. The offered PU arrival rate is $\lambda_p = 0.02$ arrivals/s.

[4] D.T.C. Wong, Anh Tuan Hoang, Ying-Chang Liang, and F.P.S. Chin, "Dynamic Spectrum Access with Imperfect Sensing in Open Spectrum Wireless Networks," in *WCNC'08*, Apr. 2008, pp. 2765–2770.

[5] W. Ahmed, J. Gao, H. A. Suraweera, and M. Faulkner, "Comments on "analysis of cognitive radio spectrum access with optimal channel reservation"," *Wireless Communications, IEEE Transactions on*, vol. 8, no. 9, pp. 4488–4491, Oct. 2009.

[6] D. Cabric, A. Tkachenko, and R. W. Brodersen, "Spectrum sensing measurements of pilot, energy, and collaborative detection," in *Military Communications Conference, 2006. MILCOM 2006. IEEE*, February 2007, pp. 1–7.

[7] A. Ghasemi and E.S. Sousa, "Collaborative spectrum sensing for opportunistic access in fading environments," in *DySPAN'05*, Nov. 2005, pp. 131–136.

[8] F. F. Digham, M. S. Alouini, and M. K. Simon, "On the energy detection of unknown signals over fading channels," in *Communications, 2003. ICC '03. IEEE International Conference on*, June 2003, vol. 5, pp. 3575–3579 vol.5.

[9] T. Hobfeld, A. Mader, and D. Staehle, "When Do We Need Rate Control for Dedicated Channels in UMTS?," in *Vehicular Technology Conference, 2006. VTC 2006-Spring. IEEE 63rd*, May 2006, vol. 1, pp. 425–429.

[10] Remco Litjens, Hans van den Berg, and Richard J. Boucherie, "Throughputs in processor sharing models for integrated stream and elastic traffic," *Perform. Eval.*, vol. 65, no. 2, pp. 152–180, 2008.

[11] R. B. Cooper, *Introduction to Queueing Theory*, North-Holland (Elsevier), 2nd edition, 1981.

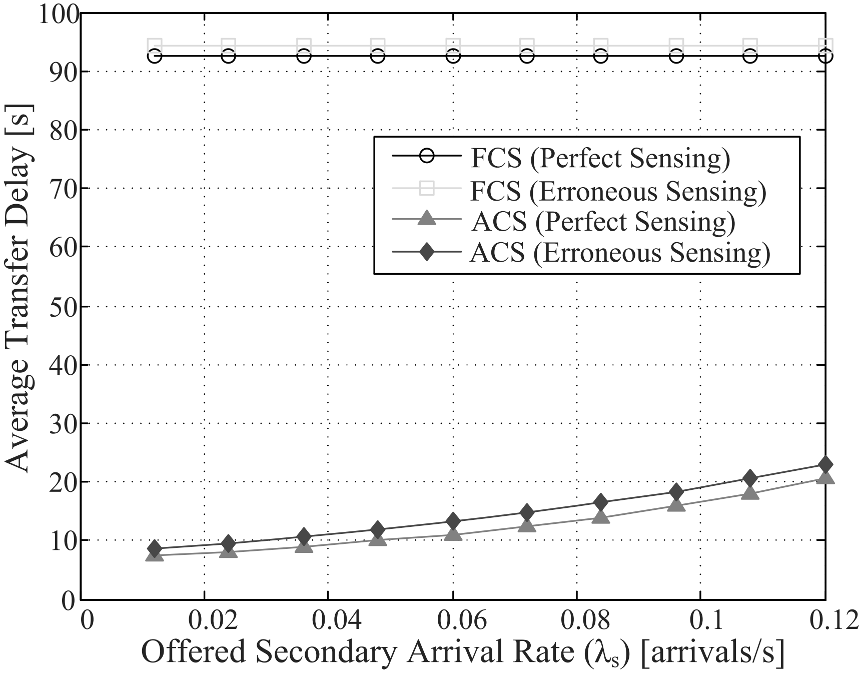


Figure 6.11: Average transfer delay between ACS and FCS (with $M_s = 32$) for VBS ($E[L] =$ 2Mbytes) when perfect sensing and erroneous sensing applies. The offered PU arrival rate is $\lambda_p = 0.02$ arrivals/s.

[12] William J. Stewart, *Introduction to the numerical solution of Markov chains*, Princeton University Press, 1994.

Chapter 7

Operating Point Selection for Primary and Secondary Users in Cognitive Radio Networks

This chapter addresses the problem of opportunistic access of secondary users to licensed spectrum in cognitive radio networks. In order to avoid interference to the licensed primary users, efficient spectrum detection methods need to be developed. For this purpose, in recent years several sensing techniques have been proposed to monitor and regulate the spectrum access to the shared spectrum resources. However, spectrum sensing may be affected by errors in the form of missed-detections (i.e., an occupied spectrum is erroneously detected as free) or false-alarms (i.e., a free spectrum is erroneously detected as occupied). These two magnitudes pose a tradeoff on the design of the spectrum sensing mechanisms meaning that low missed-detection can only be achieved at the expense of high false-alarm and vice versa. Thus, the network designers should adaptively tune the sensing techniques such that the highest perceived Quality of Service (QoS) is achieved by both primary and secondary users. In this chapter, a framework is introduced for determining the sensing operating points. Also the definition of Grade-of-Service (GoS) metrics is adopted to the case of primary/secondary users spectrum sharing. It is shown that the operating points of the sensing mechanisms can be easily adjusted according to the current traffic load of both primary and secondary users so that the perceived GoS is maximized. In addition, the Erlang Capacity of the spectrum sharing system for both primary and secondary users is also evaluated considering the effects of erroneous sensing.

7.1 Motivation and Problem Statement

Sensing-based spectrum discovery mechanisms, [1], may be affected by errors and, consequently, provide false information to the SU about PU spectrum occupancy. These errors are typically in the form of false-alarm (i.e., a free channel is erroneously sensed to be occupied) and miss-detection (i.e., an occupied channel is erroneously sensed to be free). As explained hereafter, by adequately choosing the operating points of the sensing mechanisms, a trade-off may be achieved between these two errors. As shown in the following, this means that low miss-detection is attained at the cost of increased false-alarm and, conversely, low false-alarm is achieved at the cost of high missed-detection.

According to the above, the missed-detection error will mainly affect the interference of PUs with SUs, that is, it will cause SUs accessing the spectrum already occupied by a PU. Consequently, resulting in a degraded operation for both PUs and SUs. On the other hand, false-alarm error will prevent SUs from accessing non-utilized spectrum, thus degrading the performance of these users. Bearing this in mind, a common approach has been to impose low values on the missed-detection probability so as to protect the PUs a the cost of reduced SU performance. Nevertheless, despite the PUs having strict access priority to spectrum resources, it may be necessary to guarantee some Quality of Service (QoS) requirement not only for PUs but also for SUs [2]. This becomes particularly true if the license holder (i.e., the entity ruling primary operation) demands payment for secondary access to the spectrum. Accordingly, the secondary system will expect some minimum return in terms of perceived service quality by SUs.

According to the above, the main contributions of this chapter are summarized in the following:

- The definition of the framework presented in Chapter 5 will enable to assess the potential gains that can be achieved by correctly selecting the sensing operating point which determines a particular value of the false-alarm and missed-detection probabilities.
- The suitability of the sensing operation points is determined using the Grade-of-Service (GoS) concept from “classical” telephone networks properly adapted to the primary/secondary spectrum sharing scenario. In this way, a metric is built taking into consideration the perceived service quality for both PUs and SUs.

7.1.1 Related Work

With respect to Grade-of-Service (GoS) definition in the context of opportunistic secondary access, some efforts were provided in [8, 12]. Nevertheless, the GoS metrics in [8] are strictly related to the blocking probability for both PUs and SUs disregarding other cross-effects between PUs and SUs such as the interruption probability (i.e., the service disruption of an SU due to PU activity) and the interference probability (i.e., the probability that both an SU and a PU share the same channel and thus cause interference). Also in [12], where a queueing framework is presented accounting for the dynamic allocation of primary and cognitive users, the GoS concept is exclusively related to the blocking probability. In this chapter, an improved definition for GoS is provided capturing the aforementioned effects, i.e., in addition to the blocking probability, the GoS accounts for the interference and the interruption probability of primary and secondary users respectively.

The remainder of the chapter is organized as follows. In Section 7.2, we present the considered spectrum sensing model and address some issues regarding the operation points of the spectrum sensing mechanisms. The performance metrics of interest for numerical evaluation purposes are explained and detailed in Section 7.3. Section 7.4 deals with the performance evaluation of the proposed model. Finally, conclusions are drawn in Section 7.5.

7.2 Spectrum Sensing Model

We assume that spectrum sensing over a given frequency band (or channel) is performed using energy detection techniques [10]. Such method consists in measuring the energy of the received waveform over a given bandwidth W (Hz) during an observation time-window T (s). The product $m = T \cdot W$ is usually referred to as the *time-bandwidth product*. Several works, among them [10, 11], have been devoted to determine closed-form analytical expressions for the false-alarm and misdetection (or, conversely, detection) probabilities under various channel conditions. Basically, the energy detection scheme performs a binary hypothesis on the occupancy of a band or channel: $\mathcal{H}_0$ if the channel is free and $\mathcal{H}_1$ if the channel is occupied. Then, the false-alarm and misdetection, ε and δ accordingly, can be defined as:

$$\varepsilon = \Pr\left[Y > \lambda | \mathcal{H}_0 \text{ is true}\right] \triangleq G_\varepsilon(\lambda) \tag{7.1}$$

$$\delta = \Pr\left[Y < \lambda | \mathcal{H}_1 \text{ is true}\right] \triangleq G_\delta(\lambda)\,, \tag{7.2}$$

where the decision statistic Y is compared to the decision threshold λ in order to determine the occupancy status of the channel. Accordingly, expressions for $G_\varepsilon(\lambda)$ and $G_\delta(\lambda)$ can be determined by accounting several channel conditions and cooperation schemes [10, 11]. For example, in the case of spectrum sensing in Rayleigh fading environments we have [11]:

$$\begin{aligned} G_\varepsilon(\lambda) &= \frac{\Gamma(m, \lambda/2)}{\Gamma(m)} && (7.3) \\ G_\delta(\lambda) &= e^{-\frac{\lambda}{2}} \sum_{k=0}^{m-2} \frac{1}{k!} \left(\frac{\lambda}{2}\right)^k + \left(\frac{1+\gamma}{\gamma}\right)^{m-1} \times \\ & \left(e^{-\frac{\lambda}{2(1+\gamma)}} - e^{-\frac{\lambda}{2}} \sum_{k=0}^{m-2} \frac{1}{k!} \left(\frac{\lambda\gamma}{2(1+\gamma)} \right)^k \right), && (7.4) \end{aligned}$$

where $\Gamma(.)$ and $\Gamma(.,.)$ are the complete and incomplete gamma functions respectively, and γ is the average signal-to-noise ratio.

Of particular interest is to determine the relationship between ε and δ through the so-called *Receiver Operating Characteristic* (ROC) curves where ε is plotted against δ for some given average signal-to-noise ratio γ and time-bandwidth product m. Formally, from (7.1) we can express $\lambda = G_\varepsilon^{-1}(\varepsilon)$ and by using (7.2) we obtain $\delta = G_\delta\left(G_\varepsilon^{-1}(\varepsilon)\right)$ which results in the ROC curve in Fig. 7.1a for the particular case of sensing in Rayleigh fading. More specifically, we isolate λ from expression (7.3) and substitute it in expression (7.4). Each point of such curve, hereon indicated by the pair $(\delta_0, \varepsilon_0)$, denotes a possible Operating Point (OP) for the sensing mechanism. In Fig. 7.1a a possible set of feasible OPs is marked by circles. Note that the existing trade-off between false-alarm and misdetection probability where the low values of ε are attained at high values of δ and vice versa.

By appropriately selecting a specific decision threshold value $\lambda = \lambda_0$ we obtain a particular value for the OP $(\delta_0, \varepsilon_0)$. It is worth mentioning that the function mapping between λ_0 and $(\delta_0, \varepsilon_0)$ is bijective, i.e., there is a one-to-one correspondence between λ_0 and $(\delta_0, \varepsilon_0)$ values in both directions.

For the sake of representation, rather than using the decision threshold λ (which depends on the decision statistic Y and, consequently, on the measured signal energy) we define the operating-point mix Θ, with $0 \leq \Theta \leq 1$, as:

$$\Theta \triangleq \frac{\log(\varepsilon/\varepsilon_{\min})}{\log(\delta/\delta_{\min}) + \log(\varepsilon/\varepsilon_{\min})}, \tag{7.5}$$

where $\varepsilon_{\min}$ and $\delta_{\min}$ are the minimum operating values for the false-alarm and misdetection probabilities respectively given by the ROC curve (see that $\varepsilon_{\min} = \delta_{\min} =$

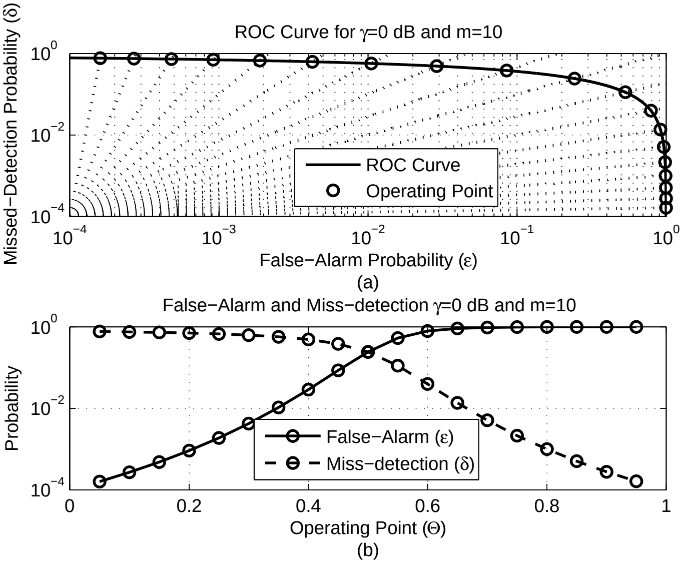


Figure 7.1: (a) ROC curves in Rayleigh fading channel and (b) tradeoff between false-alarm and miss-detection against the operating point.

10^{-4} in Fig. 7.1a). The values of $\varepsilon_{\min}$ and $\delta_{\min}$ can be regarded as the resolution of the sensing mechanism and consequently they are determined by the sensing equipment characteristics. Then, after some algebra manipulation, it follows that:

$$\delta = \delta_{\min} \left(\frac{\varepsilon}{\varepsilon_{\min}} \right)^{\left(\frac{1}{\Theta}-1\right)}, \tag{7.6}$$

which is plotted in Fig. 7.1a, for different values of $0 \leq \Theta \leq 1$, which results in the set of dashed lines crossing the origin of coordinates at $(\delta_{\min}, \varepsilon_{\min})$. For each particular value of $\Theta = \Theta_0$ we obtain a particular OP $(\delta_0, \varepsilon_0)$ which is represented by the circles in Fig. 7.1a denoting the intersection of the line equation given by (7.6) with the ROC curve.

In this way, we have a normalized parameterization through parameter Θ for the feasible OPs of the sensing mechanism. Note that, see Fig. 7.1b, for $0 < \Theta < 0.5$ we have that $\delta > \varepsilon$; for $\Theta = 0.5$ we obtain $\delta = \varepsilon$; and finally, for $0.5 < \Theta < 1$ we have $\delta < \varepsilon$. Then, the value of Θ will be used to represent the full range of possible cases and determine the most suitable OP for different traffic conditions.

In addition, for a longer time T devoted to sensing purposes, lower false-alarm and missed-detection probabilities can be attained. Indeed, this can be seen in Fig. 7.2, which plots the ROC curve for several values of the time-bandwidth product

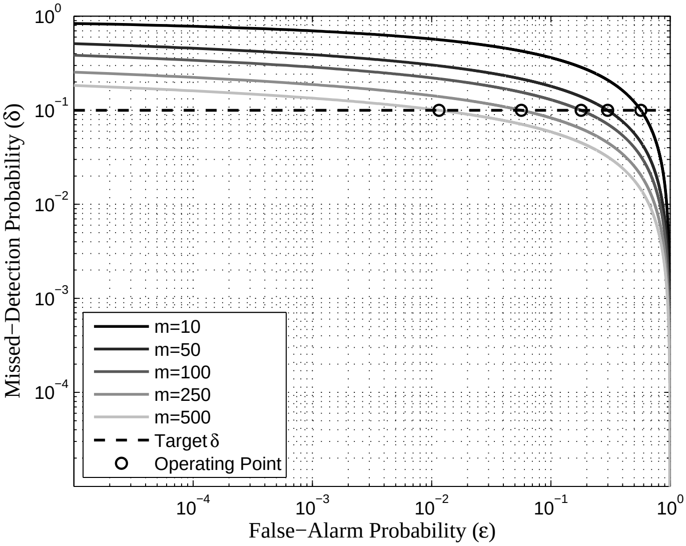


Figure 7.2: ROC curve varying the time-bandwidth product (m) for SNR of γ=0 dB.

(m). For a particular target missed-detection probability value ($\delta = 10^{-1}$) several corresponding false-alarm values are obtained as indicated by the OPs in Fig. 7.2.

7.3 Performance Metrics

The classical Grade-of-Service (GoS) concept in classical telephone networks [14] is adopted to the opportunistic spectrum access scenarios. The GoS metrics, introduced hereafter, will be computed from performance metrics derived from the steady state probabilities $P_{(i,j)}$ and $P'_{(i,j)}$ obtained as specified by the DTMC model in Chapter 5. In particular, the performance metrics of interest for the GoS computation are: primary and secondary users blocking probabilities, interruption probability and interference probability.

7.3.1 Blocking Probability

Blocking occurs whenever a new user cannot be assigned a channel given all channels are occupied, in the case of a PU, or thought to be occupied, in the case of an

SU. Accordingly, the blocking probability for PUs, P_B^P, can be computed from the true steady state probabilities, $P_{(i,j)}$, as:

$$P_B^P = \sum\nolimits_{j=0}^{C} P_{(C,j)} \, . \tag{7.7}$$

On the other hand, the SU blocking probability, P_B^S, is given by:

$$P_B^S = \sum\nolimits_{i=0}^{C} \sum\nolimits_{j=C-i}^{C} P'_{(i,j)} \, . \tag{7.8}$$

Notice that $P'_{(i,j)}$ is used instead of $P_{(i,j)}$ to indicate that secondary blocking may occur due to the sensing of all channels as occupied while this may in fact not be true.

7.3.2 Interruption Probability

Interruption of secondary service occurs whenever an SU is forced to release a channel, before its session has ended, due to primary activity. To compute the interruption probability, P_D, the average number of secondary users, N_s, can be considered:

$$N_s = \sum\nolimits_{S_{(i,j)} \in \mathcal{S}} j {\cdot} P_{(i,j)} \, , \tag{7.9}$$

can be interpreted as the average served SU traffic, i.e. $T_S^{served} = N_s$ [13]. Furthermore, it can be expressed as:

$$T_S^{served} = T_S \cdot \left(1 - P_B^S\right) \cdot \left(1 - P_D\right) \, , \tag{7.10}$$

meaning that the served traffic (T_S^{served}) is the offered traffic ($T_S = \lambda_S / \mu_S$) which is not blocked nor interrupted. By re-arranging (7.10) we obtain:

$$P_D = 1 - \frac{T_S^{\text{served}}}{T_S \left(1 - P_B^S\right)} = 1 - \frac{N_S}{\frac{\lambda_S}{\mu_S} \left(1 - P_B^S\right)} \, , \tag{7.11}$$

with P_B^S defined in (7.8) and N_S given in (7.9).

7.3.3 Interference Probability

The interference probability, P_I, is defined as the probability of being in state $S_{(i,j)} \in \mathcal{S}_c$ with the set $\mathcal{S}_c$ defined in (5.7), i.e., the probability that at least a channel is simultaneously occupied by both a PU and an SU, then:

$$P_I = \sum\nolimits_{S_{(i,j)} \in \mathcal{S}_c} P_{(i,j)} \, . \tag{7.12}$$

7.3.4 Grade-of-Service Definitions

Primary GoS (GoS^P) is derived from the blocking probability given in (7.7) and the interference probability in (7.12) as follows[1]:

$$GoS^P = \left(P_B^P + \omega_P \cdot P_I\right)/(1+\omega_P)\,, \tag{7.13}$$

where $\omega_P > 1$ is a weight factor indicating a higher penalty of interference with respect to blocking from the PUs' perspective.

We consider the SU blocking probability P_B^S, as in (7.8), along with the interruption probability P_D, as in (7.11), to define the secondary GoS (GoS^S) as follows:

$$GoS^S = \left(P_B^S + \omega_S \cdot P_D^S\right)/(1+\omega_S)\,, \tag{7.14}$$

with $\omega_S > 1$ the corresponding secondary weight factor indicating that interruption is more harmful than blocking.

Finally, we may define the aggregate GoS (GoS^A) as:

$$GoS^A = \left(GoS^S + \omega_A \cdot GoS^P\right)/(1+\omega_A)\,, \tag{7.15}$$

which jointly accounts for the individual GoS of both PUs and SUs and where we consider that the weight factor $\omega_A > 1$ will prioritize PU quality since they have strict precedence as primary (licensed) users of the shared spectrum. Note that ω_P, ω_S and ω_A should be chosen adequately in accordance to the expected perceived GoS of each user type (i.e., PU or SU). Nevertheless, note that these values are empirical and depend on the subjective perception of the grade of service metric.

7.4 Performance Evaluation

In this section we evaluate the proposed model by considering the performance metrics presented in Section 7.3. We consider a spectrum partition with $C = 8$ channels. The spectrum sensing periodicity, unless otherwise stated, is $\Delta T = 2$ seconds. For the spectrum sensing model, the cases considered in Figs. 7.1 and 7.2 are employed, i.e., the Rayleigh channel with signal-to-noise ratio of $\gamma = 0$dB. The weight factors for GoS computation in (7.13), (7.14), (7.15) are given by $\omega_S = 10$, $\omega_P = 20$ and $\omega_A = 10$.

[1] For convenience, a normalized version of the GoS $\in [0,1]$ is used, where $GoS \to 1$ means degraded operation while $GoS \to 0$ means improved operation.

7.4.1 Erlang Capacity

The Erlang Capacity of a system with limited resources refers to the maximum amount of offered traffic it may handle provided some quality of service requirements are met [15]. In the case of primary/secondary spectrum sharing, the interest is on the maximum primary and secondary traffic that can be offered such that some aggregate GoS requirement (i.e., accounting for both PUs and SUs) is satisfied. Mathematically, this can be expressed as:

$$\mathbb{E} = \left\{ (T_S, T_P) : QoS^A \leq QoS^{A*} \right\} , \tag{7.16}$$

with QoS^A defined in (7.15) and where QoS^{A*} indicates a target value for the aggregate GoS.

According to the above definition, the Erlang Capacity may be regarded as a region comprising the pairs of (T_S, T_P) offered traffic values which yield satisfactory GoS requirements. Accordingly, Fig. 7.3 shows the Erlang Capacity regions (defined as the areas below the Erlang capacity limits plotted in the figure) considering several time-bandwidth product (m) values. For the sake of comparison, the case where ideal sensing, and thus full awareness of PU activity, is also considered. As expected, the ideal sensing case translates into a larger Erlang Capacity region provided a better use of unoccupied spectrum resources can be achieved by SUs given their fully-aware information about PU spectrum usage. For the case of non-ideal sensing, the higher the time-bandwidth product (m) the better spectrum resources are being utilized. This is due to the fact that higher m values imply lower false-alarm probabilities (given the missed-detection probability is fixed) as can be observed from Fig. 7.2. In addition, it can be observed that for increased primary offered traffic values (e.g. $T_P > 2.5$ Erlangs), the Erlang Capacity region narrows towards lower values of offered secondary traffic (T_S), indicating that if a high number of PUs are occupying the spectrum (which have strict priority) then hardly no SUs are able to access the spectrum.

It is worth mentioning that the increase of the time-bandwidth product to reduce the false-alarm probability comes at the cost of increased sensing times which in turn degrade the achievable throughput as identified in [16, 17]. Therefore, although Fig. 7.3 indicates that the higher the value of m the better, the effect on throughput should be also taken into consideration.

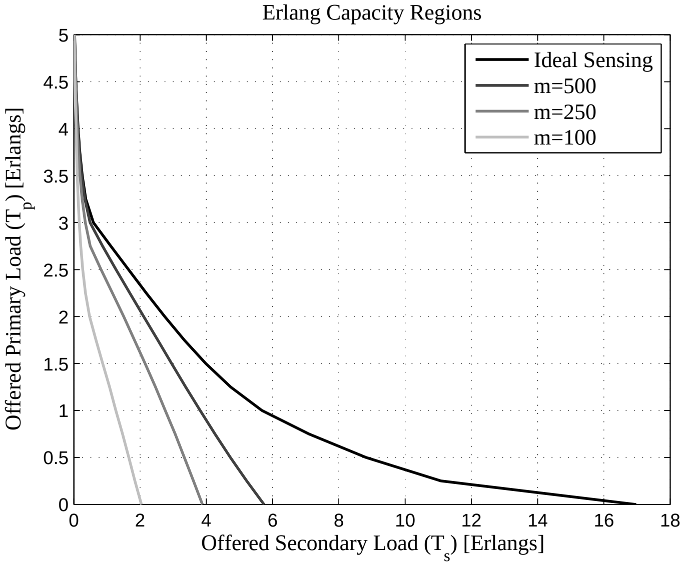


Figure 7.3: Erlang Capacity regions varying the time-bandwidth product (m) for target $GoS^{A*} = 5 \cdot 10^{-3}$.

7.4.2 Quality-of-Service Provisioning in Sensing-based Spectrum Sharing Scenarios

Concerning the experienced QoS for both PUs and SUs, GoS metrics, defined in Section 7.3.4, indicate that the interference and interruption probabilities are the major causes for PU and SU dissatisfactions, respectively. Then, we are interested in finding the most suitable OP of the sensing mechanism so that some satisfaction balance between PUs and SUs can be achieved. Accordingly, Figs. 7.4 and 7.5 show the interference and interruption probabilities against the sensing OP (Θ) under varying SU traffic and an offered PU traffic of $T_P = 2$ Erlangs. As for the interference probability in Fig. 7.4, the OP values of $\Theta \to 0$ indicate higher missed-detection probabilities (δ) as opposed to lower false-alarm probabilities (ε). Consequently, we note an increased interference due to an excess of SUs accessing the spectrum and erroneously detecting occupied channels as free. Conversely, when the value of Θ is increased towards 1, missed-detection decreases, thus, causing lower interference. The opposite behavior can be seen in Fig. 7.5 where the interruption probability is plotted. In this case, the low false-alarm (i.e., $\Theta \to 0$) benefits SUs since higher spectrum access chances are experienced. On the other hand, if false-alarm is increased (meaning $\Theta \to 1$), the detection of free channels as occupied will force SUs to defer their communication, thus, causing the interruption probability to rise. For both the interference and interruption probabilities,

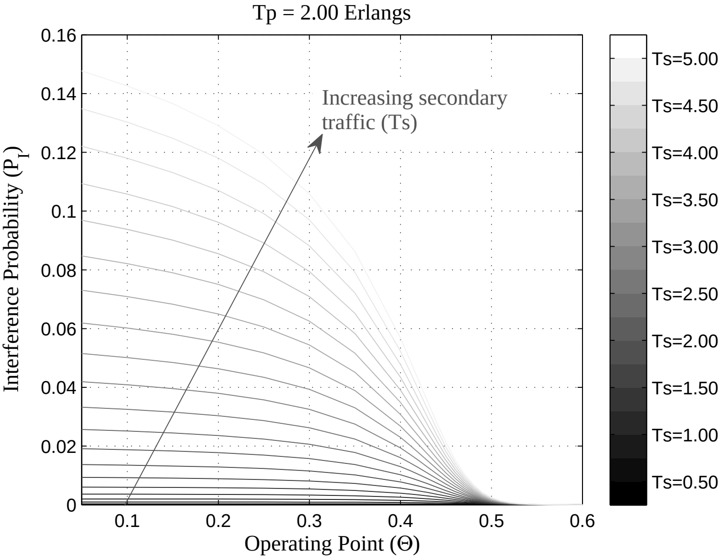


Figure 7.4: Interference probability against the OP for several traffic conditions.

the higher the secondary offered traffic (T_S), the higher degradation is observed.

In Fig. 7.6, the aggregate GoS (GoS^A) as defined in (7.15) is plotted for different offered traffic configurations (see Figs. 7.6(a), 7.6(b) and 7.6(c)).

By observing Fig. 7.6, we realize that by appropriately choosing the sensing OP (Θ) a minimum value of aggregate GoS can be achieved, thus improving the perceived satisfaction of both PUs and SUs. In addition, note that as the offered primary traffic increases as $T_P = 1, 2, 3$ Erlangs in Figs. 7.6(a), 7.6(b) and 7.6(c) respectively, the suitable OP value moves towards increased values of Θ in order to protect the increasing number of PUs in the system. This is in line with what depicted in Fig. 7.4 where values of $\Theta \to 1$ are required in order to lessen the interference probability experienced by PUs.

In addition to Fig. 7.6, the suitable OP values (i.e. those that minimize the perceived aggregate GoS) are provided in Table 7.1 for specific primary and secondary offered traffic. These values correspond to some of the star-shaped marks in Figs. 7.6(a), 7.6(b) and 7.6(c). It is worthwhile noting that, for low offered primary and secondary traffic, a range of OP values, denoted as [.,.] in Table 7.1, provides the minimum GoS^A. In these cases, interference with PUs is kept low and, thus, values of $\Theta \to 0$, which benefit SUs, can be selected. However, as secondary and primary traffic increases, so does the probability of interference (see Fig. 7.4), therefore

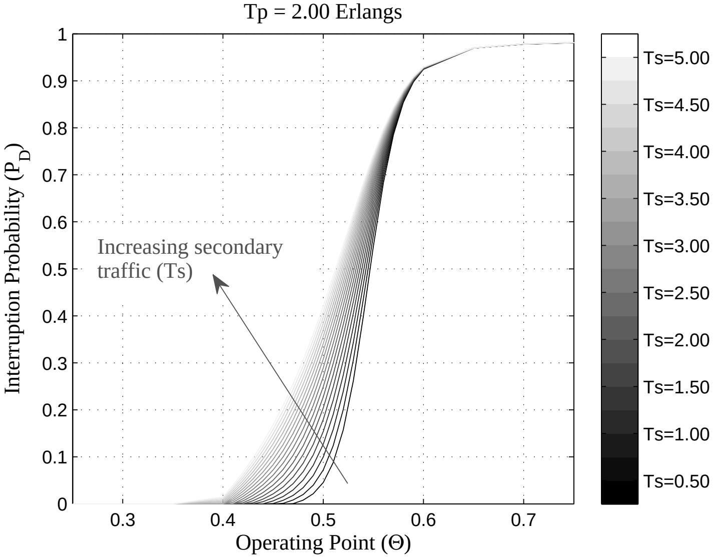


Figure 7.5: Interruption probability against the OP for several traffic conditions.

Table 7.1: Suitable Operating Points for several offered primary and secondary traffic values, T_s and T_p (expressed in Erlangs).

	Suitable OP (Θ^*)		
T_S	$T_P = 1.00$	$T_P = 2.00$	$T_P = 3.00$
0.50	[0, 0.45]	[0, 0.45]	0.44
1.00	[0, 0.45]	[0, 0.44]	0.44
1.50	[0, 0.43]	0.43	0.46
2.00	[0, 0.41]	0.43	0.46
2.50	0.40	0.44	0.47
3.00	0.40	0.45	0.48
3.50	0.41	0.46	0.48
4.00	0.42	0.46	0.49
4.50	0.42	0.46	0.49
5.00	0.43	0.47	0.49

increased values of Θ are needed in order to protect the PUs. Then, the higher the primary traffic, the less flexible is the selection of the suitable OP, which is, on the other hand, somewhat expected.

Finally, in Fig. 7.7, the effect of the sensing periodicity value (ΔT) on the perceived GoS is plotted. As the sensing periodicity increases, so does the interference probability given that secondary access is based on an older, and potentially out-of-date, spectrum occupancy information. Then, as ΔT increases, the suitable sensing

OP is shifted towards values of $\Theta \rightarrow 1$ given this protects PUs by decreasing the missed-detection probability.

7.5 Chapter Summary

In this chapter, a framework for the evaluation of sensing-based secondary spectrum access has been motivated and further presented. The main purpose of the framework is to determine the suitable sensing operating point so that requirements in terms of Grade-of-Service could be satisfied for both primary and secondary users. In this sense, the operating point of a sensing mechanism using threshold-based energy detection has been parameterized, given by Θ, in order to capture the existing tradeoff between missed-detection and false-alarm probabilities which negatively affect spectrum awareness. This tradeoff is tackled by means of defining a set of Grade-of-Service metrics which account for both the satisfaction level of PUs and SUs, and also on some aggregate satisfaction. In this way, performance results reveal that, by choosing an appropriate sensing operating point (Θ), the aggregate GoS can be minimized thus improving PU and SU perceived service quality. Moreover, the suitable operating point can be adjusted according to the current traffic load conditions and sensing periodicity cycles leading to an overall improved primary/secondary operation.

Bibliography

[1] D. Cabric, A. Tkachenko, and R.W. Brodersen, “Spectrum sensing measurements of pilot, energy, and collaborative detection,” *Military Communications Conference, 2006. MILCOM 2006*, pp. 1–7, Oct. 2006.

[2] Sooksan Panichpapiboon and Jon M. Peha, “Providing secondary access to licensed spectrum through coordination,” *Wirel. Netw.*, vol. 14, no. 3, pp. 295–307, 2008.

[3] Lianfeng Shen Xiaorong Zhu and Tak-Shing Peter Yum, “Analysis of Cognitive Radio Spectrum Access with Optimal Channel Reservation,” *IEEE Communications Letters*, vol. 11, no. 4, pp. 304–306, Apr. 2007.

[4] M. Raspopovic and C. Thompson, “Finite Population Model for Performance Evaluation Between Narrowband and Wideband Users in the Shared Radio Spectrum,” in *DySPAN’07*, Apr. 2007, pp. 340–346.

[5] Shensheng Tang and B.L. Mark, “An Analytical Performance Model of Opportunistic Spectrum Access in a Military Environment,” in *IEEE Wireless Communications and Networking Conference, 2008. (WCNC’08)*, Apr. 2008, pp. 2681–2686.

[6] Yiping Xing, R. Chandramouli, S. Mangold, and S.S. N, “Dynamic spectrum access in open spectrum wireless networks,” *IEEE Journal on Selected Areas in Communications*, vol. 24, no. 3, pp. 626–637, Mar. 2006.

[7] Beibei Wang, Zhu Ji, and K.J.R. Liu, "Primary-Prioritized Markov Approach for Dynamic Spectrum Access," in *DySPAN'07*, Apr. 2007, pp. 507–515.

[8] Pak Kay Tang, Yong Huat Chew, Ling Chuen Ong, and Francois Chin, "On the grade-of-services in the sharing of radio spectrum," *Cognitive Radio Oriented Wireless Networks and Communications, 2007. CrownCom 2007. 2nd International Conference on*, pp. 85–89, Aug. 2007.

[9] D.T.C. Wong, Anh Tuan Hoang, Ying-Chang Liang, and F.P.S. Chin, "Dynamic Spectrum Access with Imperfect Sensing in Open Spectrum Wireless Networks," in *WCNC'08*, Apr. 2008, pp. 2765–2770.

[10] F.F. Digham, M.-S. Alouini, and M.K. Simon, "On the energy detection of unknown signals over fading channels," *Communications, 2003. ICC '03. IEEE International Conference on*, vol. 5, pp. 3575–3579 vol.5, May 2003.

[11] A. Ghasemi and E.S. Sousa, "Collaborative spectrum sensing for opportunistic access in fading environments," in *DySPAN'05*, Nov. 2005, pp. 131–136.

[12] Huang Zhen, Huang Hai, Lin Fei, Yue Guangxin, and Xu Daxiong, "Dynamic channel allocation scheme for hybrid cognitive network," *Anti-counterfeiting, Security, Identification, 2007 IEEE International Workshop on*, pp. 179–183, April 2007.

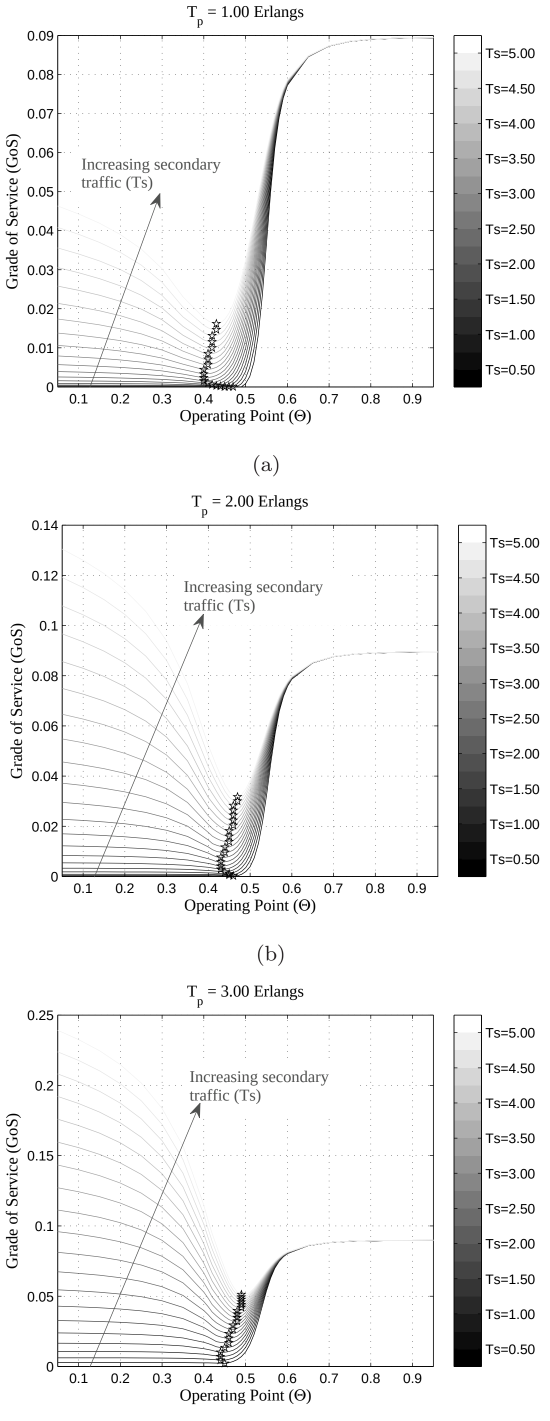
[13] D. Bertsekas and R. Gallager, *Data Networks*, Prentice-Hall, 1992.

[14] Jens Zander and S-L Kim, *Radio Resource Management for Wireless Networks*, Artech House, Inc., Norwood, MA, USA, 2001.

[15] Debasis Mitra and John A. Morrison, "Erlang capacity and uniform approximations for shared unbuffered resources," *IEEE/ACM Trans. Netw.*, vol. 2, no. 6, pp. 558–570, 1994.

[16] Won-Yeol Lee and Ian. F. Akyildiz, "Optimal spectrum sensing framework for cognitive radio networks," *Wireless Communications, IEEE Transactions on*, vol. 7, no. 10, pp. 3845–3857, October 2008.

[17] Ying-Chang Liang, Yonghong Zeng, E.C.Y. Peh, and Anh Tuan Hoang, "Sensing-throughput tradeoff for cognitive radio networks," *Wireless Communications, IEEE Transactions on*, vol. 7, no. 4, pp. 1326–1337, April 2008.



(a)

(b)

(c)

Figure 7.6: Grade of Service (GoS) as a function of the operating point for different PU and SU traffic load conditions.

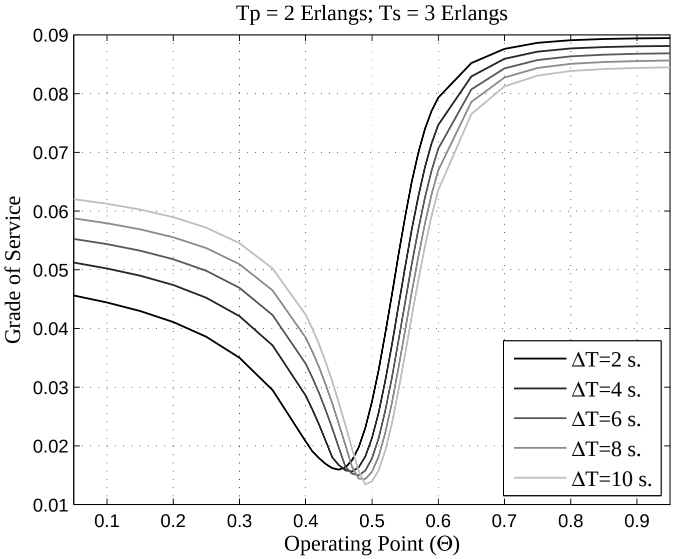


Figure 7.7: Grade of Service (GoS) as a function of the operating point for different sensing periodicities.

Part III

Concluding Remarks

Chapter 8

Conclusions

In summary, the results of this dissertation indicate that user capacity gains can be obtained by efficiently exploiting the usage of both radio and spectrum resources. Although each chapter includes a stand-alone summary regarding specific contributions, in what follows, and according to the problem formulation split *P1* and *P2* defined in Section 1.3, the summary of conclusions of this dissertation is provided.

8.1 How to select an appropriate RAT?

It is widely established that RAT selection procedures in multi-service/multi-access scenarios play a key role in the provision of CRRM functionalities.

Results have confirmed the validity and suitability of the proposed Markovian model which has been evaluated for several RAT selection policies and in different access-constrained scenarios including coverage limitation, multimode terminal availability and RAT-service compliance.

Results obtained in a non-access-constrained TDMA/WDCMA scenario with voice and data services (see Chapter 2) indicate that a trade-off between the average data throughput-per-user and the total blocking probability arises when comparing two counteracting service-based policies. That is, increased user capacity comes at the cost of reduced throughput-per-user for data services. This trade-off may be suitably handled using the load balancing (LB) policy which is able achieve both increased capacity while maximizing the throughput-per-user attained by data services.

In scenarios with access limitations due to coverage, terminal availability and RAT-service compliance (see Chapter 3), three key parameters influencing the performance of RAT selection were identified. First, the ratio between demanding services, or *traffic-mix*, largely affects the suitability of a particular service-based

RAT selection process. Second, *resource contention* among services in a single RAT, as opposed to RATs that only uphold one service-type, causes service-greedy allocation principles to appear. Both service-mix and resource contention issues must be jointly considered when allocating multiple services on multiple RATs. Third, *RAT-Service eligibility* determines if a particular RAT is is "selectable" in the RAT selection procedure provided coverage availability, terminal support and RAT-service compliance is satisfied.

Based on the influencing parameters described above, specific guidelines in a multi-access/multi-service scenario considering GSM, UMTS and WLAN RATs along with voice and data services were provided. Summarizing (the reader is referred to Chapter 3 Section 3.8), when perfect eligibility applies: Service-greedy, i.e. either voice-greedy or data-greedy, allocation principle is desirable when favorable traffic-mix conditions apply. Hence, voice-greedy (data-greedy) allocation is suitable in voice-dominant (data-dominant) traffic situations. Additionally, in the same perfect-eligibility scenario, late-contention is preferable to early-contention in RATs where services compete for resources. As for the case of limited-access scenarios the rule *access-limited-first* provides a good guiding principle for the allocation of multiple services to multiple RATs. This rule should be used in combination with the aforementioned service-greedy rule making sure that favorable service-mixes apply. Finally, LB policy exhibits an overall improved behavior regardless of the traffic-mix and RAT eligibility. It is then suggested when the service distribution is unknown or vaguely estimated.

An analytical probabilistic characterization for radio access congestion in multi-RAT environments has also been presented (see Chapter 4). Specific analytical expressions for the congestion probability were provided for both TDMA and WCDMA in the presence of voice and data service requests. Performance evaluation, considering perfect-eligibility, was carried out using three RAT selection policies: Load Balancing (LB), Service-Based (SB) and Congestion-Aware (CA). With the proposed framework, the impact of different RAT selection policies on the congestion probability can be measured and QoS degradation is assessed. The higher flexibility exhibited by LB causes the congestion probability and throughput-per-user is improved with respect to SB. Nonetheless, LB disregards the impact of congestion probability on the throughput attained in the WCDMA RAT. Then, using congestion information, as with CA policy, can lead to a better performance in terms of both congestion probability and throughput. In addition, the use of congestion information as a guiding principle for initial RAT selection can also prevent from high blocking situations which result from applying tighter AC mechanisms in order to reduce such congestion.

8.1.1 Future Work

Future efforts in the context of problem *P1* should be devoted to mobility aspects by means of the implementation of vertical handover strategies into the Markov model. In addition, a further characterization of the radio propagation and interference environment would be also of great interest, which could eventually lead to alternative RAT selection criteria to those proposed in this dissertation. For example, considering a mix of outdoor and indoor users, along with the availability of femto-cells, may result into new allocation principles. Regarding the considered service-type case studies, i.e. generically voice and data, these could be extended to include specific features and effects of real services such as, e.g., TCP for web-browsing or FTP sessions. In addition, the studied RAT cases should be extended to adopt other specific technologies such as HSPA, WiMAX, LTE, LTE-Advanced, etc. This dissertation has developed allocation principles under the assumption of a single operator, whereas multi-operator environments with certain degree of cooperation would be interesting.

8.2 How to efficiently share spectrum between licensed and unlicensed users?

In this dissertation, a generalized and flexible DTMC-based framework for the definition and evaluation of opportunistic shared spectrum scenarios has been presented. The framework considers an uncoordinated operation between primary and secondary networks where primary spectrum occupancy information is retrieved through sensing mechanisms. Model validation was assessed by comparison with a system-level simulator. Furthermore, model limitations were also determined and boundaries on limiting parameters provided.

The existing tradeoff between the sensing accuracy and the exhibited secondary throughput has also been studied. As expected, an increased sensing accuracy through longer sensing periods will, at a given point, not payoff the degradation obtained in terms of throughput since less time is then devoted to the actual data transmission. Results revealed that the sensing time (equivalently, the time-bandwidth product) can be conveniently adjusted in order to maximize throughput. In addition, if the number of channels to be sensed is large, sensing procedures will take longer to determine the spectrum occupancy of the whole band, consequently reducing the sensing efficiency. Then, the time-bandwidth product should be reduced when increasing the bandwidth on which PUs and SUs operate.

The impact of the spectrum awareness periodicity (*how often do we sense?*) has also been evaluated. As expected, increased sensing periodicity causes higher collision events between PUs and SUs to happen. In addition, improved sensing accuracy degrades the experienced interference by allowing an increased number of SUs in the system. As shown in Chapter 5, secondary operation can be optimized by choosing adequate values for the time-bandwidth product (i.e. the time devoted to sensing) in such way that the throughput is maximized.

The impact of channelization schemes in a primary-secondary opportunistic spectrum sharing system has been also addressed in this dissertation (see Chapter 6). Two channelization alternatives, namely Fixed Channelization Scheme (FCS) and Adaptive Channelization Scheme (ACS), have been proposed, modeled and evaluated in a DTMC framework. Secondary user service characterization has also been addressed considering a time-based service (TBS) and a volume-based service (VBS). Both service characterizations have been evaluated in this framework under the FCS and the ACS. Results suggest the suitability of the ACS with respect to the FCS given its ability to provide increased bit-rates at lower blocking probability (i.e. higher user capacity) for both TBS and VBS. In addition, ACS has proven to be more resilient to sensing errors.

The problem of appropriately selecting the operating point of a sensing-based secondary spectrum access scheme has also been addressed. The main objective was to determine the suitable sensing operating point so that Grade-of-Service requirements could be satisfied for both primary and secondary users. Accordingly, parameterization was used to capture the existing trade-off between missed-detection and false-alarm probabilities in an energy detector spectrum awareness mechanism. Missed-detection increases interference between PUs and SUs, thus degrading the PUs which require non-harmful operation. On the contrary, false-alarm causes spectrum overlook and thus spectrum opportunities are missed causing a degraded operation for the SUs. This trade-off is tackled by means of defining a set of Grade-of-Service metrics which account for both the satisfaction level of PUs and SUs, and also on some aggregate satisfaction. Performance results reveal that, by choosing an appropriate sensing operating point, the aggregate GoS can be minimized thus improving PU and SU perceived service quality. Moreover, the suitable operating point can be adjusted according to the current traffic load conditions and sensing periodicity cycles leading to an overall improved primary/secondary operation.

8.2.1 Future Work

Future work regarding spectrum sharing should be devoted to further investigate practical cases, with specific technologies. For example, spectrum sharing over TV bands based on the IEEE 802.22 standard. In addition, cooperative mechanisms for spectrum awareness and spectrum sharing are also of concern since they are expected to introduce capacity gains at the cost of increased complexity. The work in this dissertation considers ideal and static radio conditions concerning the use of radio channels shared between primary and secondary users. It would be then desirable to introduce some dynamics on the availability of spectrum bands due to propagation and spatial location of terminals.

Part IV

Appendices

Appendix A

Project Involvement

This appendix briefly describes the projects in which the author of this dissertation has been involved with.

A.1 Project COSMOS

A.1.1 Project Information

Relevant project information is given in the following:

- Project Name: Calidad de servicio extremo a extremo y flexibilidad espectral en redes móviles heterogéneas (COSMOS) (Proyecto CICYT Ref. TEC2004-00518)
- Funding Entity: Ministerio de Educación y Ciencia, Spain.
- Duration: 12/2004 - 12/2007.
- Principal Researcher: Dr. Josep Oriol Sallent Roig
- URL: `http://www.cosmos.upc.edu/`

A.1.2 Project Description

This project focuses on mobile wireless heterogeneous networks (including UMTS, GERAN and WLAN) with the ultimate goal to provide end-to-end QoS for IP-based multimedia services in an efficient way. Within this framework, the purpose of this project is to address the different problems at different levels leading to gain insight into an ultimate global solution. In particular, the objectives of the project can be stated as:

a. To analyze, evaluate and propose architectures to support end to end QoS, with special emphasis in the component dealing with common radio resource management (CRRM), the core network QoS management as well as the inter-working architecture among the different radio access technologies.

b. To analyse, develop and propose resource management strategies in the framework of integrated heterogeneous networks with the ability to provide quality, capacity and coverage targets in the most efficient way. Prior to define CRRM strategies it is required to analyse, develop, propose and evaluate radio resource management strategies for every single technology considered (i.e. GSM/GPRS/EDGE, UMTS and WLAN). The proposed algorithms will consider the impact of flexible spectrum management and allocation through Spectrum Management (SM) algorithms.

c. To demonstrate, by means of IP multimedia applications, the suitability of the developed elements through a laboratory testbed implementation, supporting the main outcomes of the project.

On the other hand, since the successful deployment of heterogeneous networks architectures as well as RRM, CRRM and SM algorithms largely depends on their economical viability, this component will be closely taken into account in the envisaged research, thus considering the technic-economic perspective.

A.2 Project COGNOS

A.2.1 Project Information

Relevant project information is given in the following:

- Project Name: Gestión cognitiva de recursos radio y espectro radioeléctrico en redes móviles heterogéneas con provisión de calidad de servicio extremo a extremo (COGNOS) (Proyecto CICYT ref. TEC2007-60985)
- Funding Entity: Ministerio de Educación y Ciencia, Spain.
- Duration: 10/2007 - 09/2010.
- Principal Researcher: Dr. Josep Oriol Sallent Roig
- URL: `http://www.cognos.upc.edu/`

A.2.2 Project Description

This project departs from a heterogeneous mobile network scenario where end-to-end QoS requirements should be efficiently provided to IP-based multimedia services. In this framework, several global and integrated solutions are proposed to tackle the problems involved in these scenarios. In particular, the main objectives of this project can be stated as:

a. Propose and develop an integrated framework for radio resource and spectrum management (RRM and SM) that allow an efficient and flexible use of available resources, clearly identifying the involved mechanisms and the interactions among them, and also considering multi-operator environments. Due to the inherent dynamics of the system operation (traffic, propagation, mobility, etc.) adopted strategies will incorporate cognitive features, thus exhibiting auto-adaptation and robustness.

b. Propose and develop supporting end-to-end QoS architectures consistent with the capacities and particularities of the different involved communication segments: (1) a flexible and heterogeneous radio interface, (2) a transport network from the access point towards the IP-based backbone network, and (3) an IP-based backbone network bearing in mind the possibility that (2) and (3) may be regarded as a single segment.

c. Analyze, propose, develop and evaluate practical, precise and flexible radio and spectrum resource management strategies that jointly fulfill requirements in terms of quality, capacity and coverage in the most efficient way as possible. Prior to the definition of such integrated strategies it is necessary to analyze, propose, develop and evaluate individual RRM strategies for the different access technologies forming the heterogeneous network. Only after this step, the integrated vision will be successfully accomplished.

d. Demonstrate, through multimedia IP-based applications, the usefulness of the considered strategies by developing a testbed that is able support the aforementioned features.

This project will ensure that the proposed solutions regarding heterogeneous network architecture, inter-working between technologies, development of the transport and backbone networks along with their corresponding algorithms for resource management are aligned with the main standardization forums, in particular the 3GPP (3rd Generation Partnership Project) and the IETF (Internet Engineering Task Force).

A.3 Project EVEREST

A.3.1 Project Information

Relevant project information is given in the following:

- Project Name: Evolutionary Strategies for Radio Resource Management in Cellular Heterogeneous Networks (Ref. IST-2002-001858)
- Funding Entity: EU Sixth Program Framework (FP6)
- Duration: 01/2004 - 12/2005.
- Principal Researcher: Dr. Fernando Casadevall Palacio (UPC)
- URL: `http://www.everest-ist.upc.es/`

A.3.2 Project Description

The objective of the EVEREST project is to devise and assess a set of specific strategies and algorithms for access and core networks, leading to an optimized utilization of scarcely available radio resources for the support of mixed services with end-to-end QoS mechanisms within heterogeneous networks beyond 3G.

The provision of beyond 3G heterogeneous network topologies is conceptually a very attractive notion; however, it is a challenge to accomplish an efficient network design. In this context, Radio Resource Management (RRM) strategies are responsible for an utmost efficient utilization of the air interface resources in the available Radio Access Networks (RANs). EVEREST will provide tangible contributions towards a heterogeneous realization of 2G/2.5/3G (e.g. GERAN, UTRAN) and 3.5G networks with the inclusion of newly emerging RANs (e.g. WLAN for vertical coverage extensions). The potential inclusion of location information in RRM design, as well as some forms of RAN sharing, will be considered as additional examples of the medium and long term research focus of EVEREST.

In order to accomplish these objectives, the project evolves around two main activities:

- Algorithmic development and simulation by means of advanced simulation tools.
- Demonstration of the technology by means of implementing real-time testbeds for proof of concepts

It is a further purpose of the project to contribute actively to the different standardization fora. In that sense, the proposed solutions will be compliant with and aligned to standardization activities in the field, e.g. 3GPP, IETF, IEEE. Moreover, the results obtained in EVEREST are expected to be of significant momentum, the beneficiaries to which are service-providers, operators, manufacturers and end-users.

The research challenges, to be tackled by EVEREST project, can thus be summarized as follows:

- To identify, propose, simulate, assess and validate advanced RRM algorithms for GERAN and UMTS as well as novel radio concepts beyond 3G.
- For heterogeneous networks, to develop Common RRM (CRRM) algorithms between access technologies focused on UTRA and GERAN. Both for tight and very tight coupling will be considered.
- To consider other technologies that can be a complement to GPRS/UMTS, such as:
  - WLAN for indoor hotspots.
  - Different types of repeaters, acting as coverage extensions.

- To support end-to-end QoS in a heterogeneous wired and wireless mobile environment. To this end, the investigation about the relationship between the core network Band-width Broker (BB) and the RRM & CRRM entities for a plethora of RANs (UMTS, GERAN and WLAN) becomes of prime importance.
- To demonstrate the benefits of the developed RRM and CRRM algorithms by means of multimedia IP based applications over a real time testbed.

A.4 Project AROMA

A.4.1 Project Information

Relevant project information is given in the following:

- Project Name: Advanced Resource Management Solutions for future all IP Heterogeneous Mobile Radio Environments (Ref. IST-4-027567)
- Funding Entity: EU Sixth Program Framework (FP6)
- Duration: 01/2006 - 12/2007.
- Principal Researcher: Dr. Fernando Casadevall Palacio (UPC)
- URL: `http://www.aroma-ist.upc.edu/`

A.4.2 Project Description

The objective of the AROMA project is to devise and assess a set of specific resource management strategies and algorithms for both the access and core network part that guarantee the end-to-end QoS in the context of an all-IP heterogeneous network.

AROMA project aims not only to asses and maximize the potential benefits coming from the medium-term evolution of the considered radio-access technologies (e.g. HSDPA/HSUPA; MBMS) but in parallel also to promote and investigate potential benefits coming from a long-term evolution towards an all IP heterogeneous mobile and wireless network architecture. In that context, the RAN architecture should be also evolved to accommodate future IP-based networks, which allow a common transport even in different access networks, simple resource management, and easy heterogeneous inter-working.

On the other hand, in order to support end-to-end QoS in a heterogeneous wired and wireless mobile environment, an appropriate interaction between the QoS management entities of the core network (CN) and the Common Radio Resource Management (CRRM) in the radio part is crucial. These kinds of issues are extensively covered in the project.

Last but not least, it is also prime important to carry out economic evaluation on the impacts of the novel architecture solutions considered by the project

In summary AROMA aims at providing tangible contributions, in terms of resource management, for the future all IP heterogeneous wireless systems, which will take into account 2G/2.5/3G (e.g. GERAN, UTRAN) and 3.5G networks (e.g. HSDPA), including the newly emerging RAN technologies (e.g. WLAN , WIMAX) and services, for the 2010-2015 time frame.

In order to accomplish these objectives, the project evolves around two main activities:

1. Algorithmic development and simulation by means of advanced simulation tools, and

2. Demonstration of the technology by means of implementing real-time testbeds for proof of concepts.

It is a further purpose of the project to contribute actively to the different standardization fora. Results obtained in AROMA are expected to be of significant momentum, the beneficiaries to which are service-providers, operators, manufacturers and end-users.

A.5 Project E^3

A.5.1 Project Information

Relevant project information is given in the following:

- Project Name: End-to-End Efficiency (Ref. ICT-2007-216248)
- Funding Entity: EU Seventh Program Framework (FP7)
- Duration: 01/2008 - 12/2009.
- Principal Researcher: Mr. Wolfgang König (Alcatel-Lucent)
- URL: `https://ict-e3.eu/`

A.5.2 Project Description

The End-to-End Efficiency (E^3) project is an ambitious FP7 EC Large Scale Integrating Project (IP) aiming at integrating cognitive wireless systems in the Beyond 3G (B3G) world, evolving current heterogeneous wireless system infrastructures into an integrated, scalable and efficiently managed B3G cognitive system framework. The key objective of the E^3 project is to design, develop, prototype and showcase solutions to guarantee interoperability, flexibility and scalability between existing legacy and future wireless systems, manage the overall system complexity, and ensure convergence across access technologies, business domains, regulatory domains and geographical regions.

Cognitive radio systems are seen by many actors of the wireless industry as a core technical evolution towards exploitation of the full potential of B3G systems. It is under way to revolutionize wireless communications just as the PC revolution did in its domain. E^3 will ensure seamless access to both applications and services as well as exploitation of the full diversity of corresponding heterogeneous systems, in order to offer an extensive set of operational choices to the users (e.g. seamless experience), application and service providers (e.g. fast deployment of enhanced features and services in reduced time frames), operators (e.g. network management, operation and maintenance), manufacturers (e.g. wider market and migration to new standards) and regulators (e.g. increasing spectrum efficiency). E^3 will optimize the use of the radio resources and spectrum, following cognitive radio and cognitive network paradigms (autonomic management, learning, experience, knowledge as well as context, profiles, policies). The management functions will be distributed over different network elements at various levels of the system topology. A corresponding management agility will be required for supporting the most efficient use of the cooperating technologies, at local, regional, and global levels.

E^3 will make converge both cognitive radios and cognitive networks from technical, business, regulatory and standardization perspectives. Consequently, future wireless deployments will be conceived on a fully cognitive system basis. The business models research will permit selection of the most relevant concepts and solutions ensuring development and future deployment of sustainable cognitive radio systems. The regulatory research will further and support the adoption of E^3 concepts and solutions in the world radio regions. The research will also help the evolution of the regulatory framework in order to cope with the future development of more flexible spectrum usage (e.g. based on modified administrative mechanisms or market or technology driven), that will only be possible if suitable solutions for managing and controlling complex heterogeneous systems are in place. E^3 will build on the IST E2R research results on reconfigurable equipment extending the corresponding concepts towards the design of a wireless cognitive radio system in which network entities will be able to self-adapt to a dynamically changing context (i.e., user traffic demands, etc.). The focus will furthermore be on the evolution of wireless systems in an evolutionary, non-disruptive way, by integrating existing wireless radio standards into a common framework with user devices be able to reconfigure and maintain one or multiple links simultaneously, and contributing to currently active/emerging standardization bodies with a focus on key convergence enablers. In particular, ongoing standardization on IMT-Advanced related radio and cognitive systems are targeted, with contributions enabling the convergence towards a future harmonized and interoperable wireless landscape. E^3 will devise structuring rules for the definition and design of next generation of various standards (e.g. IEEE 802.16/11, P1900, ETSI, 3GPP....) allowing a seamless use of these standards to fulfill the scenarios of the current definitions of 4G systems at lower cost and complexity, and for a better spectrum efficiency (plug and play *lego-blocks* for standards).

List of Abbreviations

The list of used abbreviations in this dissertation follows. In the cases where a single acronym has several meanings, the chapter in which each meaning is adopted appears. In addition, throughout the dissertation, pluralization of abbreviations is denoted by adding a lower-case ’s’ at the end, as in, e.g., RAT vs. RATs.

2G	Second Generation
2.5G	Second-point-Five Generation
2.75G	Second-point-Seventy-Five Generation
3G	Third Generation
3GPP	Third Generation Partnership Project
3.5G	Third-point-Five Generation
4D	Four-Dimensional
4G	Fourth Generation
ABC	Always Best Connected
AC	Admission Control
ACK	Acknowledgment
ACS	Adaptive Channelization Scheme
AIFSN	Arbitration Interframe Space Number
AP	Access Point
B3G	Beyond 3G
BC	Book Chapter
BE	Best-Effort
BP	Blocking Probability
BS	Base Station
BSC	Base Station Controller
CA	Conference Article (Chapter 1)
	Congestion Aware (Chapter 4)
CAC	Common Admission Control (Chapter 1)
	Call Admission Control (Chapters 2, 3 and 4)

CBR	Constant Bit Rate
CC	Congestion Control
CCAC	Common Call Admission Control (a.k.a CAC, first meaning)
CCC	Common Congestion Control
CCPCH	Common Control Physical Channel
CD	Congestion Detection
CDMA	Code Division Multiple Access
CP	Congestion Probability
CPC	Cognitive Pilot Channel
CPICH	Common Pilot Channel
CPS	Common Packet Scheduling (a.k.a. CTS)
CR	Cognitive Radio
	Congestion Resolution (Chapter 4)
CRN	Cognitive Radio Networks
CRV	Congestion Recovery
CS	Coverage Scenario (used in Chapter 3 as CS1, CS2 and CS3.)
CSCC	Common Spectrum Coordination Channel
CRRM	Common Radio Resource Management
CTMC	Continuous Time Markov Chain
CTS	Common Traffic Scheduling
	Clear to Send (Chapter 3)
DCH	Dedicated Channel
DFS	Dynamic Frequency Selection
DL	Downlink
DSA	Dynamic Spectrum Access
DTMC	Discrete Time Markov Chain
E	Erlang
ECG	Erlang Capacity Gain
EDGE	Enhanced Data rates for GSM Evolution
ES	Spanish
ETSI	European Telecommunications Standards Institute
EU	European Union
FACH	Forward Access Channel
FCC	Federal Communications Commission
FCS	Fixed Channelization Scheme
FOMA	Freedom of Mobile Multimedia Access
FSA	Fixed Spectrum Access
GERAN	GSM/EDGE Radio Access Network
GoS	Grade of Service
GPRS	General Packet Radio Service
GSM	Global System for Mobile communications

HOC	Handover Control
HSDPA	High Speed Downlink Packet Access
HSPA	High Speed Packet Access
IEEE	Institute of Electrical and Electronics Engineers
IP	Internet Protocol
JA	Journal Article
JRRM	Joint RRM (a.k.a. CRRM)
LB	Load Balancing
LTE	Long-Term Evolution
MAC	Medium Access Control
MCL	Minimum Coupling Loss
MMTD	Multi-Mode Terminal Driven RAT selection policy
MRRM	Multiple RRM (a.k.a. CRRM)
NB	Narrow Band
NO	Network Operator
OFDMA	Orthogonal Frequency Division Multiple Access
OP	Operating Point
OSA	Opportunistic Spectrum Access
OVSF	Orthogonal Variable Spreading Factor
PC	Power Control
PCH	Paging Channel
PDC	Pacific Digital Cellular
PHY	Physical Layer
PN	Primary Network
PS	Packet Scheduling (a.k.a. TS)
PU	Primary User (i.e. licensed user)
QM	Queue Management
QoS	Quality of Service
RAN	Radio Access Network
RAT	Radio Access Technology
RC	Radio Resource Consumption
RNC	Radio Network Controller
RND	Random RAT selection policy
ROC	Receiver Operating Characteristic
RRA	Radio Resource Allocation
RRM	Radio Resource Management
RRU	Radio Resource Units
RTS	Request to Send
SCA	Supporting Conference Article
SB	Service Based (also used as SB#1, SB#2, etc. when referring to a particular service-based policy)

SBS	Secondary Base Station
SDR	Software Defined Radio
SF	Spreading Factor
SIFS	Short Inter-Frame Space
SM	Spectrum Management
SN	Secondary Network
SNR	Signal to Noise Ratio
SpHO	Spectrum Handover
SSBE	Steady-State Balance Equation
SU	Secondary User (i.e. unlicensed user)
	Service Unavailability (Chapter 3)
TBS	Time-Based Service
TDMA	Time Division Multiple Access
TD-SCDMA	Time Division-Synchronous Code Division Multiple Access
TS	Traffic Scheduling
	Terminal Scenario (Chapter 3, used as TS1, TS2 and TS3)
TSL	Time-Slot
TV	Television
UHF	Ultra High Frequency
UL	Uplink
UMTS	Universal Mobile Telecommunications System
UTRAN	UMTS Radio Access Network
UWB	Ultra-Wideband
VBR	Variable Bit Rate
VBS	Volume-Based Service
VHF	Very High Frequency
VHO	Vertical Handover
WB	Wide Band
WCDMA	Wideband CDMA
WiMAX	Worldwide Interoperability for Microwave Access
WLAN	Wireless Local Area Network
WMAN	Wireless Metropolitan Access Network
WWAN	Wireless Wide Area Network

List of Figures

List of Tables

Short Biography

Xavier Gelabert (Alaior, 1978) received the Telecommunications Engineering degree (equivalent to BSc plus MSc) from the Universitat Politècnica de Catalunya (UPC), Barcelona, in 2004. He also holds an MSc degree in electrical engineering, with major in wireless communications, from the Royal Institute of Technology (KTH), Stockholm, 2003. In 2004, he joined the Radio Communication Research Group in the Department of Signal Theory and Communications, UPC, where he is pursuing his PhD. From August to December 2008 he was a visiting researcher at the Broadband Wireless Networking Laboratory (BWN-Lab) at Georgia Institute of Technology. From January to July 2009, he was a visiting researcher at the Instituto de Telecomunicaciones y Aplicaciones Multimedia (iTEAM), Universidad Politécnica de Valencia (UPV). His current research interests are in the field of mobile radio communication systems, with a special emphasis on Common Radio Resource Management (CRRM) strategies in multi-access networks, quality of service provisioning, and opportunistic/cognitive spectrum management. He has been actively involved in European-funded projects EVEREST, AROMA, and E3 along with Spanish projects COSMOS and COGNOS. He is a member of the IEEE.