Design and evaluation of admission control policies in mobile cellular networks

ELENA BERNAL MOR

EDITORIAL
UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Universitat Politècnica de València

Departamento de Comunicaciones

UNIVERSITAT
POLITÈCNICA
DE VALÈNCIA

Doctoral Thesis

« Design and Evaluation of Admission Control Policies in Mobile Cellular Networks »

Author: Elena Bernal Mor

Directed by: Dr. Jorge Martínez Bauset
Dr. Vicent Pla

Valencia
February 2013

First Edition, 2013

© of the present edition:
Editorial Universitat Politècnica de València
www.editorial.upv.es

ISBN: -978-84-9048-049-6 (printed version)

Concern for man and his fate must always form the chief interest of all technical endeavors. Never forget this in the midst of your diagrams and equations.

Albert Einstein

Agradecimientos

A través de estas líneas me gustaría mostrar mi agradecimiento a aquellas personas que han hecho que esta tesis sea posible. Mi más sentido agradecimiento a mis directores de tesis, Jorge y Vicent, por su dirección, su ayuda y por todos los conocimientos y entusiasmo que me han transmitido. A Vicente, por su experiencia investigadora y por el inestimable trabajo que realiza dia a dia en el grupo. También, a todos los miembros del grupo GIRBA.

Me gustaría mostrar mi agradecimietno por su amistad y apoyo a los miembros del laboratorio con los que he compartido esta andadura dia a dia. Aquellos con los que empecé, María José, José Manuel y David G., por mostrarme el camino a seguir. Aquellos con los que finalizo, Diego, Salva, David T., por mantener la ilusión y el entusiasmo siempre. También, a los compañeros de otros laboratorios que han compartido toda esta etapa.

También quiero mostrar mi agradecimiento por el apoyo y la hospitalidad recibida a Prof. Chris Blondia durante la estancia en el grupo PATS. A Kathleen y Bart por su entusiasmo, dedicación y profesionalidad. A todos los miembros del grupo PATS por la acogida recibida. Así mismo, quiero agradecerle a Prof. Ian Akyildiz la hospitalidad y los consejos recibidos durante la estancia en BWN lab. A Josep Miquel. A todos los miembros de BWN lab por su amabilidad y su talento investigador.

Finalmente, me gustaría agradecer el apoyo, la paciencia y comprensión de mi familia. Ellos me han acompañado durante todo este tiempo dándome fuerzas y ánimo en cada paso. Gracias.

Acknowledgment

I would like to thank all the people who made possible this thesis. I would like to express my sincere gratitude to my directors, Jorge and Vicent, for their direction, their help and all the knowledge and eagerness transmitted. To Vicente, for his research experience and the incalculable work done every day. My gratitude to all the members of the group GIRBA.

Also, I want to express my gratitude to my labmates for their friendship and support. To Maria José, José Manuel and David G. for leading me through the first steps. To Diego, Salva and David T. for keeping the eagerness. My gratitude to the colleagues of other laboratories who have shared this stage.

I would like to express my gratitude for his support and hospitality to Prof. Chris Blondia during my time at PATS group. To Kathleen and Bart for their eagerness, kindness and professionalism. To all the members of PATS group for the warm welcome. As well, I want to express my gratitude to Prof. Ian Akyildiz for his hospitality and advice during my time in the BWN lab. To Josep Miquel. To all the members of BWN lab for their friendliness and research expertise.

Finally, I would like to thank my family for their support, patience and understanding. They have been with me during all this time, giving me strength and courage on every step. Thank you.

Abstract

In the last decade, mobile cellular networks have experienced a major growth and progress due to a change in the way today's society creates, shares and consumes information. The enormous impact and penetration of mobile phone services on the society, as well as the introduction of a large variety of multimedia and data services, has lead to a spectacular growth of the traffic volume carried by these type of networks. This trend will continue in the coming years as new applications are continuously appearing with higher QoS and bandwidth requirements. However, current mobile cellular networks have to face strong bandwidth limitations due to the scarcity of frequencies in the radio spectrum. Thus, these new requirements have established new challenges for the telecommunication industry. It is necessary to manage an increasing number of demanding services together with the scarcity of the spectrum in order to offer services that meet the user needs in an efficient and economical manner. In this context, the radio resource management arises as a key strategy to deal with those network requirements. Specifically, the admission control is a key mechanism to efficiently use the available radio resources, while providing the required QoS guarantees for all users.

This work aims at designing and evaluating admission control policies implemented in mobile cellular networks that support different bearer services. Moreover, this thesis is a contribution to the development of models that evaluate different admission control policies in the challenging context introduced by the forthcoming 4G networks. Thus, in this thesis, the design

and evaluation of admission control policies is tackled for current and forthcoming cellular networks. In the first part of the thesis, the development of admission control policies for current mobile cellular networks is studied, while in the second part of the thesis, novel admission control policies are proposed to overcome the challenges introduced by forthcoming mobile networks, such as Long Term Evolution networks or Cognitive Radio technologies.

Resum

Durant els darrers anys les xarxes mòbils cel·lulars han experimentat un considerable creixement degut al nou mode en que la societat crea, comparteix i consumeix informació. L'enorme impacte i penetració dels serveis mòbils telefònics en la societat, així com la introducció d'una amplia varietat de nous serveis de dades i serveis multimèdia, han portat a un creixement espectacular del volum de tràfic transportat per aquest tipus de xarxes. Aquesta tendència es mantindrà en els pròxims anys ja que constantment van apareixent noves aplicacions que demanden major qualitat de servei i ample de banda. Tot i així, les xarxes mòbils cel·lulars actuals tenen fortes limitacions en quant a amples de banda, degut a l'escassetat de freqüències en l'espectre radioelèctric. Amb tot, aquests nous requeriments estableixen nous reptes per a la industria de les telecomunicacions. És necessari gestionar un nombre creixent de serveis que demanden una elevada quantitat de recursos conjuntament amb l'escassetat de l'espectre radioelèctric per a oferir serveis que satisfacin les necessitats dels usuaris de un mode eficaç i econòmic. Dins d'aquest context, la gestió dels recursos ràdio es presenta com una estratègia clau per a fer front a les característiques pròpies d'aquestes xarxes. Concretament, el control d'admissió es un mecanisme clau per utilitzar eficientment els recursos radio disponibles, proporcionant al mateix temps les garanties de qualitat de servei requerides a tots els usuaris.

El present treball es centra en el disseny i avaluació de polítiques de control d'admissió implementades en xarxes mòbils cel·lulars multiservei que transporten diferents serveis portadors. D'altra banda, la present tesis també

és una contribució al desenvolupament de models amb els que avaluar diferents polítiques de control d'admissió dins del desafiant context introduït per les xarxes 4G de pròxima aparició. En la primera part d'aquesta tesis es tracta l'implementació de polítiques de control d'admissió per a xarxes utilitzades actualment, mentre que en la segona part d'aquesta tesis es proposen polítiques de control d'admissió innovadores amb l'objectiu de fer front als nous reptes introduïts per les xarxes de pròxima aparició, tal com les xarxes *Long Term Evolution* o les de tecnologia *Cognitive Radio*.

Resumen

Durante los últimos años las redes móviles celulares han experimentado un considerable crecimiento y desarrollo debido al nuevo modo en que la sociedad crea, comparte y consume información. El enorme impacto y penetración de los servicios móviles telefónicos en la sociedad actual, así como la introducción de un amplio abanico de nuevos servicios de datos y servicios multimedia, han llevado a un crecimiento espectacular del volumen de tráfico transportado por este tipo de redes. Esta tendencia se mantendrá en los próximos años ya que constantemente van apareciendo nuevas aplicaciones que demandan mayor calidad de servicio y ancho de banda. Sin embargo, las redes móviles celulares actuales tienen fuertes limitaciones de ancho de banda debido a la escasez de frecuencias en el espectro radioeléctrico. Así, estas nuevas necesidades establecen nuevos retos para la industria de las telecomunicaciones. Es necesario gestionar un creciente número de servicios que demandan elevadas cantidades de recursos, conjuntamente con la escasez del espectro radioeléctrico, para ofrecer servicios que satisfagan las necesidades de los usuarios de un modo eficaz y económico. Dentro de este contexto, la gestión de los recursos radio se presenta como una estrategia clave para hacer frente a las características especiales de estas redes. Concretamente, el control de admisión es un mecanismo clave para utilizar eficientemente los recursos radio disponibles, proporcionando al mismo tiempo las garantías de calidad de servicio requeridas para todos los usuarios.

El presente trabajo se centra en el diseño y evaluación de políticas de control de admisión implementadas en redes móviles celulares multiservicio que

transportan diferentes servicios portadores. Además, la presente tesis es una contribución al desarrollo de modelos con los que evaluar diferentes políticas de control de admisión en el desafiante contexto introducido por las redes 4G de próxima aparición. En la primera parte de esta tesis se trata el desarrollo de políticas de control de admisión para redes utilizadas actualmente, mientras que en la segunda parte de esta tesis se proponen políticas de control de admisión novedosas con el objetivo de hacer frente a los retos introducidos por las redes de próxima aparición, tales como las redes *Long Term Evolution* o las de tecnología *Cognitive Radio*.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

During the last years, mobile cellular networks have witnessed an enormous growth in the carried traffic volume mainly due to two reasons. First, the high penetration of mobile phone services in the society, and second, 3G networks and forthcoming 4G networks have established a new paradigm with a variety of services which have different Quality of Service (QoS) requirements and traffic characteristics. This trend will continue in the coming years as mobile systems are expected to support a larger variety of multimedia services that are carried over different bearer services. Unlike wired networks, current mobile wireless networks have to face strong bandwidth limitations due to the scarcity of frequencies in the radio spectrum. Therefore, the main challenge for the telecommunication industry is to offer services that meet the user needs in an efficient and economical manner.

In order to increase the capacity of the mobile networks, the geographical reuse of frequencies by means of the use of the cellular architecture has been established [HR97]. This architecture has become very efficient to this date, but it is not enough to face the development paradigm that is expected in this type of networks due to the tremendous growth of mobile telecommunications services. Moreover, this architecture entails some problems such as the high economical cost associated to its development or the additional

complexity introduced by the terminal mobility. Users move from one cell to another and the continuity of active sessions must be guaranteed.

Taking into account the limited bandwidth together with the user's mobility and the requirements of the new services, different mechanisms will be needed to provide a certain degree of QoS in mobile cellular networks. The performance parameters that define the QoS can be considered at two levels: *packet-level* and *session-level*. The *packet-level* QoS corresponds to the correct transmission of packets of data. Some parameters used to describe the packet-level QoS are for example, the maximum delay, the variability of the delay or *jitter*, loss, the throughput, etc. In order to deal with these QoS requirements, efficient medium access protocols and packet scheduling schemes as well as bandwidth reservation must be in place. The *session-level* QoS corresponds to connection establishment and management of the connectivity and continuity of services in a mobile cellular network. Although the QoS is a global concept that must be provided at every level, this work is focused on the session-level.

In order to provide an acceptable level of QoS to the subscribers, one of the main issues that must be considered is the Radio Resource Management (RRM) which is the set of mechanisms needed to have an efficient use of the available radio resources [PRSADG05]. An important mechanism for RRM is the Admission Control (AC). In this context, the AC is a key mechanism in the design and operation of multiservice mobile cellular networks to maximize the network usage while providing the QoS guarantees to all ongoing sessions in the system. The admission controller decides if a session is admitted to the cell or not. It bases its decisions on the availability of resources considering the resources needed to guarantee the QoS of the new session and the already accepted sessions. These decisions are more complex due to the limited bandwidth and the terminal mobility. This work is fundamentally based on the development and evaluation of models to study the performance of AC policies, and on the design of such AC policies.

Modeling and evaluation of RRM in mobile cellular networks have been

widely studied for 2G and 3G technologies, but technologies like *Long Term Evolution* (LTE) developed by the *3rd Generation Partnership Project* (3GPP) or the future 4G of mobile cellular networks introduce new challenges in the design of AC policies. These technologies increase the capacity and bit rate of mobile cellular networks by employing Orthogonal Frequency-Division Multiple Access (OFDMA) and using Adaptive Modulation and Coding (AMC) [3GP10a]. This means that different Modulation and Coding Schemes (MCSs) can be used at different points in time depending on the signal quality. One of the most dominant factors in the reduction of the transmitted signal power is the pathloss, which is linked to the distance between the transmitter and the receiver. As a consequence, data can be sent to users at different bit rates, which are determined by the MCSs used, as they move around in a cell. Therefore, in order to maintain a service with a fixed bit rate the number of resource units that a user needs is time-varying, while in 2G models the number of resources needed is considered a fixed amount. It is for example possible that at one point in time there is sufficient capacity to provide the desired bit rate to all users while at another point in time, without having accepted new users in the system, the cell capacity can drop below the required amount due to the varying number of resources needed. The varying of this cell capacity might influence the decision taken by the AC policy. Thus, it is important when designing AC policies in 4G networks to take this fact into account.

Moreover, due to the new paradigm established by the forthcoming 4G networks, reducing the cell size has become a hot topic introducing the novel concept of femtocells [CAG08] [CHS08]. The main trend in these new technologies to increase the system capacity is to reduce the cell size without deploying more infrastructure, which is achieved by using femtocells. These are data access points designed for indoor usage to improve the indoor data and voice coverage and reduce the traffic managed by the macrocell network. This new concept has also introduced some challenges related to the RRM and the AC mechanisms needed to guarantee the QoS requeriments, which makes it necessary to adapt the current AC algorithms and develop new ones.

Despite the fact that many efforts have been made on the development of AC policies and the characterization of mobility patterns of terminals in multiservice mobile cellular networks, the gain in efficiency is still insufficient. Hence, the *Cognitive Radio* (CR) technology [DAR03] has been proposed.

Today's cellular mobile networks are characterized by fixed spectrum assignment policies. A study by the Federal Communication Commission (FCC) Spectrum Task Force [Com02] showed high temporal and geographic variations in the spectrum utilization, these variations range from 15% to 85% in the bands below 3 GHz. Most of the spectrum that could be reasonably utilized for communications is licensed and these licenses are allocated for very long periods of time. As a result, the spectrum allocation itself is nowadays rigid and the spectrum is underutilized which makes necessary a new communication paradigm to exploit the existing wireless spectrum opportunistically [AAFS04]. Current research efforts in this field are devoted to the study of technologies that enable dynamic spectrum access. The CR technology has been proposed as a concept that provides the ability to detect idle frequencies that are not occupied by licensed users and enables non-licensed users to use these idle bands in an opportunistic manner [MGM99, ALVM08, PPPMB09, MBPPP12]. Additional functionalities are required for licensed and non licensed users to share the licensed spectrum band and therefore, an efficient channel sharing strategy and an AC policy should be in place for secondary users when QoS guarantees for them are required.

1.1 Thesis objectives

This work aims at designing and evaluating AC policies implemented in multiservice mobile cellular networks that support both real-time and non-real-time traffic. It is also a contribution to the development of models that evaluate different AC policies in the challenging context introduced by forthcoming 4G networks, such as the OFDMA based networks, the femtocell

concept or the CR technology. To that end, the lines of work which have been developed in this thesis are the following:

- Characterization of the flow duration and handover probability of non-real-time traffic in mobile cellular networks.
- Design and evaluation of different AC policies, considering the computational cost of their configuration, the robustness against overloads of traffic, and the insensitivity to distributions which model the system.
- Study of analytical models in order to evaluate AC policies in forthcoming networks such as OFDM based networks and femtocell deployments.
- Design of optimal AC policies for non-licensed users in CR networks.

Similarly, simulation models have been developed in order to validate the modeling assumptions made in the analytical models.

1.2 Thesis structure

This thesis is structured in 10 chapters, of which the first two chapters are introductory, the last chapter concludes the thesis and the rest of chapters are structured in two parts. The first part includes four chapters related to multiservice mobile cellular networks and the second part includes three chapters that aim to study the AC in forthcoming mobile networks.

Specifically, in Chapter 2 we introduce a model of a cellular mobile network from the traffic perspective, that will be used along the thesis. The most common assumptions done when modeling this type of networks are detailed and the state of the art related to AC techniques and methodologies is described.

The first part of the thesis includes four chapters that study current problems in multiservice cellular networks. In Chapter 3 the non-real-time traffic

is studied in order to find a model to characterize the flow duration and handover probability of elastic users. In Chapter 4 the computational cost of different algorithms used to design the parameters of trunk reservation AC policies is evaluated. Later, in Chapter 5, the robustness and efficiency of different AC policies is studied and compared in different multiservice mobile cellular scenarios. Next, in Chapter 6, we study the reversibility of the continuous time Markov chain which models the system generated by different AC policies, and the insensitivity of its stationary distribution to the channel holding time distribution.

The second part of the thesis analyzes the AC challenges introduced by forthcoming mobile networks, such as the 4G networks. In Chapter 7, AC in OFDMA based mobile cellular networks is studied. A model is presented and validated, and then it is used to evaluate some proposed AC policies. Next, in Chapter 8, the femtocell concept is considered. A model to study the activity profile in the femtocells is presented and an AC policy for macrocell users is evaluated. In Chapter 9, a cognitive radio scenario is evauated where different strategies to rent the resources from the licensed network are applied. In addition, an optimal AC policy for a given cost function is proposed.

Finally, Chapter 10 summarizes the conducted work in this thesis. The main contributions and conclusions are highlighted and the possible future lines of work are suggested .

Chapter 2

Cellular model

In this chapter, we introduce a model of a cellular mobile network from the traffic perspective, that will be used along the thesis. The assumptions done in this thesis when modeling the cellular network are detailed and the state of the art related to AC techniques and methodologies is described.

In the mobile cellular systems, the service area is divided in smaller areas named cells. Each of them is assigned a certain number of frequencies and is covered by a Base Station (BS), which enables the communication of the users who are in the cell with the rest of the network. The main idea under the cellular concept is the reuse of frequencies, i.e., the same frequencies allocated to a cell can be reused in other different cells.

The users roam around the service area moving from one cell to other. The signal transmitted between the Mobile Terminal (MT) and the BS becomes weak when the MT is far from the BS due to the signal attenuation. At a given point the signal power can be too low to assure a reliable communication and it is necessary to connect the MT to a closer BS which provides better service. The mechanism used to manage the continuity of the service in progress when a MT moves from one cell to another is named *handover* and must be transparent to the user. To ensure the continuity of communications when a mobile terminal moves from one cell to another, i.e., a handover

occurs, adjacent cells overlap with each other. The mobility of users adds more complexity to the operation of the cellular network since the QoS of all new and ongoing sessions must be provided considering that during the session lifetime a MT can change its location. This means that the AC policy must guarantee that the required resources are available in the destination cell where the MT moves to. The most common approach used by the AC policy to guarantee the continuity of the service is giving priority to the handover requests by using the resource reservation. The basic idea is to reserve a certain number of resources for requests with higher priority and admit requests with lower priority only if the amount of free resources exceeds the number of reserved resources.

When mobile cellular systems are analyzed, a common approach is to consider homogeneous traffic, i.e., the offered traffic by new sessions is equal in all cells and the arrival and departure rate of handovers are also equal in all cells. Then, in a homogeneous cellular system, the system performance can be evaluated considering only one isolated cell [OR01]. When the scenario under consideration is not homogeneous and presents hot-spots, for example scenarios with big supermarkets or densely populated streets, a multicell approach is more appropriate [BBP01] since the offered traffic in each cell can have high variations from one cell to another. In this thesis, we consider homogeneous scenarios and therefore, we evaluate the system performance considering one isolated cell.

2.1 Teletraffic random variables

In order to model appropriately a mobile cellular network, random variables are generally used to describe the most important magnitudes that are associated with the cellular environment. These magnitudes are: the inter-arrival time for new and handover request, the session duration, the cell residence time and the channel holding time. For the sake of the tractability of the model, particularly if it is an analytical model, the general trend is to assume

that all random variables are exponentially distributed.

1. *Inter-arrival time*.

   For new requests, if the inter-arrival times are independent and identically distributed (iid) according to an exponential distribution, the arrival process is of Poisson type. Considering a Poisson arrival process for new request is the most accepted assumption in the literature. Nevertheless, the infinite population, which is inherent to the Poisson arrival process, has been questioned in [BB97]. However, this study is based on field measurements in a Public Access Mobile Radio (PAMR) network. A PAMR is a wireless network which does not have the same characteristics than a mobile cellular network since the PAMR networks provide services for a limited number of users.

   For handover request, the studies to validate the exponential assumption are not unanimous. Chlebus and Ludwin [CL95] show that the arrival process for handover traffic is Poisson when the process for new arrival requests is Poisson and the blocking phenomena is not considered. They conclude that the Poisson assumption for handover traffic is a reasonable approximation. Sidi and Starobinski [SS97] also conclude that the Poisson assumption is reasonable for homogeneous traffic between a large number of cells. Rajaratnam and Takawira [RT01] show that the Poisson assumption for handover traffic may not be appropriate when the terminal mobility is high. Orlik and Rappaport [OR01] showed that small differences are found between Poisson handover traffic and non-Poisson, specially for heavy loads. In the light of these studies and taking into account that we consider homogeneous traffic in our model and non-extreme user mobility patterns, we consider a Poisson arrival process for handover requests.

2. *Cell residence time*, *channel holding time* and *session duration*.

   The Cell Residence Time (CRT) is the time that a MT spends inside a cell. The Channel Holding Time (CHT) is the time a MT occupies the

resources of a given cell. Clearly, the CHT is given by he minimum value of the CRT and the session duration. The CRT depends on the cell geometry and the user mobility pattern, while the session duration depends on the specific applications. The statistical characterization of CRT and CHT have been widely studied in the literature.

For CHT, in [ZD97] the authors show that if the session duration is considered exponentially distributed, then the CHT can be approximated by the exponential distribution. Many distributions have been studied in different works for the CHT, such as the lognormal [JL96], the deterministic negative exponential and the Gamma [RT01], a linear combination of lognormal distributions [BJ00] or a phase-type distribution [CNI04], but there is no unanimity about the validity of the exponential assumption. In [XT03], the exponential assumption is presented as a good approximation, while the results in [HSSK02] show some situations with significant divergences.

For CRT, a series of empirical studies in [HSSK01, HSSK02] and references therein, show that CRT and CHT have self-similar properties. In [ZD97] the authors conclude that the CRT can be described by a gamma distribution. More recent studies [ZBA09] indicates that the lognormal distribution is a reasonable assumption. In [Mac05], the authors show that if the new and handover sessions are treated equally, the system performance is insensitive to the distribution of the CRT and session duration.

2.2 Streaming and elastic traffic

Mobile systems support a large variety of services which demand different QoS requirements. Depending on the QoS demanded, applications have to be carried by bearer service classes with different characteristics. The traffic generated by the applications can be mainly classified in two different groups, namely, elastic traffic and streaming traffic [BR03].

1. *Streaming traffic* requires a minimum transfer rate in order to work properly as well as some time related requirements such as bounded delay and *jitter*. It corresponds to real-time services such as voice or streaming video.

2. *Elastic traffic* has loose time requirements and its transfer rate can be adapted to the available resources. It is generated by applications that support the transfer of data traffic, such as web navigation, or file transfers, i.e., non-real time services. The data traffic is bursty, i.e., sometimes the data transmission rate is idle, while at other times it might be very high, appearing sudden traffic peaks. Thus, an elastic session can generate several traffic flows, i.e, several sequences of data packets.

   Elastic flows are generally transported over TCP which takes care of rate adaptation and bandwidth sharing among the different flows. If the total traffic demand of elastic flows exceeds the available capacity some flows might be aborted due to impatience. Flow impatience can arise from human impatience or because TCP of higher-layer protocols interpret that the connection is broken. Abandonments due to impatience are useful to cope with overload and serve to stabilize the system but, this phenomenon will have a negative impact on the efficiency because capacity is wasted by non-completed flows [BR03]. Hence, an AC policy should be enforced for elastic traffic [BR03].

In the light of the above arguments, it seems logical to give priority to streaming traffic and leave elastic traffic use the remaining resources, reserving a small quantity of resources for elastic traffic to prevent starvation in case of overload of streaming traffic.

2.3 Admission control

Although different mechanisms can be implemented to manage the radio resources, such as *queuing* [HR86] or reducing the transmission rate when it

is necessary [LNH96], the most common approach is to use AC policies to limit the access to a certain number of resources depending on the priority of the arrival session. Such schemes where introduced in the mid 80's [PG85] and since then a great deal of variants, generalizations and improvements have been proposed. The aim of any AC algorithm is to achieve the best possible QoS using the available resources.

A classical scheme for resource sharing is the *Complete Sharing* (CS) policy, which admits a request provided there are enough free resources units available in the system. This leads to share the resources indiscriminately, i.e., it is equivalent to implement no AC. Another classical scheme is the *Complete Partitioning* (CP) policy, which statically divides the resource units among the service classes allowing each class the use of its allocated resources, i.e, the resources are not shared as each service class uses its portion of resources.

AC policies can be classified according to three different points of view: the type of information used to take admission decisions, the method employed to adjust the parameters, and the general structure of the model.

Different AC strategies have been proposed which differ in the available information that is needed to decide on the acceptance or rejection of user's requests. The most common approach bases the admission decision on the local state of the cell at which the AC operates. In general such information consists on the number of ongoing sessions in the cell or the quantity of resources being used, either aggregated or detailed by the service class. Another approach bases the admission decision on some kind of information obtained from the state of neighboring cells, mobility of terminals, history-based patterns, etc. With this information the future handovers can be predicted. Several works that study AC based on movement prediction can be found in the literature [HF01, YL02, SK04, MBGGP08, GG09, MBPPP12].

When an AC is designed, two methods can be followed in order to adjust the parameters. The first and most common approach is heuristic, i.e., a new AC policy is proposed and then it is evaluated by comparing its performance to the performance of previously existing AC policies [DS02, PCG05].

The second approach consists in formulating the admission problem as an optimization problem. The optimization problem has to be defined accurately in order to find the optimal AC policy among the possible AC policies, so that the system performance goal is maximized. The optimization problem is often formulated using the framework of *Markov Decision Processes* (MPD) or *Semi-Markov Decision Processes* (SMDP) [Ros70, PCG03]. Alternatively, other optimization methods can be employed, such as *linear programming* [PCG03], *genetic algorithms* [YR97], *hill-climbing* [GRMBP05] or *reinforcement learning* [PGGMCG04, GGMBP07, GG09].

Finally, the AC policies can be classified in three different families with respect to their general structure [GMF08]:

1. *Product-form* policies.

   The decision to accept a session depends only on the number of resources occupied by the ongoing sessions of the same class [GRMBP04]. The AC policies from this family produce a Markov process whose stationary state distribution can be computed as the product of the marginal distributions of each class, subject to a normalization constant. The product-form policies show a lower computational complexity at the cost of a reduced capacity and generally are insensitive to the CHT distribution [GRMBP05, MBPBM11]. Some product-form policies have been proposed in the literature. For example, the policy *Integer Limit* (IL) which limits the number of resources that a flow can occupy at one time [Ive87]. In [CLW95] the *Upper Limit and Guaranteed Minimum* (ULGM) policy is described, where each flow has a dedicated number of resources and competes for a common portion of resources. The policy *Fractional Limit* (FL) is presented in [LA95], which is similar to IL policy but the number of resources allocated is a fractional number.

2. *Trunk reservation* policies.

   The decision to accept a session depends on the number of free resources units in the system [RTN97]. These policies outperform *product-*

form policies in terms of system capacity, defined as the maximum aggregated offered traffic that the system can handle while meeting certain QoS requirements [GRMBP05]. However, *trunk reservation* policies do not produce a product-form stationary distribution and higher precision is required when the optimal configuration is determined with respect to *product-form* policies. Therefore, the computational complexity required to numerical evaluations is also higher.

Trunk reservation policies constitute the most common approach to give priority to certain flows. For a single service scenario, two *trunk reservation* policies named *Guard Channel* (GC) and *Fractional Guard Channel* (FGC) are optimum for common QoS objective functions [RTN97]. The GC policy was introduced in [HR86], where a static reserve of a certain number of resources was made for handover requests. In the FGC policy [RTN97], a fractional reservation is set by means of probabilistic admission decisions, thus allowing a finer adjustment of the policy. A generalization of GC to the multiservice scenario, *Multiple Guard Channel* (MGC) policy, was proposed in [CC97] and [LLC98], and the FGC was extended to the multiservice scenario, *Multiple Fractional Guard Channel* (MFGC), in [HUCPOG03a]. The main difference between single service and multiservice schemes is that while in the former there is only one admission threshold, in the later several admission thresholds are set, one for each service class. Other variations of trunk reservation policies have been proposed afterwards, such as the adaptive trunk reservation policies, which implement adaptive reservation of resources depending on the network load [GRDBMBP05, DGRP05, MBGRDB+09].

3. *General Stationary* policies.

   For single service scenarios the optimum AC policy belongs to the family of *trunk reservation* policy [RTN97, Bar01]. However, these optimality results cannot be extended to their counterparts in multiservice scenarios [BC02, PCG03], where the optimum AC policy belongs to *Stationary* policies or to *Ramdomized Stationary* (RS) policies [Ros70, GMF08]. For

Stationary policies, the admission decision depends on the current state of the system, which is expressed as the number of sessions of each service class in progress. In the RS policies, the decision also depends on a random component. The solutions provided by RS policies in general outperform the previous families in terms of system capacity and are considered as upper bounds in comparative studies [PCG03]. Note that, both *product-form* and *trunk reservation* policies are subclasses of the family of RS policies, where the decision depends, respectively, on the number of resources occupied by the sessions of the same service class, and the total quantity of resources occupied in the system, but not on the quantity of resources occupied by each service class.

Part I

Multiservice mobile cellular networks

Chapter 3

Elastic traffic characterization

3.1 Introduction

The teletraffic analysis of mobile networks, as well as the design of precise mobility models, play an important role in the network dimensioning and resource planning [FC02]. In this context, a good knowledge of the mobility characteristics of terminals is important for research and system design issues, for instance, knowing the probability of handover or the number of handovers that a session will execute is useful for resource dimensioning [Zha10]. Accurate models are required in order to appropriately characterize the mobility of terminals.

For streaming traffic, specifically voice traffic, the CHT and the CRT distributions have already been widely studied [BJ00, HSSK01, CNI04, Mac05, ZBA09] (see Section 2.1). In [GLZ07], the session duration distribution is studied in a GSM system. The handover probability for streaming traffic has been also studied in [FC02] where the session duration and the CRT follow a general distribution with a rational Laplace transform. The results for CRT considering streaming traffic can be extended to the case of elastic traffic, but the results for session durations in the aforementioned works are

not applicable to data applications. Modeling the elastic flow duration and handover related metrics for elastic traffic is qualitatively different and more complex than modeling their streaming traffic counterparts, as the duration of an elastic flow is heavily dependent on the network load. Despite the enormous surge in volume and economical relevance of mobile data traffic caused by the introduction of smartphones [CJP+11], to the best of our knowledge, such type of studies have not been carried out for elastic flows so far.

In this chapter, we aim at obtaining the distribution of both the flow duration and the number of handovers an elastic user has to go through. To make the model more realistic and of broader application, we consider impatient elastic users and a generic AC policy. As the flow duration of elastic flows in mobile networks depends on the network load, its complexity comes from the fact that, unlike streaming sessions, it depends on the load of the cells visited along a flow life time. Thus, it is reasonable to consider that the flow duration is composed of a number of phases with different rates given by the network load. We assume exponentially distributed sojourn times in each phase for elastic flows and exponentially distributed CRT and CHT for streaming traffic. That can be modeled by a phase-type (PH) distribution [Neu81]. Thereby, we start by constructing a *Continuous-Time Markov Chain* (CMTC) from which we derive the PH distribution for the duration of the flow and, from this, the distribution of the number of handovers. We provide exact results under the assumption that both flow sizes (in bits) and the CRT are exponentially distributed. Then, the model is extended by introducing an approximate technique to deal with more realistic distributions of the CRT and the flow size. We apply the obtained model to determine the handover probability under different CRT distributions and validate the results by simulation. This work resulted into the publication in [BMPMB12].

This chapter is structured as follows. In Section 3.2, we describe the scenario and the CTMC which models the system under study. Then, we derive the flow duration distribution of elastic flows and the handover related metrics. In Section 3.3, we compare the results obtained with the analytical model with simulation results. Finally, Section 3.4 concludes the chapter.

3.2 Analytical model

3.2.1 System model

We consider the homogeneous case where all cells are statistically identical and independent. Consequently the global performance of the system can be analyzed focusing on a single cell. Each cell has a total of C resource units, each of them has a capacity of R bits per second. The system offers N_s different Service Classes (SCs) that carry streaming traffic and N_e SCs that carry elastic traffic. Thus, the total number of SCs is $N = N_s + N_e$, where by $i = 1, \ldots, N_s$ we refer to streaming SCs and by $i = N_s + 1, \ldots, N$ to elastic SCs. Elastic flows use the capacity not occupied by streaming traffic. To avoid starvation, the system reserves 1 resource unit for elastic traffic. For each SC, new and handover arrivals are distinguished, so that there are N SCs and $2N$ arrival types.

For the sake of mathematical tractability we make the common assumptions of Poisson arrival processes for all SCs and exponentially distributed session durations and CRTs for streaming SCs. Let λ_i^n (λ_i^h) be the arrival rate for new (handover) streaming sessions or elastic flows of the ith SC. Let $\mu_i^{d,s}$ and $\mu_i^{r,s}$ be the rates of the session duration and the CRT of the ith streaming SC, $1 \leq i \leq N_s$. Hence, the CHT in a cell for a streaming SC is exponentially distributed with rate $\mu_i^s = \mu_i^{d,s} + \mu_i^{r,s}$. For streaming SCs, a request of the ith SC consumes b_i resources, $b_i \in \mathbb{N}$. The system state is described by the N-tuple $\boldsymbol{x} = (x_1, \ldots, x_N)$, where x_i is the number of ongoing streaming sessions or elastic flows of the ith SC, regardless of whether they were initiated as new or handover arrivals. The number of resources occupied at state $\boldsymbol{x}$ by streaming traffic is:

$$b(\boldsymbol{x}) = \sum_{i=1}^{N_s} x_i b_i.$$

We model elastic traffic at the flow level and ignore interactions at the packet level (scheduling, buffer management, etc.). The flow content is then viewed as a fluid that is transmitted as a continuous stream with rate changes

occurring only at streaming sessions or elastic flows arrivals and departures. An elastic flow of the ith SC is assumed to be rate-limited either by terminal capabilities or because it is bottlenecked at the radio link. We denote its maximum bandwidth by r_i^M. Also, elastic flows require a minimum bandwidth denoted by r_i^m. This can be guaranteed by an appropriate AC policy. We assume that flow sizes are exponentially distributed with mean L (bits). Let $\mu_i^M = r_i^M/L$ and $\mu_i^m = r_i^m/L$ (in flow/s) be the maximum and minimum service rates of an ith SC flow. Without lost of generality, we consider in this work that all the elastic SCs have the same minimum service rate μ^m and hence, the maximum number of flows in the system is

$$n_M = \left\lfloor \frac{C \cdot R}{r^m} \right\rfloor. \tag{3.1}$$

For elastic flows of the ith SC, the CRT is here assumed to be exponentially distributed with rate $\mu_i^{r,e}$. However, we extend the model for other distributions in Section 3.2.2 and validate this assumption in Section 3.3.

Without loss of generality, we consider that flows are ordered in increasing value of their rate limits r_i^M. The total number of elastic flows in the system in state $\boldsymbol{x}$ is denoted by

$$c(\boldsymbol{x}) = \sum_{i=N_s+1}^{N} x_i.$$

Let us define $\mu_i^{d,e}(\boldsymbol{x})$ as the service rate, i.e. the rate of the flow duration, of the ith SC flow at state $\boldsymbol{x}$ and

$$V(\boldsymbol{x}) = (C - b(\boldsymbol{x}))\frac{R}{L}$$

as the available aggregated service rate at state $\boldsymbol{x}$. Flows share the available resources fairly according to the following rule.

For $i = N_s + 1$,

$$\mu_i^{d,e}(\boldsymbol{x}) = \min\left(\mu_i^M, \frac{V(\boldsymbol{x})}{c(\boldsymbol{x})}\right), \tag{3.2}$$

while for $i > N_s + 1$,

$$\mu_i^{d,e}(\boldsymbol{x}) = \min\left(\mu_i^M, \frac{V(\boldsymbol{x}) - \sum_{j=N_s+1}^{i-1} x_j \mu_j^{d,e}(\boldsymbol{x})}{\sum_{j=i}^{N} x_j}\right). \tag{3.3}$$

Clearly, in (3.3) the bandwidth that a SC cannot use because it reached its rate limit is used by other SCs with higher rate limit.

Flows become impatient and might leave the system due to a very low throughput. The patience time at state $\boldsymbol{x}$ is modeled by an exponential distribution with rate:

$$\beta_i(\boldsymbol{x}) = \beta_i^1 \left(\frac{\mu_i^M}{\mu_i^{d,e}(\boldsymbol{x})} - 1\right) + \beta_i^0, \tag{3.4}$$

where β_i^1 is a scaling factor that relates the throughput degradation and the patience rate and β_i^0 is a factor that determines the patience rate when a flow is served at its maximum rate. When the service rate of the ith SC is less than its maximum, its patience rate increases when the service rate decreases.

We consider a non-preemptive AC policy. A streaming session or elastic flow is accepted if there are enough free resources to support it and if, after acceptance, all ongoing elastic flows obtain a service rate equal or bigger than their minimum μ^m. We denote by $a_i^n(\boldsymbol{x})$ ($a_i^h(\boldsymbol{x})$) the probability of accepting a new (handover) arrival of the ith SC in state $\boldsymbol{x}$. Clearly, the system can be modeled as a CTMC, specifically as a multidimensional birth and death process with state space,

$$\mathcal{W} := \left\{\boldsymbol{x} : x_i \in \mathbb{N}; \sum_{i=1}^{N_s} x_i b_i \leq C - 1; \mu_i^{d,e}(\boldsymbol{x}) \geq \mu_i^m\right\}. \tag{3.5}$$

As an example, Fig. 3.1 shows the transition diagram of the CTMC that models a system where $N_s = 1$, $N_e = 1$. Note that transitions from states with $j = 0$ or $k = 0$ to states $j - 1$ or $k - 1$, respectively, are not possible. For clarity, the notation has been simplified writing a_i^n, a_i^h β_2 and $\mu_2^{d,e}$ instead of $a_i^n(\boldsymbol{x})$, $a_i^h(\boldsymbol{x})$, $\beta_2(\boldsymbol{x})$ and $\mu(\boldsymbol{x})_2^{d,e}$, respectively.

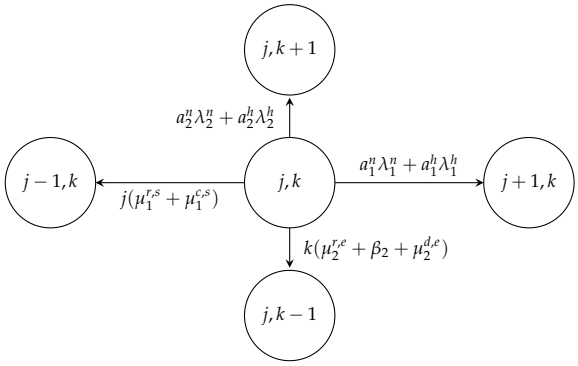

Figure 3.1: Transition diagram of the CTMC with SC=2. The state (j,k) obeys $0 \leq j \leq \lfloor (C-1)/b_1 \rfloor$ and $0 \leq k \leq n_M$.

If the state $\boldsymbol{x}'$ represents the state achieved by the system after a state transition and $q_{xx'}$ is the transition rate from $\boldsymbol{x}$ to $\boldsymbol{x}'$. The stationary distribution $\pi(\boldsymbol{x})$ of this CTMC can be obtained by solving the global balance equations and the normalization equation,

$$\pi(\boldsymbol{x}) \sum_{\forall \boldsymbol{x}' \in \mathcal{W}} q_{xx'} = \sum_{\forall \boldsymbol{x}' \in \mathcal{W}} q_{x'x} \pi(\boldsymbol{x}') \qquad \forall \boldsymbol{x} \in \mathcal{W}, \tag{3.6}$$

$$\sum_{\forall \boldsymbol{x} \in \mathcal{W}} \pi(\boldsymbol{x}) = 1. \tag{3.7}$$

From now on we will refer to this CTMC as the original CTMC.

3.2.2 Duration of elastic flows and handover metrics

In the system under study, the duration of an elastic flow is composed of a number of exponentially distributed phases with different rates. Thus, the elastic flow duration follows a PH distribution, (see Appendix C.1.2). A PH distribution defines the time until absorption in an Absorbing Markov Process (AMP) [Neu81], where an *absorbing* state is a state which is impossible to leave. It is commonly represented by a pair $(\boldsymbol{\alpha}, \boldsymbol{S})$, where matrix $\boldsymbol{S}$ de-

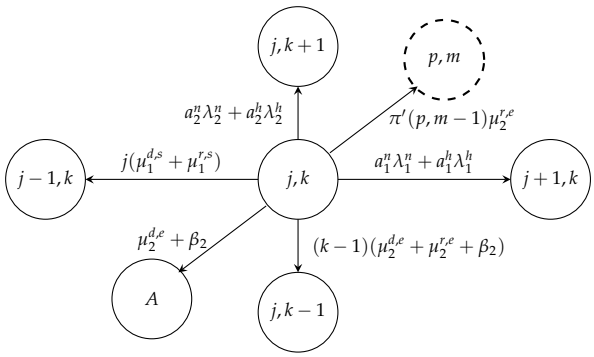

Figure 3.2: Transition diagram of the AMP for SC=2. State A is the absorbing state, $0 \leq j,p \leq \lfloor (C-1)/b_1 \rfloor$, $1 \leq k,m \leq n_M$ and $(p,m) \neq (j,k)$.

fines transition rates among the transient states, and vector $\boldsymbol{\alpha}$ the probabilities that the process is started in any of the states. The transition rates from the transient states to the absorbing state is defined by vector $\boldsymbol{\tau}$, which satisfies $\boldsymbol{\tau} = -\boldsymbol{S}\mathbf{1}$, where $\mathbf{1}$ is a column vector of 1s.

A different AMP is defined for each elastic SC i, PH$(\boldsymbol{\alpha}_i, \boldsymbol{S}_i)$. It contains the states that might be visited by a tagged elastic flow of the ith SC until it abandons or terminates successfully. The AMP must consider that flows might be handed over multiple times to adjacent cells. We assume that the tagged flow has a normal progression and therefore, it is never blocked. The state space $\mathcal{W}'$ of the AMP of the ith SC is defined by:

$$\mathcal{W}'_i := \{\boldsymbol{x} \in \mathcal{W}, x_i > 0\}. \tag{3.8}$$

The initiation vector probabilities $\boldsymbol{\alpha}_i$ is derived from the stationary probabilities $\pi(\boldsymbol{x})$ of the original CTMC considering that the state space is restricted to $\mathcal{W}'_i$. As an example, Fig. 3.2 shows the AMP associated to the elastic flows of a system with $N_s = 1, N_e = 1$. State A represents the absorbing state and it is visited when the tagged flow abandons or ends. The dashed circle (p,m) represents a set of states composed of all the feasible states that can be reached immediately after the tagged flow is handed over to an adjacent

cell, i.e, (p, m) represents any state where $(p, m) \neq (j, k)$ and $m \neq 0$. The probability $\pi'(p, m-1)$ is used to take into account the state that the target cell is in immediately before the handover and it is obtained from the stationary distribution of the original CTMC $\{\pi(\boldsymbol{x})\}$ after removing the blocking states and then re-normalizing. Finally, note also that the transitions to states $(j, k-1)$, where $k-1=0$ are not possible.

When a flow is initiated as a handover request, it spends all the CRT as an active flow. Let T_d and T_r be random variables that denote the duration and CRT of an elastic flow, respectively. At this point, no assumptions are made regarding the distributions of these random variables. When a flow enters a cell as a handover request, it is handed over again if T_d is longer than T_r, and therefore the probability P^h that it is handed over again is determined by

$$\begin{aligned} P^h &= 1 - P(T_d < T_r) = \\ &= 1 - \int_0^\infty P(t_d < T_r) f_d(t_d) dt_d = \\ &= 1 - \int_0^\infty (1 - F_r(t_d)) f_d(t_d) dt_d \ , \end{aligned} \tag{3.9}$$

where F_r is the distribution function of T_r, and f_d stands for the probability density function of T_d.

When a flow is initiated as a new request, it does not spend all the CRT as an active flow. Let $\hat{T}_r$ be the residual life of the CRT of a flow [HLL04], i.e., the time elapsed since a flow is initiated until the terminal leaves a cell, (see Fig 3.3). Thus, when a flow is initiated as a new request, it is handed over when T_d is longer than $\hat{T}_r$. Then, the handover probability of sessions which enter the cell as new requests, $\hat{P}^h$, has the same expression as in (3.9), but substituting F_r by the distribution function of the residual CRT, $\hat{F}_r$. From the residual life theorem [Ros85], the probability density function of the residual life of a CRT, $\hat{f}_r(t)$, is given by:

$$\hat{f}_r(t) = \frac{1}{E[T_r]} \left[1 - F_r(t)\right], \tag{3.10}$$

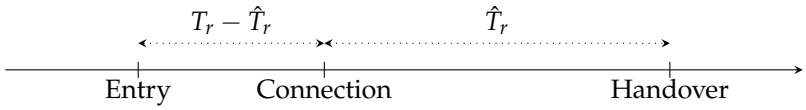

Figure 3.3: Residual life of the CRT

and therefore,

$$\hat{F}_r(t) = \int_0^x \frac{1}{E[T_r]} [1 - F_r(t)]dt. \tag{3.11}$$

Finally,

$$\hat{P}^h = 1 - \int_0^\infty (1 - \hat{F}_r(t_d)) f_d(t_d) dt_d. \tag{3.12}$$

Let N_i be the random variable given by the number of handovers executed by an elastic flow of SC i during its lifetime. Its distribution is given by:

$$P[N_i = n] = \begin{cases} \hat{P}_i^h (1 - P_i^h)(P_i^h)^{(n-1)} & n > 0 \\ 1 - \hat{P}_i^h & n = 0. \end{cases} \tag{3.13}$$

Then, N_i follows a geometric distribution with mean

$$E[N_i] = \frac{\hat{P}_i^h}{(1 - P_i^h)}.$$

In the system under study, the flow duration follows a PH distribution and therefore, the probability density function of the CRT is given by (see Appendix C.1.2):

$$f_d(t) = \boldsymbol{\alpha}_i e^{t\mathbf{S}_i} \boldsymbol{\tau}_i. \tag{3.14}$$

As a particular case, if we consider exponentially distributed CRT, from (3.10) we have:

$$\hat{f}_r(t) = \mu_i^{r,e}[1 - (1 - e^{-\mu_i^{r,e} t})] = \mu_i^{r,e} e^{-\mu_i^{r,e} t} = f_r(t)$$

and hence,

$$\hat{F}_r(t) = F_r(t) = 1 - e^{-\mu_i^{r,e} t}. \tag{3.15}$$

Therefore, when the CRT is considered exponentially distributed, the probabilities of handover for elastic flows of SC i, $\hat{P}_i^h$ and P_i^h have the same value and from (3.9), (3.14) and (3.15) we have:

$$\begin{aligned}\hat{P}_i^h = P_i^h &= 1 - \int_0^\infty e^{-\mu_i^{r,e} t_d} \boldsymbol{\alpha}_i e^{t_d \boldsymbol{S}_i} \boldsymbol{\tau}_i dt_d = \\ &= 1 - \boldsymbol{\alpha}_i \int_0^\infty e^{-t_d(\mu_i^{r,e}\boldsymbol{I} - \boldsymbol{S}_i)} dt_d \boldsymbol{\tau}_i = \\ &= 1 - \boldsymbol{\alpha}_i (\mu_i^{r,e}\boldsymbol{I} - \boldsymbol{S}_i)^{-1} \boldsymbol{\tau}_i \,,\end{aligned} \tag{3.16}$$

where $\boldsymbol{I}$ is the identity matrix.

3.3 Validation of the analytical model

In this section the analytical results are compared with simulation results in order to validate the correctness of the analytical model and the assumptions of exponential distributions for CRTs and flow sizes. For the simulation, we choose the lognormal distribution, as it models more realistically the CRT [ZBA09] and the fow size [ZSXX11].

We consider, unless otherwise indicated, a system with parameters indicated in Table 3.1. The notation has been simplified writing $\mu^{r,e}$, $\hat{P}^h$, P^h and N

Table 3.1: Definition of system parameters.

Parameter	Value	Parameter	Value
N_s	2	N_e	1
C	10	R	100
L	200	$\mathbf{b}$	$[1, 2]$
$\boldsymbol{\lambda^n}$	$[0.008, 0.012, 5]$	$\boldsymbol{\mu^{d,s}}$	$[0.008, 0.01]$
$\boldsymbol{\mu^{r,s}}$	$[0.004, 0.006]$	r_3^m	50
r_3^M	500	β_3^0	0
β_3^1	1		

instead of $\mu_3^{r,e}$, $\hat{P}_3^h$, P_3^h and N_3 respectively. The values of the arrival rates for handover arrivals, λ^h, are calculated by simulating a system with the same parameters than the analytical model under study in each case.

In Fig. 3.4 and Fig. 3.5, the evolution of the handover probability, both $\hat{P}^h$ and P^h, with $\mu^{r,e}$ are shown. Curves labeled with 'Exp' correspond to analytic results considering exponentially distributed CRTs and obtained using (3.16). Curves labeled with 'Logn P^h' and 'Logn $\hat{P}^h$' are obtained considering exponentially distributed CRTs to determine $(\boldsymbol{\alpha}, \boldsymbol{S})$, and then using lognormal distributions and their residual distributions to model the CRT when computing P^h and $\hat{P}^h$ from (3.9) and (3.12) respectively, as explained in Section 3.2.2. Its mean is set to $1/\mu^{r,e}$, while its Coefficient of Variation (CV[1]) is set to 1 in Fig. 3.4 and to 0.5 and 2 in Fig. 3.5. Curves corresponding to simulation results are labeled with 'SR'. They are obtained by considering a multi-cell scenario with a central cell and two outer rings of cells, which make a total 19 cells. Upon CRT termination, terminals select a neighbor cell with equal probability. We consider wraparound to avoid abnormal terminations at the edges. For more details see Appendix D.2. For these simulations, the random variable $\hat{T}_r$ for the residual life of the CRT has to be generated from its distribution function $\hat{F}_r$. For that, we use the method known as *acceptance-rejection method*. For more details about this method, see also Appendix D.2.

Observe the excellent agreement between the analytical and simulation results, i.e. they practically overlap. We conclude that the PH distribution models appropriately the flow duration, even when the CRT are lognormally distributed. We can also see that in both Fig. 3.4 and 3.5, P^h and $\hat{P}^h$ increase with $\mu^{r,e}$ as expected, because the CRT decreases as the rate increases. In Fig. 3.4, also as expected, P^h and $\hat{P}^h$ decrease as r_3^M increases. This is because the higher r_3^M is, the earlier the flow transfer will terminate.

Results in Fig. 3.6 show the evolution of the distribution of N (number of handovers) with $\mu^{r,e}$ for lognormally distributed CRTs with $CV = 1$. Analytical results, obtained using (3.13), are represented using lines, while simula-

[1]The Coefficient of Variation (CV) of the random variable X is defined as the ratio of the standard deviation σ_X to the mean $E[X]$, i.e., $CV_X = \sigma_X / E[X]$

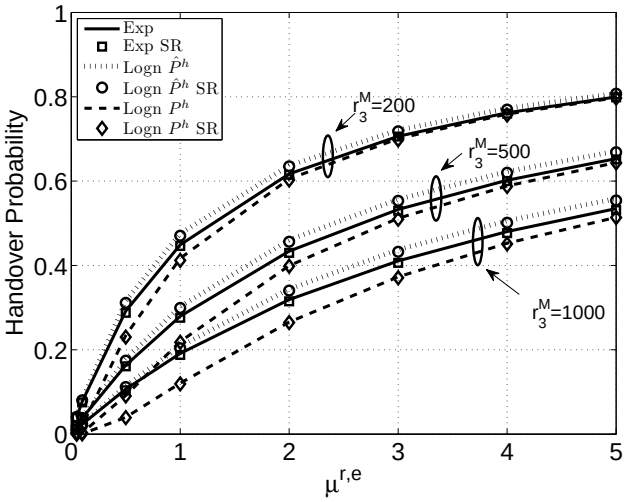

Figure 3.4: Handover probability as function of $\mu^{r,e}$, $CV = 1$.

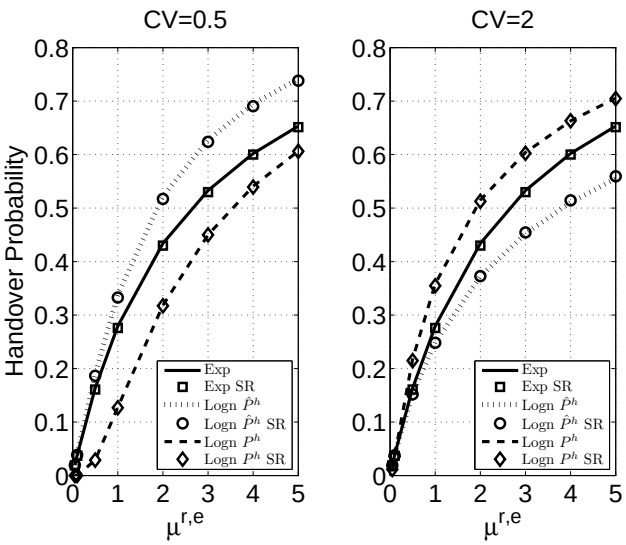

Figure 3.5: Handover probability as function of $\mu^{r,e}$ for b_3^M=500, $CV = 0.5$ and $CV = 2$.

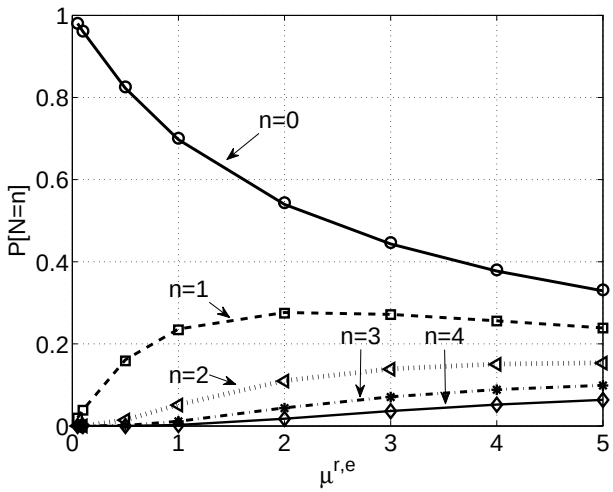

Figure 3.6: Probability of having n handovers as function of $\mu^{r,e}$.

tions results are represented only by markers at the evaluated points. Clearly, when the mobility increases, i.e. $\mu^{r,e}$ increases, $P[N = n]$, $n \geq 1$, increases, while $P[N = 0]$ decreases. That is, as the mobility increases it is more probable that a flow executes 1 or more handovers, while it is less probable that executes 0 handovers. Note that the curve for $n = 1$, first increases with $\mu^{r,e}$ and then decreases. This is because, the higher $\mu^{r,e}$ is, the less likely that a flow executes exactly 1 handover and the more likely that it executes more than 1.

Fig. 3.7 shows the impact of the flow size distribution on the distribution of N, for lognormal CRT times with $CV = 1$. Analytical results, represented by lines, are obtained considering an exponential flow size distribution and using (3.13). Simulation results, represented only by markers at the evaluated points, are obtained considering a lognormal flow size distribution with $CV = 2$. Note that when the flow size is exponentially distributed, the memoryless property for the flow size holds and, after a handover, the residual flow size maintains the original distribution. However, when the flow size

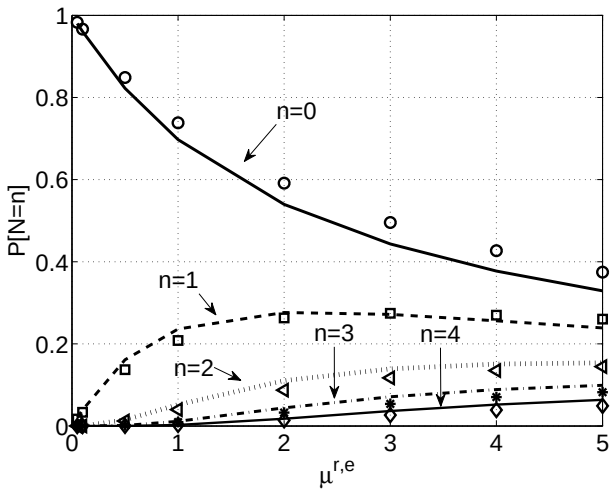

Figure 3.7: Probability of having n handovers as function of $\mu^{r,e}$. Lognormal flow size distribution, $CV = 2$.

distribution is lognormal, the residual flow size does not maintain the original distribution after a handover, i.e. handover probabilities depend on the number of previous handovers. This fact has been taken into account in the simulation model. Results in Fig. 3.7 show an excellent agreement between the analytical and simulation results. As expected, the agreement is closer for low mobility scenarios (low $\mu^{r,e}$). That is, when the number of handovers experienced is 0 or close to 0, the difference in the residual distributions of the flow size has no impact. However, in high mobility scenarios, the proposed model is still able to capture with high precision the handover performance of the elastic flows.

3.4 Conclusions

We studied the flow duration distribution for elastic traffic in cellular networks and obtained handover related metrics, such as the distribution of the

number of handovers a flow has to go through. To make the model more realistic and of broader application, we considered impatient elastic users, a generic admission policy that guarantees a minimum rate for elastic flows, and scenarios where CRTs and flow sizes were modeled by different distributions.

The flow duration was characterized by a PH distribution and the probabilities of the first and successive handovers were determined. We provide exact results under the assumption that both flow sizes and CRTs are exponentially distributed. Then, the model was extended by introducing an approximate technique to deal with more realistic distributions of the CRT and the flow size. We used lognormal distributions because, as suggested in the literature, they model realistically both CRTs and flow sizes. Note the modeling approach can be used with any other distribution. Results obtained by the proposed technique were validated by simulation and very close agreement was found.

We conclude that the proposed method models appropriately the flow duration and handover performance under rather general assumptions.

Chapter 4

Optimal design of the MFGC policy

4.1 Introduction

Efficient AC policies must be used to optimize the resource utilization and fulfill QoS constraints in multiservice mobile networks. Dropping a session in progress is generally considered to have a more negative impact from the users' perception than blocking a new session. What is actually done is reserving resources for potential future handover requests. In this sense, AC policies consider that handover sessions have higher priority than new requests and therefore, one of the key goals is to minimize the handover blocking probability. Moreover, the goal of a network service provider is to maximize revenue by improving network resource utilization. This is achieved by maximizing the carried traffic, i.e. by keeping the blocking probabilities for new sessions low. When more resources are reserved for handover requests (lower handover blocking probability), more new sessions are blocked. Therefore, the main challenge in the design of an efficient AC policy is to balance these two conflicting requirements.

In a single service scenario, the GC and the FGC trunk reservation policies are known to be optimal under different criteria [RTN97, Bar01] (see Sec-

tion 2.3). However, these results cannot be extended to their counterparts in multiservice scenarios [BC02, PCG03, GRMBP05], the MGC and the MFGC policies [HUCPOG03b]. In [BC02], it is shown that the optimal AC policy (with respect to a certain cost function) in multiservice scenarios belongs to the wider family of RS policies [Ros70, Ros95]. In spite of that, the MFGC policy is still an efficient AC policy when compared to product-form policies [GRMBP04, GRMBP05]. In the MFGC policy, the policy parameters controls the number of resources that each SC can access. The optimal parameter setting maximizes the carried traffic that the system can handle while meeting certain QoS requirements. High precision is required when determining this optimal configuration and therefore, computing the optimal parameter setting is computationally costly [CC97]. As a consequence, in large systems with a high number of SCs the computational complexity of adjusting these parameters could be intractable.

The design of the optimal configuration of MFGC policy is typically based on an iterative analysis procedure that adjusts the configuration parameters of the MFGC to maximize the offered traffic that the system can handle while meeting certain QoS objectives. In [PMCG05] a methodology to determine the exact parameter values of the MFGC policy is proposed, including some speed up suggestions, but its computational cost is prohibitive for practical systems. In [CPVAOG04], an approximation based on the Kaufman and Roberts (K&R) recursion is proposed to compute the parameters of the MFGC policy. Although the computational cost greatly is reduced, no indication is provided about its precision. In [BM07], a new procedure to determine the parameters of the MFGC policy is proposed. It is an enhancement of the one proposed in [PMCG05] that uses the approximation based on the K&R recursion to quickly approximate to the surroundings of the solution and then a modified version of [PMCG05] is developed to achieve the final solution. Henceforth, we refer to the algorithm proposed in [PMCG05] as PMC, to the approximation based in K&R as CVO and to the methodology in [BM07] as BGMP. In this chapter, an algorithm to determine the optimal parameters of MFGC, which is based on a modified version of the adaptive AC proposed

in [GRDBMBP07], is proposed. Then, it is compared the computational cost necessary to obtain the optimal configuration of the MFGC policy using the CVO approximation, the PMC, the BGMP and the adaptive algorithms. This work resulted into the publication in [BMGRPMB08].

The policy MFGC has been proposed to deal with only streaming traffic since the elastic traffic uses the capacity not occupied by streaming traffic, and only the streaming SCs are competing directly for the resources. Thus, in this chapter, only SCs which carry streaming traffic are considered in order to find the optimal configuration of the MFGC parameters.

This chapter is structured as follows. First, in Section 4.2 the system model and the QoS objectives are described. In Section 4.3, the aforementioned algorithms are briefly described and the adaptive algorithm is presented. Some results and conclusions are discussed, respectively, in Sections 4.4 and 4.5.

4.2 System model and MFGC policy definition

We consider that all the SCs in the system carry streaming traffic. The system model and the streaming traffic characterization are the same as in the system model described in Section 3.2. Each cell has a total of C resource units. The system offers N different streaming SCs, where for each SC, new and handover arrivals are distinguished, so that there are N SCs and $2N$ arrival types.

For the sake of mathematical tractability we make the common assumptions of Poisson arrival processes for all SCs and exponentially distributed session durations and CRTs for streaming SCs. Let λ_i^n (λ_i^h) be the arrival rate for new (handover) streaming sessions of the ith SC. We define the aggregated arrival rate as

$$\lambda^T = \sum_{1 \leq i \leq N} \lambda_i^n, \tag{4.1}$$

where if f_i is the fraction of λ^T that corresponds to SC i,

$$\lambda_i^n = f_i \lambda^T, \qquad 0 \leq f_i < 1, \qquad \sum_{1 \leq i \leq N} f_i = 1. \tag{4.2}$$

We refer to the maximum offered rate that the network can handle while meeting certain QoS requirements as λ_{max}^T.

Let $\mu_i^{d,s}$ and $\mu_i^{r,s}$ be the rates of the session duration and the CRT of the ith streaming SC, $1 \leq i \leq N$. Hence, the CHT in a cell for a streaming SC is exponentially distributed with rate $\mu_i^s = \mu_i^{d,s} + \mu_i^{r,s}$. A request of the ith SC consumes b_i resource units, $b_i \in \mathbb{N}$. The system state is described by the N-tuple $\boldsymbol{x} = (x_1, \ldots, x_N)$, where x_i is the number of ongoing sessions of the ith SC, regardless they were initiated as new or handover arrivals. The number of resources occupied at state $\boldsymbol{x}$ is denoted by $b(\boldsymbol{x}) = \sum_{i=1}^N x_i b_i$.

The system can be modeled as a CTMC, specifically, as a multidimensional birth and death process with state space,

$$\mathcal{W} := \left\{ \boldsymbol{x} : x_i \in \mathbb{N}; \sum_{i=1}^{N} x_i b_i \leq C \right\} .$$

In order to describe the *session-level* QoS in multiservice mobile cellular networks, the most common parameters used are: *new session blocking probability*, $P_i^{b,n}$, which is the probability that a new request of SC i is not accepted in the system, and the *handover blocking probability*, $P_i^{b,h}$, which is the probability that a handover request of SC i is not accepted in the destination cell.

The coefficients $a_i^n(\boldsymbol{x})$ and $a_i^h(\boldsymbol{x})$ denote the probabilities of accepting a new and a handover arrival of SC i respectively, and $\pi(\boldsymbol{x})$ is the stationary distribution. Then, the new session and handover blocking probabilities for SC i, respectively, are obtained as

$$\begin{aligned} P_i^{b,n} &= \sum_{\boldsymbol{x} \in \mathcal{W}} (1 - a_i^n(\boldsymbol{x})) \pi(\boldsymbol{x}), \\ P_i^{b,h} &= \sum_{\boldsymbol{x} \in \mathcal{W}} (1 - a_i^h(\boldsymbol{x})) \pi(\boldsymbol{x}). \end{aligned} \tag{4.3}$$

We consider that the QoS requirements are given in terms of upper-bounds for the new session blocking probability, B_i^n, and the handover blocking probability, B_i^h.

If the system is in statistical equilibrium, the rate at which handover sessions enter the cell is equal to the rate at which handover sessions exit the cell. Then, handover arrival rates are related to the new session arrival rates and the blocking probabilities [LMN94] through the expression:

$$\lambda_i^h = \frac{\mu_i^{r,s}}{\mu_i^{r,s} + \mu_i^{d,s}} \left((1 - P_i^{b,n})\lambda_i^n + (1 - P_i^{b,h})\lambda_i^h \right),$$

and therefore,

$$\lambda_i^h = \lambda_i^n \frac{(1 - P_i^{b,n})}{\mu_i^{d,s}/\mu_i^{r,s} + P_i^{b,h}} . \tag{4.4}$$

The blocking probabilities depend on the handover arrival rates and hence in (4.4), λ_i^h is not explicitly defined. Since the MFGC policy will be designed so that the blocking probabilities will be very close to their upper-bounds, instead of using (4.4) we use the expression:

$$\lambda_i^h = \lambda_i^n \frac{(1 - B_i^n)}{\mu_i^{d,s}/\mu_i^{r,s} + B_i^h} . \tag{4.5}$$

MFGC policy

A brief definition of the MFGC policy is given here. Two parameters are associated with SC i: t_i^n and t_i^h for new and handover arrivals, respectively. Henceforth, we use the superscript (n, h) to refer to the corresponding parameters for new or handover arrivals, for example, $t_i^{n,h}$ means t_i^n or t_i^h. These parameters are real numbers in the interval $[0, C]$. In order to decide on the acceptance of a request of SC i, upon an arrival the number of resources that will be occupied if it is accepted is compared with the corresponding threshold $t_i^{n,h}$ depending on whether the request is a new or a handover arrival.

The following decisions can be taken:

$$b(\boldsymbol{x}) + b_i \begin{cases} \leq \lfloor t_i^{n,h} \rfloor & \text{accept} \\ = \lfloor t_i^{n,h} \rfloor + 1 & \text{accept with probability} \quad t_i^{n,h} - \lfloor t_i^{n,h} \rfloor \\ > \lfloor t_i^{n,h} \rfloor + 1 & \text{reject.} \end{cases} \tag{4.6}$$

The policy parameters $t_i^{n,h}$ control the number of resource units that a SC can access. Thus, the coefficients $a_i^n(\boldsymbol{x})$ and $a_i^h(\boldsymbol{x})$ can be determined as follows:

$$a_i^n(\boldsymbol{x}) \begin{cases} = 1 & \text{if} \quad b(\boldsymbol{x}) + b_i \leq \lfloor t_i^n \rfloor \\ = t_i^n - \lfloor t_i^n \rfloor & \text{if} \quad b(\boldsymbol{x}) + b_i = \lfloor t_i^n \rfloor + 1 \\ = 0 & \text{if} \quad b(\boldsymbol{x}) + b_i > \lfloor t_i^n \rfloor + 1, \end{cases} \tag{4.7}$$

$$a_i^h(\boldsymbol{x}) \begin{cases} = 1 & \text{if} \quad b(\boldsymbol{x}) + b_i \leq \lfloor t_i^h \rfloor \\ = t_i^h - \lfloor t_i^h \rfloor & \text{if} \quad b(\boldsymbol{x}) + b_i = \lfloor t_i^h \rfloor + 1 \\ = 0 & \text{if} \quad b(\boldsymbol{x}) + b_i > \lfloor t_i^h \rfloor + 1. \end{cases} \tag{4.8}$$

On average, the maximum number of resource units that a new and a handover arrival of SC i can utilize are, respectively, t_i^n and t_i^h. The optimal parameters maximize λ^T under given QoS requirements. Since the parameters $t_i^{n,h}$ have an impact not only on the QoS perceived by SC i but also on the QoS perceived by the rest of SCs, the adjustment of the threshold parameters $t_i^{n,h}$ is not simple.

4.3 Parameter configuration

4.3.1 Previous algorithms

The optimal design of the MFGC policy maximizes the total offered traffic, λ^T, that the system can handle while meeting the QoS objectives.

The capacity optimization problem can be formally stated as follows

Given: C, b_i, f_i, $\mu_i^{d,s}$, $\mu_i^{r,s}$, B_i^n, B_i^h ; $i = 1, \ldots, N$

Maximize: λ^T by finding the appropriate MFGC parameters t_i^n and t_i^h; $i = 1, \ldots, N$

Subject to: $P_i^{b,n} \leq B_i^n$, $P_i^{b,h} \leq B_i^h$; $i = 1, \ldots, N$

Let us introduce the 2N-tuple $\boldsymbol{p_{max}} = (B_1^n, \ldots, B_N^n, B_1^h, \ldots, B_N^h)$ as the upper-bound vector for the blocking probabilities.

In order to find the optimal $t_i^{n,h}$ and λ_{max}^T, the PMC algorithm [PMCG05] proceeds as follows. The algorithm has a main part (Algorithm **pmc**), where λ^T is assigned an initial value. It is checked if a set of values for $t_i^{n,h}$ that fulfill the QoS requirements exists by calling the procedure **sMFGCpmc**, where the set of $t_i^{n,h}$ are initialized with small values. If this set exists, λ^T is increased and if not, it is decreased, first with big steps and later with smaller steps. The procedure **sMFGCpmc** is called again until a λ_{max}^T is found.

Algorithm:

$(\lambda_{max}^T, \boldsymbol{t}_{opt})$=**pmc**$(\boldsymbol{p_{max}}, \boldsymbol{f}, \boldsymbol{\mu^{d,s}}, \boldsymbol{\mu^{r,s}}, \boldsymbol{b}, C)$

Procedure:

(**ok**,$\boldsymbol{t}$)=**sMFGCpmc**$(\boldsymbol{p_{max}}, \boldsymbol{\lambda_n}, \boldsymbol{\mu^{d,s}}, \boldsymbol{\mu^{r,s}}, \boldsymbol{b}, C)$

The procedure **sMFGCpmc** does, in turn, call another procedure (procedure **MFGCpmc**) that calculates the blocking probabilities by solving the balance equations. Thus, at each new value of $t_i^{n,h}$ for a given λ^T a CTMC, specifically a multidimensional birth and death process, has to be solved. In PMC algorithm, the stationary distribution equations (3.6) and the normalization equation (3.7) are solved by using the Gauss-Seidel method. Solving this process is the most computationally expensive part of the algorithm. The computational cost grows enormously when the number of resource units or/and the number of different SCs is high [CC97].

To face these computational limitations the CVO numerical approximation is proposed in [CPVAOG04]. It consists of converting the multidimensional

process to a one-dimensional process where the system state is defined by $b(\boldsymbol{x})$, the total number of resource units occupied. Henceforth, let us refer to $b(\boldsymbol{x})$ as k. Then, the stationary distribution of k given by $q(k)$ can be generated recursively by the equation:

$$\sum_{i=1}^{N} \frac{a_i^n(k-b_i)\lambda_i^n + a_i^h(k-b_i)\lambda_i^h}{\mu_i^{d,s} + \mu_i^{r,s}} b_i q(k-b_i) = kq(k), \tag{4.9}$$

where $a_i^n(k)$ and $a_i^h(k)$ are the probabilities of accepting a new session or a handover of SC i, respectively, that arrives in state k. With this stationary distribution the blocking probabilities can be obtained alternatively by:

$$\begin{aligned} P_i^{b,n} &= \sum_{r=0}^{C} (1 - a_i^n(r))q(r), \\ P_i^{b,h} &= \sum_{r=0}^{C} (1 - a_i^h(r))q(r), \end{aligned} \tag{4.10}$$

and the procedure **MFGCpmc** of the PMC algorithm will be faster. This recursion is only an approximation, since the model does not take into account some reservations [SG00] and the final results may not be accurate. As shown later, the accuracy of this approximation is poor. Therefore, a new algorithm (BGMP algorithm) is proposed in [BM07] that uses the CVO approximation, among other modifications, to improve the PMC algorithm.

An enhancement of the PMC algorithm is possible by initializing the value of the parameters $t_i^{n,h}$ and the aggregated call arrival rate λ^T as close as possible to the optimal values. So, in BGMP algorithm, in a first step $t_i^{n,h}$ and λ^T are initialized with the values obtained using the CVO approximation. In the second step, these results are the initial values and the multidimensional birth and death process is solved by using the Gauss-Seidel method. Thus, the cost of the PMC algorithm is reduced considerably. In this second step, the initial interval of λ^T will be narrower, therefore the search of the optimal λ^T can be faster. Moreover, in each evaluation of each new λ^T the initial values of $t_i^{n,h}$ are the calculated values in the previous evaluation. These

changes improve the algorithm since less multidimensional birth and death processes have to be solved. For more details about the PMC and the BGMP algorithms see Appendix B.1.

4.3.2 Adaptive scheme

The adaptive scheme presented in this work is based on the adaptive scheme proposed in [GRDBMBP07] and operates in coordination with the Multiple Guard Channel (MGC) policy. In the MGC policy, one threshold parameter for new arrivals, l_i^n and another for handover arrivals l_i^h, are associated to each SC i, where $l_i^n,\ l_i^h \in \mathbb{N}$. We use $l_i^{n,h}$ to refer to the corresponding l_i^n or l_i^h. Upon a new (handover) arrival, the MGC policy takes the following decisions:

$$b(\boldsymbol{x}) + b_i \begin{cases} \leq l_i^{n,h} & \text{accept} \\ > l_i^{n,h} & \text{reject.} \end{cases} \tag{4.11}$$

Therefore, l_i^n (l_i^h) is the number of resources that a new (handover) arrival of SC i can access. In practice, we can assume without loss of generality that the QoS objective for SC i can be expressed as

$$B_i^{n,h} = \frac{c_i^{n,h}}{o_i^{n,h}},$$

where $c_i^{n,h}, o_i^{n,h} \in \mathbb{N}$. Then it is expected that if $P_i^{n,h} = B_i^{n,h}$ the SC i will experience, in average, $c_i^{n,h}$ rejected requests and $o_i^{n,h} - c_i^{n,h}$ admitted requests, out of $o_i^{n,h}$ offered requests. For example, if the QoS objective for SC i is $B_i^n = 1/100$, then $c_i^n = 1$ and $o_i^n = 100$.

It seems intuitive to think that the adaptive scheme should adjust the threshold parameters on the required direction if the perceived QoS is different from its QoS requirements, and not change the threshold parameters of those arrival SCs which meet their QoS requirements. Therefore, given that the MGC policy deploys integer values for its threshold parameters, a

probabilistic adjustment each time a request is processed is proposed in the following way. First of all, we choose an arrival type that, for simplicity, we assume that it is the new arrival of SC $i = 1$, which is the SC with the lowest priority.

1. For new arrivals of SC $i \in [2, N]$ and handover arrivals of SC $i \in [1, N]$:
   - If accepted, do $\{l_i^{n,h} \leftarrow (l_i^{n,h} - \Delta l)\}$ with probability $1/(o_i^{n,h} - c_i^{n,h})$;
   - If rejected, do $\{l_i^{n,h} \leftarrow (l_i^{n,h} + \Delta l)\}$ with probability $1/c_i^{n,h}$, where $\Delta l \in \mathbb{N}$ is the adjustment step for the thresholds parameters.
2. For new arrival of SC $i = 1$:
   - If accepted, do $\{\lambda^T \leftarrow (\lambda^T + \Delta\lambda)\}$ with probability $1/(o_1^n - c_1^n)$;
   - If rejected, do $\{\lambda^T \leftarrow (\lambda^T - \Delta\lambda)\}$ with probability $1/c_1^n$, where $\Delta\lambda \in \mathbb{R}$ is the adjustment step for the λ^T.

The thresholds $l_i^{n,h}$ of high priority are adjusted independently from each other according to whether their requests are accepted or rejected. If a lot of requests of new or handover arrivals of SC i are rejected, it can be that $l_i^{n,h} \geq C$. In this case, l_1^n is decremented. Thus, new arrivals of SC 1 do not control their own threshold l_1^n but control their QoS like the other SCs by adjusting λ^T. When $P_1^n > B_1^n = c_1^n/o_1^n$, λ^T is decremented and if all objectives are fulfilled, λ^T is incremented.

Therefore, under stationary traffic if $P_i^{n,h} = B_i^{n,h}$, in average, $l_i^{n,h}$ is increased by Δl and decreased by Δl every $o_i^{n,h}$ offered requests, i.e. its mean value is kept constant. The optimal MFGC threshold parameters, $t_i^{n,h}$, correspond to this average value of $l_i^{n,h}$. Note also that in the operation of this simple scheme no assumption has been made concerning the arrival processes or the distribution of the session duration and CRT.

Table 4.1: Definition of systems A and B.

	System A	System B
N	2	4
$\boldsymbol{b}$	$[1,2]$	$[1,2,4,6]$
$\boldsymbol{\mu}^{d,s}$	$[1/180, 1/300]$	$[1/300, 1/300, 1/300, 1/300]$
$\boldsymbol{\mu}^{r,s}$	$[1/900, 1/1000]$	$[1/300, 1/300, 1/300, 1/300]$
$\boldsymbol{B}^n$	$[0.02, 0.02]$	$[0.05, 0.04, 0.03, 0.02]$
$\boldsymbol{B}^h$	$[0.01002, 0.00668]$	$[5.0251, 4.0161, 3.009, 2.004]10^{-3}$

4.4 Numerical evaluation

In this section we make a comparative evaluation of the CVO approximation when it is used to solve the balance equations in the PMC algorithm, the BGMP algorithm and the adaptive method, in terms of accuracy and numerical complexity given as the computational cost in seconds necessary to compute the optimal parameter setting of the MFGC policy using an Intel Pentium IV HT 3,4GHz.

For numerical examples we have considered two systems: i) System A with two streaming SCs ($N = 2$), and parameters from [PMCG05]; ii) System B with four streaming SCs ($N = 4$) and parameters from [BS97]. The parameters of system A and B are defined in Table 4.1. For system A, the fraction of the aggregated rate that corresponds to each SC i is $f_i = [0.8, 0.2]$ and for system B, it is given by

$$f_i = \hat{f}_i / \hat{F} \qquad \text{where} \qquad \hat{f}_i = \phi^{i-1}, \quad \phi = 0.2, \quad \hat{F} = \sum \hat{f}_1.$$

In Table 4.2, the results obtained in system A are shown with $C = 50$ and $C = 100$. In each column, it is shown the parameter setting that maximizes λ^T while meeting QoS requirements and this λ^T_{max}. The computational cost in seconds necessary to compute these parameters is also shown in the last column T(sec). Each column defines the results obtained using the indicated

Table 4.2: Parameter computation for system A.

	$C = 50$			$C = 100$		
	CVO	**BGMP**	**Adapt.**	**CVO**	**BGMP**	**Adapt.**
t_1^n	43.83	44.91	44.82	93.32	94.27	93.62
t_2^n	48.76	48.93	48.88	98.67	98.58	97.90
t_1^h	45.09	45.88	45.84	94.82	95.42	94.77
t_2^h	49.88	49.94	49.90	99.88	99.61	98.94
λ_{max}^T	0.146	0.151	0.150	0.336	0.345	0.339
T(sec)	11	343	511	32	3265	956

algorithm. In the column CVO, the CVO approximation is used to solve the balance equations in the PMC algorithm. When the Gauss-Seidel method is used to solve the balance equations, the PMC algorithm has quite higher computational cost to achieve similar results as algorithm BGMP. For $C = 50$, the computational cost of the PMC algorithm with the Gauss-Seidel method is 2433 sec and for $C = 100$, it is 27018 sec.

Once the parameter setting and the λ_{max}^T are computed, in order to verify that the QoS requirements are fulfilled, the new session and handover blocking probabilities, $P_i^{b,n}$ and $P_i^{b,h}$, under this design can be calculated by solving the balance equations of the system using the Gauss-Seidel method. Table 4.3 contains the relative error (%) value of the blocking probabilities calculated using the parameter design and the λ_{max}^T from each method, in relation to the upper-bounds for the new session and handover blocking probabilities, B_i^n and B_i^h. Thus, the error expressions of new session and handover blocking probabilities for SC i are given, respectively, by

$$\text{error}_i^n = \frac{P_i^{b,n} - B_i^n}{B_i^n}, \qquad \text{error}_i^h = \frac{P_i^{b,h} - B_i^h}{B_i^h}.$$

Negative errors refer to blocking probabilities lower than the objectives. Otherwise, if the error is positive, the blocking probabilities obtained are higher than the upper-bounds.

Table 4.3: System A errors (%).

	$C = 50$			$C = 100$		
	CVO	**BGMP**	**Adapt.**	**CVO**	**BGMP**	**Adapt.**
$P_1^{b,n}$	-2.98	-0.06	-0.90	-7.63	-0.22	-0.29
$P_2^{b,n}$	-45.96	-0.01	-0.37	-52.43	-0.51	-0.59
$P_1^{b,h}$	-19.34	-0.17	-1.17	-24.68	-0.68	-0.96
$P_2^{b,h}$	-55.63	-0.59	-3.46	-64.24	-0.01	-5.10

Table 4.4: Parameter computation for system B, $C = 50$.

	CVO	**BGMP**	**Adapt.**
t_1^n	39.47	40.06	39.77
t_2^n	40.93	41.55	41.26
t_3^n	43.59	44.55	43.78
t_4^n	46.40	46.81	46.52
t_1^h	43.83	43.94	43.66
t_2^h	45.14	45.22	44.90
t_3^h	47.56	47.54	47.24
t_4^h	49.97	49.85	49.69
λ_{max}^T	0.078	0.079	0.078
T(sec.)	38	325446	2298

The results presented indicate that the CVO approximation is not an accurate method to compute the parameters since the blocking probabilities calculated using the parameters obtained do not adjust to the objectives. In terms of computational cost, the CVO approximation is the fastest method. The BGMP algorithm gives good results but it can entail a high computational cost if the number of resource units is high. The adaptive method achieves good precision and although its computational cost is higher than using the BGMP algorithm when the system has few resource units, it is faster than the BGMP algorithm when the number of resource units is high.

Table 4.5: System B errors (%), $C = 50$.

	CVO	**BGMP**	**Adapt.**
$P_1^{b,n}$	3.54	−0.10	−0.22
$P_2^{b,n}$	3.52	−0.55	−0.42
$P_3^{b,n}$	−0.13	−0.67	−0.08
$P_4^{b,n}$	−4.57	−0.56	−0.31
$P_1^{b,h}$	−19.64	−0.39	−0.56
$P_2^{b,h}$	−22.64	−0.37	−0.03
$P_3^{b,h}$	−28.95	−0.52	−0.01
$P_4^{b,h}$	−40.14	−0.45	−4.52

Similarly, in Tables 4.4 and 4.5, the same results for system B are showed. For this system, the design using the PMC algorithm together with the Gauss-Seidel method is not computationally tractable. The results show that the computational cost when the CVO approximation is used is very low but the results are not accurate since some relative errors in Table 4.5 are high and some errors are positive, which means that the QoS requirements are not fulfilled for all the SCs with the parameter design computed with this approximation. We can also see that although the BGMP algorithm is very accurate, it has the highest computational cost. The computational cost of the adaptive algorithm is between the other two algorithms and the computed parameter design adjusts very accurately the blocking probabilities to the QoS requirements.

4.5 Conclusions

In this chapter, we have compared several algorithms used to design the optimal parameter configuration for the MFGC policy in order to maximize the offered traffic that the system can handle while meeting certain QoS requirements. For large systems, i.e. systems with a large number of resource units

and/or various SCs, the design of the MFGC policy can become computationally intractable, therefore it is crucial to choose a suitable method when designing these types of AC policies.

We have observed that for large systems, the PMC algorithm using the Gauss-Seidel method presents very accurate results but its computational cost can be prohibitive, while using the CVO approximation is very fast but it does not adjust to the QoS requirements. We have also observed that the BGMP algorithm presents good results, but its computational cost can be very high when the number of SCs in the system is considerably high. For large systems, the adaptive method achieves good precision and its computational cost is between the PMC algorithm using the CVO approximation and the BGMP algorithm.

Chapter 5

AC policies for time-varying traffic scenarios

5.1 Introduction

As we have shown in Chapter 4, in large systems the computational complexity of computing the optimal parameter setting of MFGC policy can be intractable. Moreover, after the design phase, this parameter configuration is static. Then, it is reasonable to expect that this AC policy may have poor performance when the offered traffic is time-varying and overload intervals appear when some SCs exceed the expected offered traffic. We can find these scenarios, for instance, in public cellular networks that support emergency services after a disaster [BF01], or in general, in networks that support high levels of congestion, where there are high priority SCs that can generate high demands of resource units. In this context, the main problem is to provide a reasonable QoS to the different SCs under high unexpected overloads. Therefore a fair, efficient and robust AC is needed.

A fair and efficient AC avoids that the different characteristics of the SCs results in a very unfair resource allocation. Robustness is a key aspect for an effective resource sharing. It is the ability to respond to statistical fluctuations,

which are inevitable even with good traffic forecasts. It is also the adaptability in an overloaded scenario where the arrival rates are higher than the expected values considered in the design of the AC policy. For instance, the CP and the CS policies are two extreme cases. Under the CP policy, when some SCs have an arrival rate higher than the expected but the overall traffic is light the resources are underutilized since each SC has its nominal allocation and can not utilize more resources, even when there are free resources in the resources allocated to other SCs. The CS policy presents the opposite case: when some SCs are overloaded, they can overwhelm all the others since the resources are shared indiscriminately.

In this chapter, the *Virtual Partitioning* (VP) policy [MZ96] is studied and compared with the MFGC policy. The VP policy can adapt to the fluctuations of the system using different static parameter configurations depending on whether one or some of the SCs are in overload. As a result a good balance between efficiency, fairness and robustness is obtained [BM98]. The performance of this policy has been studied in previous papers [BM98, MRW98, YMW+04, SNW08], but the authors consider either networks without mobility or mobile cellular networks that support only streaming traffic. In this chapter, we present a new design of the VP policy for multiservice mobile cellular networks which support elastic and streaming traffic. This work resulted into the publication in [BMPMB10b].

In the next section, the system model is presented and a basic definition of VP policy is described. In Section 5.3, the new design of the VP policy is described. In Section 5.4, the performance of the new proposed VP policy design and the performance of MFGC policy are compared. Section 5.5 concludes the chapter.

5.2 System model and VP policy definition

The system model and the traffic characterization are the same as those in Section 3.2. Each cell has a total of C resource units, each of them has a

capacity of R bits per second. The system offers N_s streaming SCs, N_e elastic SCs and $N = N_s + N_e$ total SCs. Elastic flows use the capacity not occupied by streaming traffic. To avoid starvation, the system reserves 1 resource unit for elastic traffic. For each SC, new and handover arrivals are distinguished, so that there are N SCs and $2N$ arrival types.

For the sake of mathematical tractability we make the common assumptions of Poisson arrival processes and exponentially distributed CRT for all SCs and only for streaming SCs, exponentially distributed session durations. A request of the ith SC consumes b_i resource units, $b_i \in \mathbb{N}$. For elastic traffic, the ith elastic SC can be served at a maximum rate, μ_i^M, and all the SCs must be served at least with a minimum rate, μ^m. The system state is described by the N-tuple $\boldsymbol{x} = (x_1, \ldots, x_N)$, where x_i is the number of ongoing streaming sessions or elastic flows of the ith SC. The number of resources occupied at state $\boldsymbol{x}$ by streaming SCs is denoted by $b(\boldsymbol{x})$. For elastic traffic, the session duration at state $\boldsymbol{x}$ is modeled by an exponential distribution with rate $\mu_i^{d,e}(\boldsymbol{x})$ defined in (3.2) and (3.3) and the patience time at state $\boldsymbol{x}$ is modeled by an exponential distribution with rate $\beta_i(\boldsymbol{x})$ defined in (3.4). The system can be modeled as a multidimensional birth and death process with state space $\mathcal{W}$ defined in (3.5).

Remember that $a_i^n(\boldsymbol{x})$ and $a_i^h(\boldsymbol{x})$ are the coefficients which define the AC, and $\pi(\boldsymbol{x})$ is the stationary distribution of the system. In order to describe the QoS, for streaming traffic we use the new session blocking probability, $P_i^{b,n}$, and the handover blocking probability, $P_i^{b,h}$, defined in 4.3. Again, we consider that dropping a session in progress has a more negative impact from the users' perception than blocking a new requested session, and therefore handover sessions have higher priority than new sessions. For elastic SCs, other parameters are also used to describe the QoS. In a cell, the *abandonment probability*, P_i^a, is the probability that an elastic flow of SC i is aborted due to impatience, where $i = 1 + N_s, \ldots, N$, and it is given by:

$$P_i^a = \frac{1}{\lambda_i^n(1 - P_i^{b,n}) + \lambda_i^h(1 - P_i^{b,h})} \sum_{\boldsymbol{x} \in \mathcal{W}} x_i \beta_i(\boldsymbol{x}) \pi(\boldsymbol{x}). \tag{5.1}$$

If all flows were let into the system, the abandonment probability may be considered a good and sufficient performance indicator. However, if there is some type of access restriction, both abandonment and blocking should be taken into account for characterizing the system performance. In the latter, relying only on the abandonment probability may lead to inappropriate conclusions since a low value for P_i^a can be obtained by simply using a more restrictive AC, which obviously entails a high value of the blocking probabilities. Therefore, we define the *success completion probability* P_i^c which represents the probability that a flow of SC i is not blocked and it does not leave the system due to impatience before being served.

In order to calculate P_i^c, it is necessary to know the handover probability for elastic flows, P_i^h. This parameter was studied in Chapter 3, and when the CRT is considered exponentially distributed, it is defined in (3.16). Let us define P' as the probability that in a cell, a session which has arrived after a handover terminates successfully in the cell or undergoes a successful handover to other cell. For the sake of brevity and clarity in the following expressions, let $\bar{P}$ be the complementary probability of P, $\bar{P} = 1 - P$, where P can be any of the probabilities referred here. Then, we define the success probability P_i^c as:

$$P_i^c = \bar{P}_i^{b,n} \bar{P}_i^a (\bar{P}_i^h + P_i^h P'), \tag{5.2}$$

where P' is given by:

$$P' = \bar{P}_i^{b,h} \bar{P}_i^a (\bar{P}_i^h + P_i^h P') = \bar{P}_i^{b,h} \bar{P}_i^a \bar{P}_i^h + \bar{P}_i^{b,h} \bar{P}_i^a P_i^h P',$$

and solving for P',

$$P' = \frac{\bar{P}_i^{b,h} \bar{P}_i^a \bar{P}_i^h}{1 - \bar{P}_i^{b,h} \bar{P}_i^a P_i^h}. \tag{5.3}$$

Substituting (5.3) in (5.2),

$$P_i^c = \bar{P}_i^{b,n} \bar{P}_i^a \bar{P}_i^h + \frac{\bar{P}_i^{b,h} \bar{P}_i^a \bar{P}_i^h}{1 - \bar{P}_i^{b,h} \bar{P}_i^a P_i^h} P_i^h \bar{P}_i^{b,n} \bar{P}_i^a$$

$$P_i^c = \bar{P}_i^{b,n}\bar{P}_i^a\bar{P}_i^h\left(1+\frac{\bar{P}_i^{b,h}\bar{P}_i^aP_i^h}{1-\bar{P}_i^{b,h}\bar{P}_i^aP_i^h}\right).$$

And finally,

$$P_i^c = \frac{(1-P_i^{b,n})(1-P_i^a)(1-P_i^h)}{1-(1-P_i^{b,h})(1-P_i^a)P_i^h}. \tag{5.4}$$

The QoS requirements are given in terms of upper-bounds for the new session blocking probability, B_i^n, and the handover blocking probability, B_i^h, and by lower-bounds of success completion probabilities, B_i^c.

If the system is in statistical equilibrium, the handover arrival rates for streaming traffic are related to the new session arrival rates through the expression (4.5) in Section 4.2.

Virtual Partitioning Policy

The basic definition of VP policy [BM98] is given for networks that handle only streaming traffic and when mobility is not considered. The VP policy protects SCs against overload in the system by giving, indirectly, lower priority to SCs with arrival rates higher than forecasts. At the time of design, each SC is allocated a nominal capacity C_i, where $\sum_{i=1}^{N} C_i \geq C$. The SCs which are using less than their nominal capacity are given higher priority. Conversely, the SCs that are exceeding their nominal capacity are given lower priority. The priority mechanism is implemented by a variant of the trunk reservation technique. Hence, while all the SCs are underloaded, the resources are shared indiscriminately, but when a SC is overloaded it is forced to back off if an underloaded SC needs its allocated resources. In fact, when the traffic is light VP policy behaves as the CS policy and when the system is overloaded VP policy tends to the CP policy.

The basic definition of VP policy takes the following decisions:

$$b(\boldsymbol{x}) + b_i \begin{cases} \leq C - t_i(x_i) & \text{accept} \\ > C - t_i(x_i) & \text{reject,} \end{cases} \tag{5.5}$$

where the parameter $t_i(x_i)$ of SC i may be interpreted as the parameter that introduces the trunk reservation mechanism, but in this case this parameter is dynamic and changes depending on the following rule:

$$t_i(x_i) = \begin{cases} s_i & \text{if} \quad b_i x_i + b_i \leq C_i \\ t_i & \text{if} \quad b_i x_i + b_i > C_i, \end{cases} \tag{5.6}$$

where $s_i \leq t_i$. Hence, $t_i(x_i)$ represents the resources that SC i cannot occupy when it is considered overloaded, and hence they determine the degree of isolation between SCs when the system is overloaded.

5.3 VP in Multiservice mobile cellular Networks

In this section the performance of the VP policy handling streaming and elastic traffic in a multiservice mobile cellular scenario is studied. Since mobility and multiservice networks are considered, the AC policy must take decisions distinguishing the different SCs and also the type of arrivals, new sessions or handovers. In order to simplify the design of the VP in multiservice mobile cellular networks, we first define a modified version of VP policy for streaming traffic (VPS) in Section 5.3.1. Next, a modified version of VP policy for elastic traffic (VPE) is defined in Section 5.3.2. Finally, we propose a new VP scheme (VPC) in Section 5.3.3 based on a combination of VPS, VPE and FGC [RTN97] policies for multiservice mobile cellular networks that support streaming and elastic SCs.

5.3.1 Streaming traffic: VPS policy

The objective of the VPS policy is to make a distinction among the different streaming SCs and not between the types of arrivals, i.e. new or handover arrivals. Therefore, only one type of arrivals is considered. Both types of arrivals are considered in the design of the VPC policy in Section 5.3.3.

At the time of design, each streaming SC is allocated a nominal capacity C_i. Streaming SCs which are using less than their C_i are given higher priority and otherwise, they are given lower priority. If $\boldsymbol{x}' = \boldsymbol{x} + \boldsymbol{e_i}$, where $\boldsymbol{e_i}$ denotes a vector with a 1 on the i-th position and 0's elsewhere, the VPS policy takes the following decisions when an arrival of streaming SC $i = 1, \cdots, N_s$ occurs:

$$\begin{cases} \text{accept} \quad \text{if} \quad \begin{cases} b(\boldsymbol{x}) + b_i \leq C - t_i(x_i) \\ \mu_j^{d,e}(\boldsymbol{x}') \geq \mu^m \qquad \text{for} \quad j = N_s + 1, \cdots, N \end{cases} \\ \\ \text{reject} \qquad \text{otherwise,} \end{cases} \tag{5.7}$$

where $b(\boldsymbol{x})$ is the number of resources occupied at state $\boldsymbol{x}$ by streaming traffic, b_i is the resources consumed by a request of the streaming SC i and $\mu_i^{d,e}(\boldsymbol{x})$ is the service rate for elastic SC i in state $\boldsymbol{x}$. The parameter $t_i(x_i)$ is defined as in the original VP policy in (5.6), and represents the resources that the streaming SC i cannot occupy when it is using more resources than its nominal capacity. Henceforth, to simplify the design we consider $s_i = 0 \ \forall i$ [BM98]. As it can be clearly seen, the VPS policy controls the priority of the streaming SC that wants to connect to the cell depending on its load. It also controls that after the acceptance of this streaming SC i the new service rates for all elastic SCs are still higher than the minimum required, otherwise the request is rejected.

Parameter design for Streaming Traffic

In order to design a VPS policy that provides a trade-off between efficiency and robustness against overloads, the nominal capacity C_i and the t_i parameters must be carefully chosen.

The nominal capacity, C_i, is the parameter which decides when the streaming SC i is overloaded. We define C_i as the minimum bandwidth that lets fulfill the QoS requirements of SC i in an isolated system, where only arrivals of SC i exist. From the nominal capacities, we consider that the total capacity of the system is $C = \sum_{i=1}^{N_s} C_i$.

The t_i parameters represent the resources that the SC i cannot occupy when it is overloaded. If the t_i parameters are high the VP policy tends to a CP policy and otherwise, the VP policy tends to a CS policy. In order to avoid a complex design of these parameters, which may entail high computational cost as it happens with the optimal parameter design of MFGC policy, an expression for the t_i parameters of VPS policy is chosen in a simple way by using heuristics. When studying the t_i parameter for streaming SC i, it is logical to think that when λ_j^n or b_j of the SCs $j \neq i$ are high, we need to reserve more resources for SCs j and therefore, the parameter t_i should be higher. When the CHT of the SC i is low, i.e. $\mu_i^s = \mu_i^{d,s} + \mu_i^{r,s}$ is high, sessions of SC i occupy resources during shorter times, therefore t_i should be lower for higher μ_i^s. Taking into account these facts and studying the blocking probabilities $P_i^{b,n}$ when some SCs are overloaded for different values of the system parameters, the proposed expression for the t_i parameter by using heuristics is:

$$t_i = \sqrt{\frac{3}{2}C}\frac{1}{C_i\mu_i^s}\sum_{j\neq i}\lambda_j^n b_j \qquad i = 1, \cdots, N_s. \tag{5.8}$$

The square root $\sqrt{C}$ appears as the economy of scale does not grow lineally with the total number of resources.

5.3.2 Elastic traffic: VPE policy

The objective of the VPE policy is to make a distinction among the different elastic SCs and not between the types of arrivals, i.e. new or handover arrivals. Therefore, only one type of arrivals is considered. Both types of arrivals are considered in the design of the VPC policy in Section 5.3.3.

At the time of design, each elastic SC is allocated a nominal number of flows n_i^m that will be detailed below. Elastic SCs which have less than their n_i^m flows in the system are given higher priority and otherwise, they are given lower priority. The VPE policy takes the following decision when an arrival

of elastic SC $i = N_s + 1, \cdots, N$ occurs:

$$\begin{cases} \text{accepted} & \text{if} \quad \begin{cases} b(\boldsymbol{x}) < C \\ \mu_j^{d,e}(\boldsymbol{x}') \geq \mu^m + t_i(x_i) \quad \text{for} \quad j = N_s + 1, \cdots, N \end{cases} \\ \text{rejected} & \text{otherwise,} \end{cases} \tag{5.9}$$

where the parameter $t_i(x_i)$ changes depending on the following rule:

$$t_i(x_i) = \begin{cases} s_i & \text{if} \quad x_i + 1 \leq n_i^m \\ t_i & \text{if} \quad x_i + 1 > n_i^m. \end{cases} \tag{5.10}$$

where $s_i \leq t_i$. As it can be clearly seen, the VPE policy controls the priority of the elastic SC that wants to connect to the cell depending on its load. It also controls if there are resources not occupied by streaming traffic.

Remember that n_i^m is the allocated nominal number of flows of elastic SC i. Thus, for elastic traffic, these parameters $t_i(x_i)$ represent the bandwidth that the elastic SC i cannot use when there are more flows of the SC i in the system than its nominal number of flows, n_i^m. Hence, they determine the degree of isolation between SCs when the system is overloaded. From now on, to simplify the design we consider $s_i = 0 \ \forall i$ [BM98]. Notice that for $i = 1, \cdots, N_s$, t_i is expressed in number of units of resources and for $i = N_s + 1, \cdots, N$, t_i is expressed in flows per second.

Parameter design for Elastic traffic

As it is made for streaming traffic, the nominal number of flows n_i^m and the t_i parameters for elastic traffic must also be carefully chosen. Remember that an elastic flow of the ith SC has a maximum bandwidth, r_i^M, and all the elastic flows require a minimum bandwidth, r^m. Remember also that f_i is the fraction which corresponds to SC i of the aggregated arrival rate, λ^T, defined in (4.1) in Section 4.2.

The nominal number of flows, n_i^m, is the parameter which decides when the elastic SC i is overloaded. We consider an isolated system where only arrivals of SC i exist and calculate the minimum bandwidth which lets fulfill the QoS requirements, i.e. $P_i^a \leq B_i^a$. Then, we define n_i^m as the maximum number of flows of the elastic SC i served at r_i^M in this isolated system.

The t_i parameters for elastic traffic correspond to the bandwidth that SC i can not access if it is overloaded. Again, an expression for the t_i parameters of VPE policy is chosen in a simple manner and low computational cost by heuristics. It is logical to think that for higher r_i^M, flows of SC i occupy resources during a shorter time and therefore, a lower t_i parameter is needed. Moreover, elastic SCs share all the available capacity and if the arrival rate of the elastic SC i is very high in relation with the others, ($f_i >> f_j$ where $j \neq i$), it will use more shared resources and hence, the parameter t_i should be higher to protect the other SCs. Taking into account these facts and studying the success completion probabilities P_i^c when some SCs are overloaded for different values of the system parameters, the proposed expression for t_i parameters by using heuristics is:

$$t_i = 2\frac{r^m}{L}\sqrt{\frac{C \cdot R \cdot f_i}{r_i^M}} \qquad i = N_s + 1, \cdots, N. \tag{5.11}$$

where R is the capacity in bits per second of one resource unit and L is the mean in bits of the flow size distribution.

5.3.3 VPC policy

In this section, we consider multiservice mobile cellular networks, which can handle streaming and elastic traffic and for each SC, new and handover arrivals are distinguished. The failure of a handover session is highly undesirable but reserving channels for handover traffic could increase blocking probabilities for new requests. Hence, for streaming traffic a trade-off between the two QoS measures is needed. It has to be decided whether to combine

new and handover sessions in a unique flow or not. As new and handover arrivals of the same SC have different QoS requirements ($B_i^h << B_i^n$), aggregating both types of arrivals into the same flow would be highly inefficient. But at the same time they cannot be managed as independent flows since λ_i^n and λ_i^h are related and undergo the same overloads. However, for elastic traffic, new and handover arrivals can be considered as an unique flow since the abandonment probability does not depend on whether the flow arrived at the system as new or handover request.

We propose the VPC policy, which is a combination of VPS, VPE and FGC [RTN97] policies in order to deal with the special characteristics of these networks. For streaming traffic, the different SCs are distinguished by using the VPS policy and new and handover arrivals by the FGC policy. For elastic traffic, only the different SCs are distinguished by using the VPE policy.

The FGC policy is the single service version of the MFGC policy defined in (4.6). Handover arrivals are always accepted and new session arrivals have an associated parameter $h_i \in \mathbb{R}$ that controls their acceptance through:

$$b_i x_i + b_i \begin{cases} \leq \lfloor h_i \rfloor & \text{accept} \\ = \lfloor h_i \rfloor + 1 & \text{accept with probability} \quad h_i - \lfloor h_i \rfloor \\ > \lfloor h_i \rfloor + 1 & \text{reject.} \end{cases} \tag{5.12}$$

where, remember x_i is the resources occupied by the ith streaming SC and b_i is the resources consumed by one streaming SC i.

At the time of design, each streaming SC is allocated a nominal capacity C_i and the parameter h_i, and each elastic SC a nominal number of flows n_i^m. The VPC policy takes different decisions depending on whether the arrival is either streaming or elastic traffic.

1. *Streaming traffic*

   The VPC policy is represented in Fig. 5.1 and works as follows: The handover arrivals of the SCs which do not fulfill the VPS restrictions are rejected; otherwise they are accepted.

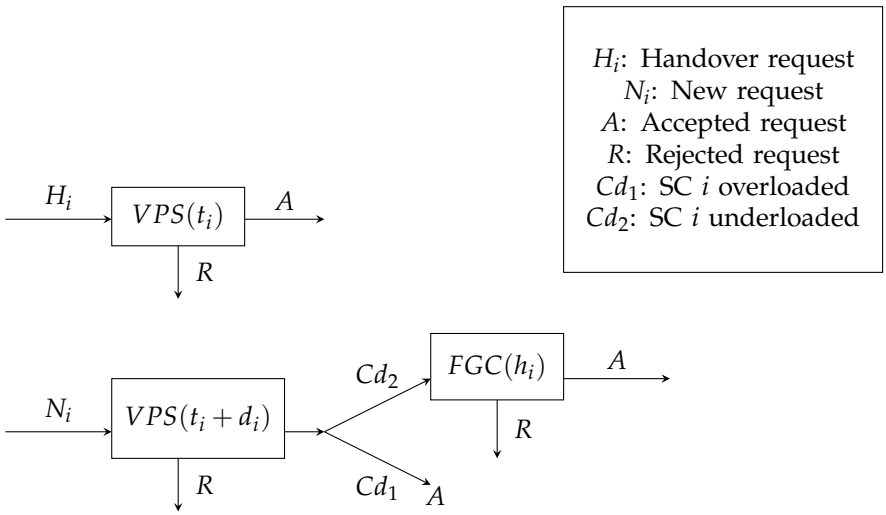

Figure 5.1: AC for multiservice mobile cellular networks for streaming traffic (VPC).

New arrivals of SCs which do not fulfill VPS restrictions are rejected. Note that for new arrivals we add the parameter $d_i = C_i - h_i$ to the VPS parameter t_i. If the new arrival passes VPS restrictions the system verifies the following conditions:

- Condition Cd_1: The SC i is overloaded and it is in the case:

$$b_i x_i + b_i > C_i,$$

$$b(\boldsymbol{x}) + b_i \leq C - (t_i + d_i).$$

Therefore, despite the fact that SC i is overloaded the new arrival of SC i is accepted because the overall traffic in the system is light. However, accepting all the new arrivals could be harmful for handover arrivals of the same SC and hence, d_i resources are reserved for handover arrivals of SC i.

- Condition Cd_2: The SC i is underloaded and it is in the case:

$$b_i x_i + b_i \leq C_i,$$

$$b(\boldsymbol{x}) + b_i \leq C.$$

Therefore, the FGC policy is applied with the corresponding parameter h_i to decide on the acceptance of the new arrival of SC i in order to protect the handover arrivals of SC i.

2. *Elastic traffic*

If the arrival is of an elastic SC, VPE is applied considering new and handover arrivals as a unique flow.

Parameter design

The nominal capacity of each streaming SC i, C_i, the FGC policy parameter of each streaming SC i, h_i, the nominal number of elastic flows of each elastic SC i, n_i^m, and the t_i parameters, must be carefully chosen.

For streaming traffic, the values of C_i and h_i are calculated considering that each streaming SC is isolated from the other SCs. Thus, we consider N_s single service scenarios with only one SC where new session and handover arrivals are possible. An AC policy is needed to provide the QoS requirements of new and handover arrivals. In this case, handover arrivals are always admitted and new arrivals are accepted or rejected depending on the FGC policy with parameter $h_i \in \mathbb{R}$. Given the system parameters, for each single service scenario we search the minimum value of C_i for which a value of h_i exists so that $P_i^{b,n} \leq B_i^n$ and $P_i^{b,h} \leq B_i^h$. Then, C_i is the minimum number of resources that fulfills objectives, and the chosen h_i are determined by adjusting the $P_i^{b,h}$ under B_i^h and the $P_i^{b,n}$ between B_i^n and 1% under B_i^n. The parameters t_i where $i = 1, \cdots, N_s$ have the same expression than in the VPS policy (5.8), but it is extended to mulitiservice mobile cellular networks

considering the handover arrivals:

$$t_i = \sqrt{\frac{3}{2}C}\frac{1}{C_i\mu_i^s}\sum_{j\neq i}(\lambda_j^n + \lambda_j^h)b_j \qquad i = 1, \cdots, N_s \tag{5.13}$$

For elastic traffic, the value of n_i^m is calculated by considering that each elastic SC is isolated from the other SCs. Thus, we consider N_e single service scenarios. In this case, the new session and handover arrivals have the same QoS requirements and hence, they have the same treatment in the system. We follow the same approach to calculate n_i^m as in the VPE policy in Section 5.3.2. The parameters t_i, where $i = N_s + 1, \cdots, N$, have the same expression as in the VPE policy defined in (5.11).

Once the parameters are determined, we can study the aggregated system which supports all streaming and elastic SCs and new sessions and handover arrivals. We consider a system with a total number of resources given by the sum of the nominal capacities obtained for streaming SCs, $C = \sum_{i=1}^{N_s} C_i$. Due to the economy of scale, the streaming SCs do not use all the resources all the time in order to fulfill the QoS requirements, and elastic SCs can use the resources that streaming SC are not using.

5.4 Numerical evaluation

In this section, the performance of the VPC policy is studied and compared with other policies in scenarios with overloaded traffic. The performance of the VPC policy for streaming traffic is compared with that of the MFGC policy and for elastic traffic is compared with that of the CS policy. In scenarios where streaming and elastic traffic exist, we compare the policy with a combination of the MFGC and the CS policies. That is, a request of streaming traffic is accepted if the requirements of the MFGC policy are fullfilled and if after accepting the streaming request, all elastic flows still receive a service rate higher than their minimum μ^m, otherwise it is rejected. The require-

ments of the MFGC policy are given by the parameters t_i^n for new sessions or t_i^h for handovers. A request of elastic traffic is accepted if after accepting the elastic flow, the service rates of all elastic flows are higher than the minimum, otherwise it is rejected.

The MFGC policy has been chosen for streaming traffic because it is commonly found in the literature and because of its flexibility. However, its design requires high precision and its computational cost can be prohibitive for some practical systems, (Chapter 4). On the contrary, the VPC policy that we propose is configured with low computational cost.

We evaluate the performance of the policies proposed in this section for different degrees of overload. We consider that both new and handover arrivals are overloaded in the same degree. If $\hat{\lambda}_i^n$ and λ_i^n are the overloaded arrival rate and the arrival rate defined by forecasts of SC i, respectively, the overload $\wedge$ is defined as the traffic that exceeds the expected traffic expressed as a percentage of the traffic forecasts, i.e.:

$$\wedge(\%) = \frac{\sum_i (\hat{\lambda}_i^n - \lambda_i^n)}{\sum_i \lambda_i^n} 100. \tag{5.14}$$

For the numerical examples we consider three different systems: i) System S, with only streaming SCs; ii) System E, with only elastic SCs; iii) System C, with both streaming and elastic SCs.

Table 5.1: Definition of system S parameters.

Parameter	Value	Parameter	Value
λ^T	30	$\boldsymbol{f}$	$[0.8, 0.2]$
$\boldsymbol{\mu^{r,s}}$	$[0.5, 0.5]$	$\boldsymbol{\mu^{d,s}}$	$[0.5, 1]$
$\boldsymbol{b}$	$[1, 2]$	$\boldsymbol{B^n}$	$[0.02, 0.01]$
$\boldsymbol{B^h}$	$[0.004, 0.002]$		

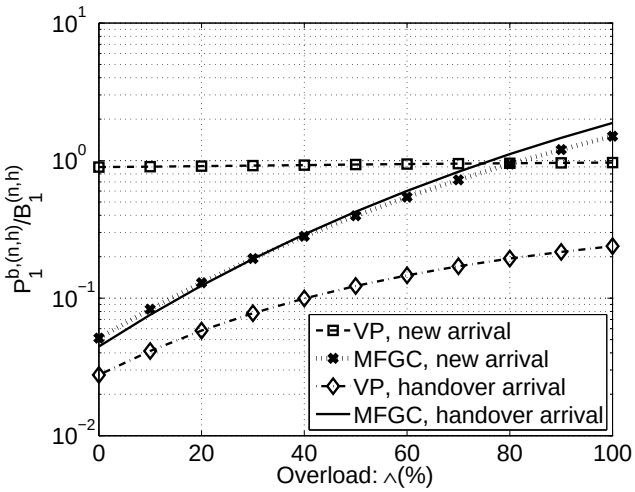

Figure 5.2: Ratios for streaming SC 1 as function of the overload undergone by SC 2 in system S.

The system parameters of system S with $N = N_s = 2$ are detailed in Table 5.1. The obtained nominal capacities are $C_1 = 62$ and $C_2 = 28$ and $C = 90$. The parameter setting of the VPC policy is: i) FGC parameters, $h_1 = 60.08$ and $h_2 = 25.46$; ii) VP parameters, $t_1 = 3.36$ and $t_2 = 13.12$. The optimal configuration of the MFGC policy for $C = 90$ is: $t^n = [86.46, 88.48]$ and $t^h = [88.46, 89.92]$ and the maximum offered traffic is $\lambda_{max}^T = 35.44$.

Results in Fig. 5.2 show the ratio of blocking probabilities to the objectives of SC 1, i.e. $P_1^{b,n}/B_1^{b,n}$ and $P_1^{b,h}/B_1^{b,h}$, when SC 2 has different degrees of overload. The ratios are calculated for both policies, the VPC and the MFGC. A ratio higher than 1 means that the QoS requirements are not fulfilled. For the VPC policy, the objectives are always fulfilled for these overloads. We can see that the blocking probabilities are lower than objectives. This is an effect of the economy of scale that appears when all the resources, dimensioned considering the nominal capacities of SC 1 and SC 2, are shared. For the MFGC policy, the maximum overload supported fulfilling objectives for new session arrivals is $\wedge = 82\%$ and for handover arrivals is $\wedge = 76\%$.

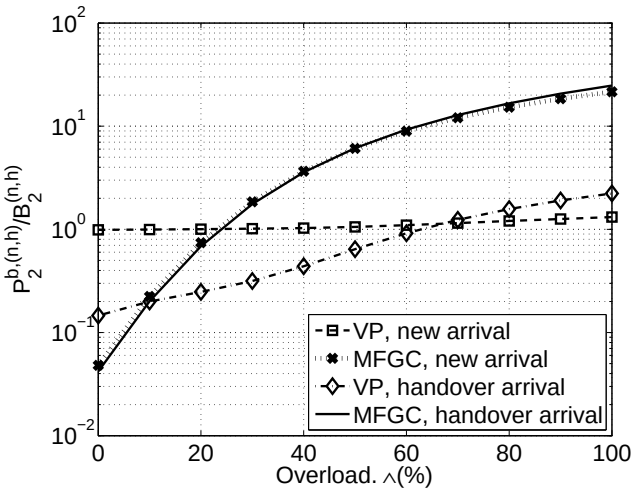

Figure 5.3: Ratios for streaming SC 2 as function of the overload undergone by SC 1 in system S.

Results in Fig. 5.3 show the same ratios for SC 2, i.e. $P_2^{b,n}/B_2^{b,n}$ and $P_2^{b,h}/B_2^{b,h}$, when SC 1 has different degrees of overload. The ratios are calculated for both policies, the VPC and the MFGC. The results show that the VPC policy achieves lower ratios than the MFGC policy. For VPC policy, the maximum overload supported fulfilling objectives for new arrivals is $\wedge = 48\%$ and for handovers arrivals is $\wedge = 63\%$. For the MFGC policy, the maximum overload supported fulfilling objectives for new arrivals is $\wedge = 23\%$ and for handovers arrivals is $\wedge = 24\%$. For overloads of $\wedge = 100\%$ and handover arrivals, which is the arrival type with the worst ratios, the achieved ratios for the MFGC policy are 11.12 times the achieved ratios for the VPC polciy.

The system paremeters of system E with $N = N_e = 2$ are detailed in Table 5.2. Since we do not consider streaming flows in this system, the total number of resources is calculated as the sum of the minimum number of resources needed to fulfill QoS requirements for each isolated system of the different SCs. Thus, when SC 1 is isolated 44 resources are needed to fulfill QoS requirements and when SC 2 is isolated we need 45 resources. The

Table 5.2: Definition of the system E parameters.

Parameter	Value	Parameter	Value
λ^T	5	$\boldsymbol{f}$	$[0.8, 0.2]$
$\boldsymbol{\lambda^h}$	$0.5\boldsymbol{\lambda^n}$	$\boldsymbol{\mu^{r,e}}$	$[0.25, 0.25]$
R	100 kbps	r^m	200 kbps
$\boldsymbol{r^M}$	$[1000, 1500]$ kbps	L	100 kbits
$\boldsymbol{\beta^0}$	$[0, 0]$	$\boldsymbol{\beta^1}$	$[1, 1]$
$\boldsymbol{B^c}$	$[0.99, 0.999]$		

total number of resources is then, $C = 44 + 45 = 89$. The obtained nominal number of flows are $n_1^m = 4$ and $n_2^m = 3$. The parameter setting of the VPC policy is: $t_1 = 9.05$ and $t_2 = 3.70$.

Results in Fig. 5.4 show the ratio of the lower-bounds of success completion probabilities of SCs 1 to the achieved success completion probabilities (B_1^c / P_1^c) when SCs 2 has different degrees of overload, and the same ratio for SCs 2 (B_2^c / P_2^c) when SCs 1 has different degrees of overload. The ratios are calculated for both policies, the VPC and the CS. The results show that the QoS requirements are not fulfilled for very high overloads. The gain of using VPC policy instead of the CS policy is noticeable for overloads higher than $\wedge = 500\%$. For SC 2, which has the worst ratio, the ratio for CS policy is 1.57 times higher than the ratio for the VPC policy when the overload is $\wedge = 1900\%$, i.e. the $\hat{\lambda}_1^n = 20\lambda_1^n$. We can conclude that in a system with only elastic SCs, the overload is not a critical feature.

The system paremeters of system C with $N = 3$ SCs, where 2 SCs carry streaming traffic, $N_s = 2$, and 1 SC carries elastic traffic, $N_e = 1$, are detailed in Table 5.3. The obtained nominal capacities are $C_1 = 10$ and $C_2 = 10$, therefore $C = 20$, and the nominal number of flows is $n_3^m = 1$. The parameter setting of the VPC policy is: i) FGC parameters, $h_1 = 8.26$ and $h_2 = 9.19$; ii) VP parameters, $t_1 = 1.47$, $t_2 = 1.30$ and $t_3 = 2.53$. The optimal parameter setting of the MFGC policy for a total capacity $C = 20$ is: $t^n = [15.53, 17.40]$ and $t^h = [17.23, 18.97]$ and the maximum offered traffic is $\lambda_{max}^T = 5.44$.

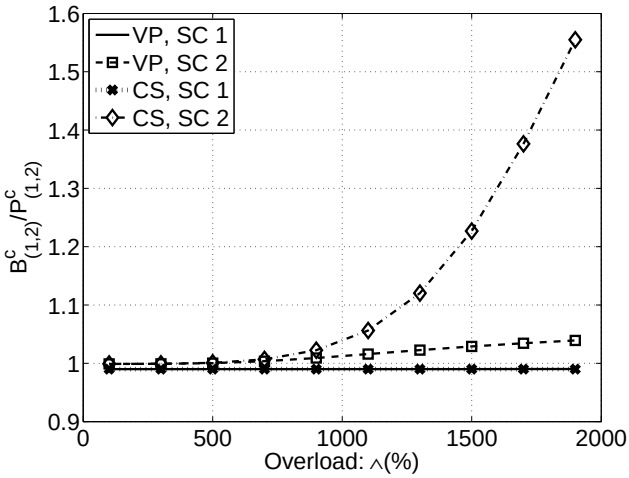

Figure 5.4: Ratios for elastic SCs as function of the overload in system E.

Table 5.3: Definition of the system C parameters.

Parameter	Value	Parameter	Value
λ^T	3	$\boldsymbol{f}$	$[0.6, 0.3, 0.1]$
$\boldsymbol{b}$	$[1, 2]$	$\boldsymbol{\mu^{d,s}}$	$[0.5, 1]$
$\boldsymbol{\mu^{r,s}}$	$[0.5, 0.5]$	$\mu^{r,e}$	0.25
R	100 kbps	r^m	200 kbps
r_3^M	500 kbps	L	100 kbits
β_3^0	0	β_3^1	1
$\boldsymbol{B^n}$	$[0.02, 0.01]$	$\boldsymbol{B^h}$	$[0.004, 0.002]$
B_3^c	0.99		

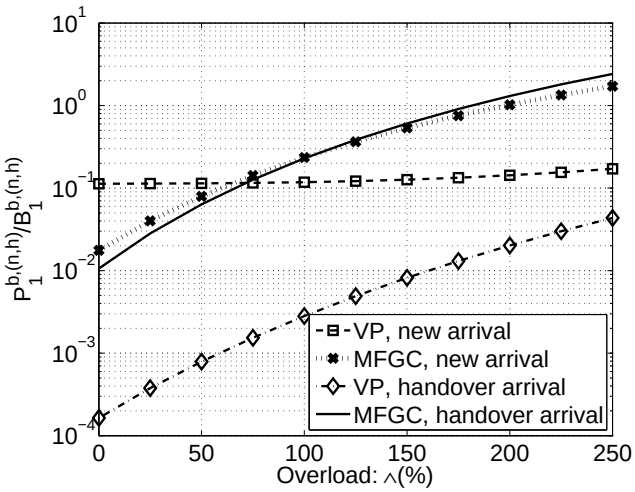

Figure 5.5: Ratios for streaming SC 1 as function of the overload undergone by SC 2 and 3 in system C.

Results in Fig. 5.5 show the ratio of blocking probabilities of SC 1 to the objectives, $P_1^{b,n}/B_1^{b,n}$ and $P_1^{b,h}/B_1^{b,h}$, when SCs 2 and 3 support the same overload $\wedge(\%)$. The ratios are calculated for both policies, VPC and MFGC. For the VPC policy, the objectives are always fulfilled for these overloads. We can see again that the blocking probabilities are lower than objectives because of the economy of scale. For the MFGC policy, the maximum overload supported fulfilling objectives for new arrivals is $\wedge = 198\%$ and for handovers arrivals is $\wedge = 182\%$. For overloads of $\wedge = 250\%$ and new arrivals, the ratio for the MFGC policy is 10 times the ratio for VPC.

Results in Fig. 5.6 show the ratio of blocking probabilities of SC 2 to the objectives, $P_2^{b,n}/B_2^{b,n}$ and $P_2^{b,h}/B_2^{b,h}$, when SCs 1 and 3 are overloaded with the same overload $\wedge(\%)$. The results show that the VPC policy achieves lower ratios than MFGC policy. For the VPC policy, the QoS requirements for new arrivals are fulfilled for these overloads and the maximum overload supported fulfilling objectives for handover arrivals is $\wedge = 235\%$. For the MFGC policy, the maximum overload supported fulfilling objectives for new

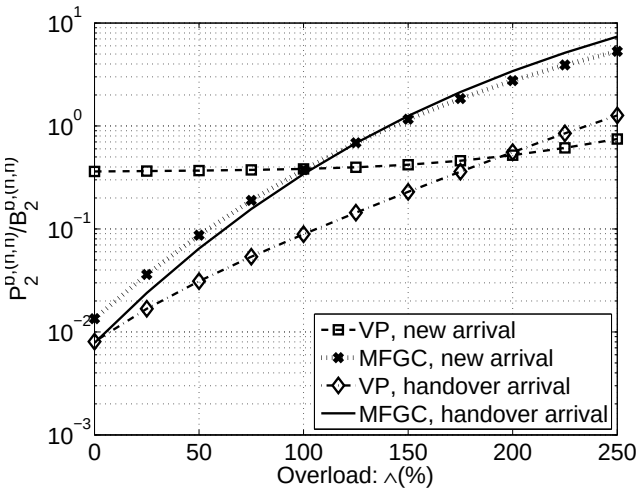

Figure 5.6: Ratios for streaming SC 2 as function of the overload undergone by SC 1 and 3 in system C.

arrivals is $\wedge = 143\%$ and for handover arrivals is $\wedge = 140\%$. For overloads of $\wedge = 250\%$ and handover arrivals, the ratio for the MFGC policy is 5.83 times the ratio for the VPC policy.

Finally, results in Fig. 5.7 show the ratio of the lower-bound of success completion probabilities of SC 3 to the achieved success completion probabilities, B_3^c/P_3^c, when SCs 1 and 2 are overloaded. The ratios are calculated for both policies, VPC and MFGC. The VPC policy can support overloads of $\wedge = 172\%$ fulfilling objectives, while the MFGC policy can support $\wedge = 122\%$ of overload. However, the ratios achieved for both policies are lower than for streaming traffic, showing that overloads are not as critical as for streaming traffic since elastic SCs can adapt their rate to the system load variability.

These figures confirm that the VPC policy is more robust than the MFGC policy when some SCs exceed the expected offered traffic. Observing the results obtained for streaming traffic in systems S and C, we can see that system S, which has higher aggregated arrival rates than system C, supports

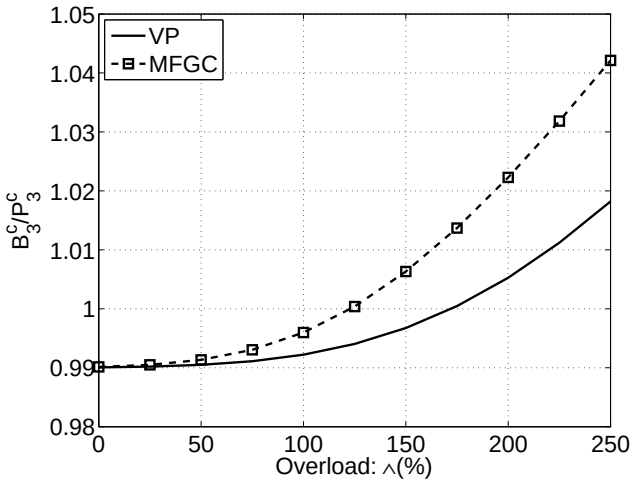

Figure 5.7: Ratios for elastic SC 3 as function of the overload undergone by SC 1 and 2 in system C.

lower degrees of overloads. Thus, as we can expect, the AC policy is more critical in systems with high aggregated arrival rates. The results obtained for elastic traffic in systems E and C show that the AC policy is less critical for elastic traffic than for streaming traffic. However, by implementing an AC policy for elastic traffic avoids wasting capacity by non-completed sessions due to impatience when the elastic flows compete with other elastic flows (system E) or with streaming sessions (system C).

5.5 Conclusions

In this chapter we have studied a new design method for the VP policy which integrates streaming and elastic traffic and considers mobility. The performance of the proposed policy, called VPC policy, has been studied and compared to that of the MFGC policy under overloaded scenarios in order to evaluate the robustness of both policies.

The results show that the VPC policy is more robust than the MFGC policy under overload conditions since the VPC policy protects better a given SC against overloads of the other SCs and at the same time all SCs fulfill objectives under normal conditions. Moreover, the design of the parameters of the VPC policy has lower computational cost than the optimal MFGC policy design. We have observed that for streaming traffic, the AC policy is more critical in systems with high aggregated arrival rates and therefore, the advantage of using the VPC policy instead of the MFGC policy is more noticeable. For elastic traffic the AC policy is not as critical as for streaming traffic when the system is overloaded since elastic SCs can adapt their rate to the variability of the system load. However, an AC policy avoids wasting bandwidth because of abandonments and hence, the VPC policy achieves better success completion probabilities than the CS policy when the system is overloaded, which is more noticeable for high overloads.

Chapter 6

Reversibility and AC policies

6.1 Introduction

As it has been pointed out in previous chapters, for streaming traffic in cellular mobile networks two important QoS measures are the fraction of new session and handover arrivals that are blocked due to the lack of enough free resources. As handover blocking is more annoying than new session blocking for subscribers, efficient AC strategies can be used to reject new sessions in order to reserve resources for future handovers, while minimizing the impact on the blocking rate of new sessions. As also pointed out in previous sections, conventional trunk reservation AC policies lead to CTMC whose state-space cardinality grows very quickly with the number of channels and SCs supported. Then, determining the stationary distribution and parameters derived from it, like new and handover probabilities, might become an unfeasible task.

Besides the efficiency or the computational cost necessary to analyze the CTMC obtained from the AC policy, an important property when studying AC policies is the reversibility of the CTMC which models the system. If the CTMC is reversible, the stationary distribution is insensitive to the

CHT, in the sense that it depends on the CHT distribution through the mean only [Bon06].

In this chapter, we propose a probabilistic AC policy for multiservice mobile cellular networks which supports different SCs and provides differentiated treatment to each arrival type (new or handover). The CTMC that models the system is reversible and its stationary distribution has a product-form, which greatly simplifies its computation. In addition, the CTMC obtained using the AC policy proposed is insensitive to the CHT distribution. On the contrary, trunk reservation policies do not lead to reversible and insensitive CTMC unless some restrictions are imposed. An interesting feature of the proposed policy is that the resource sharing among SCs, and between new and handover calls of the same SC, can be controlled independently. This study has been motivated in part by the study presented in [STKC09], although we obtain results different from the ones derived there. This work resulted into the publication in [MBPBM11].

In the next section we present and prove the reversibility and insensitivity properties of the proposed policy. In Section 6.3 we present examples of AC policies that lead to both reversible and non-reversible CTMC. In Section 6.2 we study the insensitivity property of the proposed policy, the MFGC policy and the VPC policy. Finally, Section 6.5 concludes the chapter.

6.2 Reversible and insensitive AC policy

We consider a cellular network with C resource units that supports N SCs. Only streaming SCs are considered. Since new and handover requests are distinguished, the system handles $2N$ arrival types. The new and handover arrivals of the ith SC occur according to a Poisson process with rates λ_i^n and λ_i^h, respectively. We assume that the CHT is exponentially distributed, with rates μ_i^n and μ_i^h for new and handover arrivals of the ith SC. As shown below, this assumption on the CHT has no impact on the results.

If b_i is the number of resource units required to set up a session of the ith

SC, the maximum number of ongoing sessions of the ith SC (either initiated as new or handover) in the system is given by $M_i = \lfloor C/b_i \rfloor$. Let x_i^n and x_i^h be the number of ongoing calls of the ith SC, $1 \leq i \leq N$, initiated as new or handover requests, respectively. The system state is described by the N-tuple $\boldsymbol{x} = \left(x_1^n, \ldots, x_N^n, x_1^h, \ldots, x_N^h\right)$.

For the ith SC, let us define the vectors of probabilities $\boldsymbol{c_i}$ and $\boldsymbol{d_i}$ as:

$$\begin{aligned} \boldsymbol{c_i} &= \left[c_i\left(0\right), c_i\left(1\right), \ldots, c_i\left(M_i - 1\right), 0\right] , \\ \boldsymbol{d_i} &= \left[d_i\left(0\right), d_i\left(1\right), \ldots, d_i\left(M_i - 1\right), 0\right] , \end{aligned} \tag{6.1}$$

where $0 \leq c_i\left(m\right), d_i\left(m\right) \leq 1$, and $0 \leq m \leq \left(M_i - 1\right)$.

At state $\boldsymbol{x}$, an arrival of the ith SC will be accepted depending on the following AC policy:

- *New* arrival $\rightarrow$ Accepted with probability $a_i^n(\boldsymbol{x}) = c_i(x_i^n)d_i(x_i^n + x_i^h)$.
- *Handover* arrival $\rightarrow$ Accepted with probability $a_i^h(\boldsymbol{x}) = d_i\left(x_i^n + x_i^h\right)$.

Note that the resource sharing between SCs can be controlled by configuring $\boldsymbol{d_i}$, while the resource sharing between arrival types of the same SCs by configuring $\boldsymbol{c_i}$.

The number of resource units occupied in state $\boldsymbol{x}$, $b(\boldsymbol{x})$, is given by:

$$b(\boldsymbol{x}) = \sum_{i=1}^{N} \left(x_i^n + x_i^h\right) b_i. \tag{6.2}$$

Then, the system can be modeled as a reversible CTMC with state space

$$\mathcal{S} := \left\{\boldsymbol{x} : x_i^n, x_i^h \in \mathbb{N};\ b(\boldsymbol{x}) \leq C\right\}. \tag{6.3}$$

From now on, we refer to it as CTMC $\{\boldsymbol{x}(t)\}_{t\geq 0}$. We prove its reversibility by showing that the so called arrival and service processes of an equivalent queuing network are reversible [Bon06].

Consider a queuing network with $2N$ nodes, no waiting facilities (i.e. we consider a loss network) and no internal routing, where new arrivals of the ith SC are offered to node i and handover arrivals of the ith SC are offered to node $i+N$. Assume Poisson arrivals from outside the network with rates $\lambda_i = \lambda_i^n$ for new arrivals and $\lambda_{i+N} = \lambda_i^h$ for handover arrivals, and exponentially distributed services with rates $\mu_i = \mu_i^n$ and $\mu_{i+N} = \mu_i^h$. Let $\boldsymbol{x}' = (x'_1, \ldots, x'_{2N})$ be the vector whose jth component gives the number of ongoing sessions at node j, $1 \leq j \leq 2N$. In state $\boldsymbol{x}'$, an arrival to node j is accepted with a probability given by:

$$a_j(\boldsymbol{x}') = \begin{cases} c_j(x'_j) d_j(x'_j + x'_{j+N}) & \text{if} \quad 1 \leq j \leq N \\ d_{j-N}(x'_{j-N} + x'_j) & \text{if} \quad N+1 \leq j \leq 2N. \end{cases} \tag{6.4}$$

In addition, admission decisions are subject to the capacity constraint given by the condition $b(\boldsymbol{x}') \leq C$.

We consider that after service completion at node j in state $\boldsymbol{x}'$, a session is routed to node k with probability $p_{jk}(\boldsymbol{x}') = 0$ (i.e., there is no internal routing), and leaves the network with probability $p_j(\boldsymbol{x}') = 1$. Additionally, $\gamma_j(\boldsymbol{x}')$ is the effective arrival rate to node j in state $\boldsymbol{x}'$, which takes into account the impact of the admission policy, and it is given by:

$$\gamma_j(\boldsymbol{x}') = a_j(\boldsymbol{x}') \lambda_j. \tag{6.5}$$

Then, the transition rates for the CTMC $\{\boldsymbol{x}'(t)\}_{t \geq 0}$ are given by:

$$\begin{cases} q(\boldsymbol{x}', \boldsymbol{x}' + \boldsymbol{e_j}) = \gamma_j(\boldsymbol{x}'), \quad \text{if} \quad b(\boldsymbol{x}' + \boldsymbol{e_j}) \leq C \\ q(\boldsymbol{x}', \boldsymbol{x}' - \boldsymbol{e_j}) = \mu_j(\boldsymbol{x}') p_j(\boldsymbol{x}') = x'_j \mu_j \\ q(\boldsymbol{x}', \boldsymbol{x}' - \boldsymbol{e_j} + \boldsymbol{e_k}) = x'_j \mu_j p_{jk}(\boldsymbol{x}') = 0, \quad k \neq j \end{cases} \tag{6.6}$$

where $\boldsymbol{e_i}$ is a $2N$-dimensional vector with component i set to 1 and 0 elsewhere. The CTMC $\{\boldsymbol{x}'(t)\}_{t \geq 0}$ that describes the dynamics of the queuing network is the same as the CTMC $\{\boldsymbol{x}(t)\}_{t \geq 0}$ which describes the multiservice

mobile cellular system under study.

For the considered queuing network, if there is a positive function Φ that satisfies

$$\Phi(\boldsymbol{x}') = \Phi(\boldsymbol{x}' + \boldsymbol{e_j})\mu_j(\boldsymbol{x}' + \boldsymbol{e_j}) = \Phi(\boldsymbol{x}' + \boldsymbol{e_j})(x'_j + 1)\mu_j \,, \tag{6.7}$$

$\forall j$, $1 \leq j \leq 2N$, and $\forall \boldsymbol{x}' \in \mathcal{S}$, then the service process is reversible [Bon06]. Condition (6.7) is met by the function:

$$\Phi(\boldsymbol{x}') = \prod_{j=1}^{2N} \frac{1}{x'_j!\ \mu_j^{x'_j}}. \tag{6.8}$$

Likewise, if there is a positive function Λ that satisfies

$$\Lambda(\boldsymbol{x}')\gamma_j(\boldsymbol{x}') = \Lambda(\boldsymbol{x}' + \boldsymbol{e_j}), \tag{6.9}$$

$\forall j$, $1 \leq j \leq 2N$, and $\forall \boldsymbol{x}' \in \mathcal{S}$, then the arrival process is reversible [Bon06]. Condition (6.9) is met by the function

$$\Lambda(\boldsymbol{x}') = \prod_{j=1}^{2N} \lambda_j^{x'_j} \prod_{i=1}^{N} \alpha_i(x'_i)\beta_i(x'_i + x'_{i+N}), \tag{6.10}$$

where

$$\alpha_i(u) = \prod_{k=0}^{u-1} c_i(k),$$

$$\beta_i(u) = \prod_{k=0}^{u-1} d_i(k).$$

Thus, the stationary distribution of the CTMC $\{\boldsymbol{x}'(t)\}_{t\geq 0}$ that describes the dynamics of the considered queuing network becomes

$$\pi(\boldsymbol{x}') = \pi(\boldsymbol{0})\Lambda(\boldsymbol{x}')\Phi(\boldsymbol{x}'), \qquad \boldsymbol{x}' \in \mathcal{S}\backslash\{\boldsymbol{0}\}, \tag{6.11}$$

where $\pi(\mathbf{0})$ is obtained by normalization [Bon06].

Equivalently, we obtain the stationary distribution $\pi(\boldsymbol{x})$ of the original CTMC $\{\boldsymbol{x}(t)\}_{t\geq 0}$ as:

$$\pi(\boldsymbol{x}) = \pi(\mathbf{0}) \prod_{i=1}^{N} \prod_{r=0}^{x_i-1} d_i(r) \prod_{s=0}^{x_i^n-1} c_i(s) \frac{(\rho_i^n)^{x_i^n}}{x_i^n!} \frac{(\rho_i^h)^{x_i^h}}{x_i^h!}, \tag{6.12}$$

where

$$\rho_i^n = \frac{\lambda_i^n}{\mu_i^n}, \qquad \rho_i^h = \frac{\lambda_i^h}{\mu_i^h} \qquad \text{and} \qquad x_i = x_i^n + x_i^h.$$

Then, the blocking probabilities can be determined by

$$\begin{aligned} P_i^{b,n} &= 1 - \sum_{x\in\mathcal{S}} c_i(x_i^n) d_i(x_i^n + x_i^h) \pi(\boldsymbol{x}), \\ P_i^{b,h} &= 1 - \sum_{x\in\mathcal{S}} d_i(x_i^n + x_i^h) \pi(\boldsymbol{x}) , \end{aligned} \tag{6.13}$$

where $c_i(M_i) = d_i(M_i) = 0$ as defined in (6.1).

When both the arrival and service processes are reversible, then the queuing network process $\{\boldsymbol{x}'(t)\}_{t\geq 0}$, and therefore $\{\boldsymbol{x}(t)\}_{t\geq 0}$, are also reversible. In addition, their stationary distributions are *insensitive*, in the sense that they depend on the session duration distribution at each node through the mean only. In other words, when arrivals follow Poisson processes, all key performance indicators obtained from the stationary distribution, like blocking probabilities, are independent from all traffic characteristics beyond the traffic intensity [Bon06].

6.3 Reversibility of trunk reservation policies

The CTMC $\{\boldsymbol{x}'(t)\}_{t\geq 0}$ defined in Section 6.2, which models the queuing network, is reversible if the Kolgomorov criterion is met for all possible loops of the transition diagram [Kol36, Nel95]. From the loop shown in Fig. 6.1, the

following condition is obtained:

$$\frac{a_i(\boldsymbol{x}' + \boldsymbol{e}_j)}{a_i(\boldsymbol{x}')} = \frac{a_j(\boldsymbol{x}' + \boldsymbol{e}_i)}{a_j(\boldsymbol{x}')} \,. \tag{6.14}$$

As explained in Section 2.3, the multiple guard channel policies in mobile networks are threshold type policies inside of the family of trunk reservation policies. For them, the probabilities $a_j(\boldsymbol{x}')$ are given by:

$$a_j(\boldsymbol{x}') = \begin{cases} 1 & \text{if} \quad b(\boldsymbol{x}' + \boldsymbol{e}_j) \leq t_j \\ 0 & \text{otherwise,} \end{cases} \tag{6.15}$$

The parameter t_j is the maximum number of channels that type j arrivals have access to. Then, type-j arrivals see a system limited to t_j channels and are accepted depending on the occupation at arrival time. It is not difficult to show that condition (6.14) requires that $t_j = t \; \forall j$ where $1 \leq j \leq 2N$, i.e., all SCs share the same threshold. As a consequence, no differentiated treatment can be provided, neither among SCs, nor between new and handover arrivals of the same SC. In fact, the policy degenerates into a CS policy.

If a trunk reservation policy is used, full bidirectional connectivity between adjacent states of the CTMC may be lost and therefore the detailed balance equations would not hold. As detailed balance is a necessary condi-

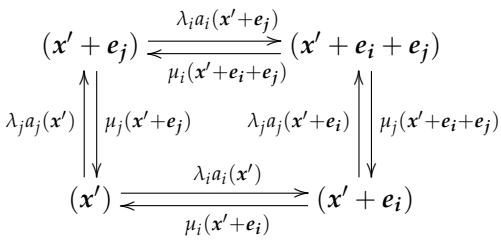

Figure 6.1: Transition diagram loop of a queuing network.

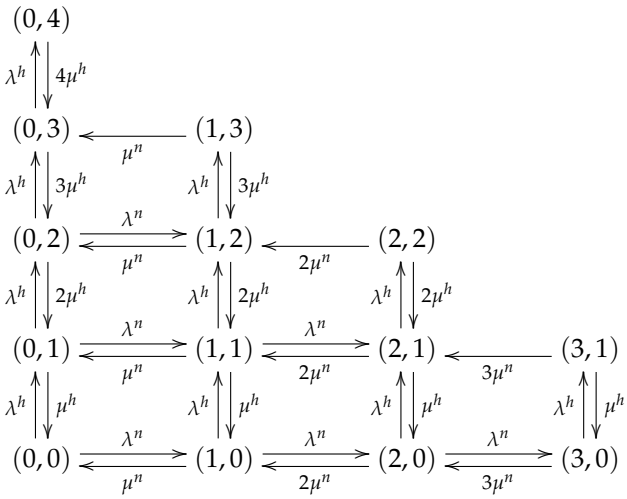

Figure 6.2: Transition diagram of a single service scenario.

tion for reversibility [Nel95], then the CTMC would not be reversible. As a consequence, the CTMC that models the multiservice system, which is constructed with processes that are not reversible, yields a non-reversible process [Nel95]. For illustrative purposes, Figure 6.2 shows the state diagram of the CTMC modeling a one-cell system enforcing a trunk reservation policy. The system parameters are: $N = 1$, $C = 4$ and $b_1 = 1$. Then, a new arrival is accepted if $x_1^n + x_1^h < 3$, and rejected otherwise. While handover arrivals are always accepted if free resources are available. Note that bidirectional connectivity is lost for the adjacent states $(0,3) \rightleftharpoons (1,3)$, $(1,2) \rightleftharpoons (2,2)$ and $(2,1) \rightleftharpoons (3,1)$.

Thus, trunk reservation policies do not lead to reversible CTMCs unless further restrictions are imposed. As an example, the *Thinning Scheme I* proposed in [Fan03], which includes the guard channel and the fractional guard channel schemes as special cases, requires that $\forall i,k$, $1 \leq i,k \leq N$, $b_i = 1$ and $\mu_i^n = \mu_i^h = \mu_k^n = \mu_k^h$. These conditions make the multidimensional CTMC to degenerate into a one dimensional birth and death process, which is known

to be reversible and for which a product-form solution exists.

Although classical trunk reservation policies are not reversible, we can obtain reversible policies based on them. Under trunk reservation policies, the acceptance of an arrival in the system depends on the total system occupation, i.e. the coefficients $a_j(x')$ depend on $\boldsymbol{x}'$. However, if the acceptance of an arrival in the system is a function of the total number of active sessions, $b(\boldsymbol{x}')$ given by:

$$b(\boldsymbol{x}') = \sum_{i=1}^{N} (x'_i + x'_{i+N}),$$

the coefficients $a_j(b(\boldsymbol{x}'))$ depend on $b(\boldsymbol{x}')$, and then, a new family of reversible policies can be obtained.

Let us define

$$\delta(m) = \frac{a_j(m)}{a_j(m-1)}$$

$$\varphi_j = a_j(0)$$

$$M = b(\boldsymbol{x}').$$

Then, the arrival rate to the jth node in state $\boldsymbol{x}'$ can be defined as:

$$\gamma_j(\boldsymbol{x}') = \lambda_j a_j(M) = \lambda_j \varphi_j \prod_{m=1}^{M} \delta(m). \tag{6.16}$$

The conditions (6.8) and (6.9) are met, respectivelly, by the functions:

$$\Phi(\boldsymbol{x}') = \prod_{j=1}^{2N} \frac{1}{(x'_j!\ \mu_j^{x'_j})} \tag{6.17}$$

$$\Lambda(\boldsymbol{x}') = \prod_{j=1}^{2N} (\lambda_j \varphi_j)^{x'_j} \prod_{m=1}^{M} \delta(m)^{M-m}. \tag{6.18}$$

Therefore, the CTMC that models the queuing network being considered, and therefore the associated multiservice mobile cellular network, is

reversible and its stationary distribution

$$\pi(\boldsymbol{x}') = \pi(\mathbf{0}) \prod_{j=1}^{2N} \frac{(\rho_j \varphi_j)^{x'_j}}{x'_j!} \prod_{m=1}^{M} \delta(m)^{M-m} , \tag{6.19}$$

where $\rho_j = \lambda_j / \mu_j$, is insensitive to the CHT distribution.

6.4 Numerical evaluation

In this section, for illustrative purposes, we study the performance of the AC policy defined in Section 6.2. We compare the blocking probabilities of the different arrival types obtained by analytical results with results obtained by simulation when several CHT distributions are considered.

Remember that the admission probabilities are given by:

$$\begin{aligned} a_i^n(\boldsymbol{x}) &= c_i(x_i^n) d_i(x_i^n + x_i^h), \\ a_i^h(\boldsymbol{x}) &= d_i\left(x_i^n + x_i^h\right). \end{aligned}$$

We now define $\boldsymbol{c_i}$ and $\boldsymbol{d_i}$ by:

$$\begin{aligned} c_i(k) &= \begin{cases} A_i^d & \text{if } 0 \le kb_i < K_i , \\ A_i^u & \text{if } K_j \le kb_i < M_i b_i , \end{cases} \\ d_i(k) &= \begin{cases} D_i^d & \text{if } 0 \le kb_i < C_i , \\ D_i^u & \text{if } C_i \le kb_i < M_i b_i . \end{cases} \end{aligned} \tag{6.20}$$

Note that when $A_i^d = D_i^d = 1$ and $A_i^u = D_i^u = 0$, the resource sharing between the SCs can be controlled by configuring $\{C_i\}$, while the resource sharing between arrival types of the same SCs by configuring $\{K_i\}$. This policy is a subclass of the one defined in Section 6.2, and therefore all previous results apply.

Table 6.1: Definition of system parameters.

Parameter	Value	Parameter	Value
N	2	C	30
K_1	7	K_2	6
C_1	20	C_2	30
$\boldsymbol{\lambda^n}$	$[1/20, 1/50]$	$\boldsymbol{\lambda^h}$	$[1/25, 1/55]$
$\boldsymbol{\mu^n}$	$[1/100, 1/5]$	$\boldsymbol{\mu^h}$	$[1/300, \mu_2^h]$
$\boldsymbol{b}$	$[1, 6]$		

Then, if we consider $A_i^d = D_i^d = 1$ and $A_i^u = D_i^u = 0$, the AC policy takes the following decision:

- A new arrival of ith SC in state $\boldsymbol{x}$ is accepted with probability $A_i^d D_i^d = 1$, if $x_i^n b_i < K_i$ and $(x_i^n + x_i^h) b_i < C_i$, and rejected otherwise.
- A handover arrival of ith SC in state $\boldsymbol{x}$ is accepted with probability $D_i^d = 1$, if $(x_i^n + x_i^h) b_i < C_i$, and rejected otherwise.

The system that we study is defined by the parameters in Table 6.1. The parameter μ_2^h is chosen to achieve that the traffic offered by handover arrivals of the SC 2, ρ_2^h, is within the interval $0.1 \leq \rho_2^h \leq 5.0$, where ρ_2^h is given by: $\rho_2^h = \lambda_2^h / \mu_2^h$.

For the proposed AC policy, we compare the blocking probabilities of the different arrival types obtained by equation (6.13), where the distribution in (6.12) is used, with those obtained by simulation when the session duration is modeled by other distributions than the exponential distribution, such as Erlang, hyper-exponential, lognormal and bounded Pareto distributions defined in Appendix C.1. For more information about the simulation environment see Appendix D.2.

The parameters of these distributions are adjusted to achieve the same mean as for the analytical model. In the case of the hyper-exponential the

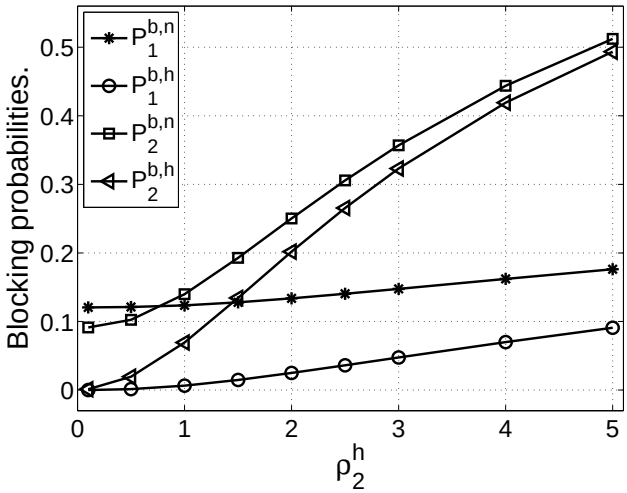

Figure 6.3: Hyper-exponentially distributed session duration with $CV = 4$.

CV is set to be $CV = 4$, for the Erlang distribution $CV = 0.25$, for the log-normal distribution $CV = 1$ and for the bounded Pareto distribution we set the shape factor to $k = 2.001$, the maximal value to $H = 10^5$ and adjusted the minimal value (L) accordingly to achieve the desired mean, obtaining CVs in the interval $[1.51, 2.33]$.

We obtain that the values of the blocking probabilities calculated with the analytical model are inside of the confidence intervals obtained from the simulation results for a level of confidence of 99%. The relative error, defined as the radius of the confidence interval divided by the blocking probability value, is lower than 5.5% for the hyper-exponential distribution, and lower relative errors, below 3%, are obtained for the other distributions. The results confirm the insensitivity property and the correctness of the stationary distribution of the CTMC defined in (6.12).

In Fig. 6.3, continuous line curves are obtained using the analytical model while simulation results are represented only by markers at the evaluated

Table 6.2: MFGC and VPC relative errors for different distributions (%).

	MFGC				**VPC**			
	Hyp	Erl	Log	Par	Hyp	Erl	Log	Par
$P_1^{b,n}$	3.54	2.43	38.77	39.11	1.39	0.70	11.32	11.88
$P_2^{b,n}$	1.06	1.01	27.88	29.25	0.94	0.62	34.17	35.51
$P_1^{b,h}$	4.60	5.90	38.32	38.53	7.89	9.37	28.57	27.08
$P_2^{b,h}$	0.73	1.10	31.25	32.93	0.84	0.63	41.25	42.54

points. The simulation results are obtained considering the hyper-exponential distribution with $CV = 4$ for the session duration distribution. We can see that simulation results overlap the analytical results. As expected the proposed AC policy protects more the handover arrivals, which have lower blocking probabilities. We can also see that SC 2 has higher blocking probabilities since one session of SC 2 uses a high number of resources $b_2 = 6$.

In order to show that the trunk reservation policies do not lead to insensitive CTMCs unless further restrictions are imposed, we compare the results obtained analytically when MFGC and VPC are applied and the session duration is exponentially distributed, with those obtained by simulation with the distributions and CV defined before. We obtain that the blocking probabilities calculated with the analytical model are not inside the confidence intervals of the simulation results for a level of confidence of 99%. In this case, we define the relative errors as the difference between the analytical and the mean of the simulation results divided by the mean of the simulation result. The percentage of relative errors for the hyper-exponential (Hyp), the Erlang (Erl), the lognormal (Log) and the bounded Pareto (Par) distributions are shown in table 6.2 for a system with $\rho_2^h = 2$. The percentages confirm that the CTMC obtained when the MFGC or the VPC policies are considered is not insensitive to the session duration distribution.

6.5 Conclusions

In this chapter, we propose an AC policy and prove that the CTMC that models the multiservice mobile cellular system that implements this AC policy is reversible. We also prove that trunk reservation policies do not lead to reversible CTMCs unless further restrictions are imposed.

In addition, the stationary distribution of the CTMC enforcing the proposed AC policy under study is insensitive to the session duration distribution, in the sense that it depends on the session duration distribution at each node through the mean only. In other words, when arrivals follow Poisson processes, all key performance indicators obtained from the stationary distribution, like blocking probabilities, are independent from all traffic characteristics beyond the traffic intensity. We show some results that confirm this insensitivity property. We also show some results that confirm that the CTMC obtained when the MFGC or the VPC policies are considered is not insensitive to the distribution of the session duration distribution.

Part II

4G mobile networks

Chapter 7

AC in OFDMA based mobile cellular networks

7.1 Introduction

Forthcoming mobile cellular networks based on *Orthogonal Frequency-Division Multiple Access* (OFDMA), such as the *Long Term Evolution* (LTE) networks defined by the *3rd Generation Partnership Project* (3GPP) or the next 4G networks, have been developed in order to face the unprecedented growth in the data-traffic volume experienced during the recent years in mobile cellular networks. In order to enhance the capacity of these OFDMA based networks, a technique called *Adaptive Modulation and Coding* (AMC) is employed. This technique allows different *Modulation and Coding Schemes* (MCSs) to be used in order to maximize the network performance. The idea is to use the most appropriate MCS depending on the signal quality and the interference received by a user at a certain point in time and space. Therefore, the data between a user and the BS can be sent at different rates as the user moves around depending on the position of the user in the cell. As a consequence, the number of resources that a user needs to guarantee its QoS requirement depends on the received signal quality and the position of the user. There-

fore, the total capacity of a cell, i.e., the total throughput achieved in the cell, is time-varying and depends on the position and the QoS requirements of each user.

The varying capacity has an impact on the QoS experienced by users. If users move towards zones with lower signal quality, the total capacity of the cell can be insufficient to serve all users, even when no more users have been accepted to the cell. Therefore, these networks introduce new challenges in the frame of RRM and specifically in the AC that controls which sessions are allowed to connect to the cell. The AC policies implemented in such systems become a key mechanism to guarantee the required QoS.

Analytical models for studying AC in cellular networks have already been widely studied in the literature, but most of them do not consider the variation of the cell capacity due to user movement. Remember for example [RTN97] and [GRMBP05], where the efficiency of several AC policies in cellular networks, where the cell capacity is assumed to be constant, is studied. In a more recent work, an AC policy is also suggested in [AJ09], where users requesting admission are gradually admitted by limiting the new user's throughput until the user is fully integrated in the system. An analytical model considering varying cell capacity is described in [KGG10], three AC policies for capacity-varying networks are discussed in [SR01] and an AC policy that takes into account the mobility of the users is proposed in [EC05]. However, the validity of the mobility assumptions made in the analytical models of these works is not discussed there.

Studying the effects of considering a varying cell capacity in wireless networks by using simulations has already been done in several publications and projects. In [JHJ05], an AC policy for *Worldwide Interoperability for Microwave Access* (WiMAX) is proposed and evaluated. In [SSB10], a self-optimisation AC policy for LTE is studied. Moreover, several algorithms that optimize the parameters of AC policies have been evaluated by simulation in several papers [PB05, SNBH00, SSB10, HHS04] and projects, such as SOCRATES [soc], MONOTAS [mon] and E^3 [e3].

The contributions of the work presented in this chapter are the proposal of an analytical model to study the AC in mobile cellular networks considering varying cell capacity and the validation of the mobility assumptions in the analytical model by using simulations. We identify the cases and the reason why both models differ. We also propose a static and a dynamic AC and study their performance using the analytical model proposed. This work resulted into the publications in [SBMS+11, SBMS+ed].

This chapter is structured as follows. In the next section, we describe the new characteristics introduced by the AMC technique. In Section 7.3, the proposed static and dynamic AC policies and the analytical models to study their performance are presented. In Section 7.4, the numerical results are discussed and validated by comparing them with simulations for both AC policies. Section 7.5 concludes the chapter.

7.2 Adaptive modulation and coding

In this work we consider the implementation of the AMC technique as it is made in LTE, but the same principles can also be applied to other network technologies. The AMC technique implemented in LTE networks defines 15 different MCSs. In order to obtain them, different code rates are combined with 3 modulation schemes (16-QAM, 64-QAM, QPSK). The MCS is determined by the signal quality and the interferences received by the MT, i.e., the *Signal to Interference Ratio* (SIR) will determine the MCS used in the transmission between an MT and the BS. The theoretical maximum bitrate for each MSC can be calculated based on the modulation scheme and the coding rate, and then, the minimum SIR which is needed to be able to use a certain MCS can be determined from the attenuated Shannon bound in (7.17), [3GP10b].

The coding rate determines how many bits corresponds to information bits among the total number of bits transmitted, which include the *Forward Error Correction* (FEC) bits. The modulation scheme determines how these bits are converted to a signal. The FEC bits are used for controlling errors

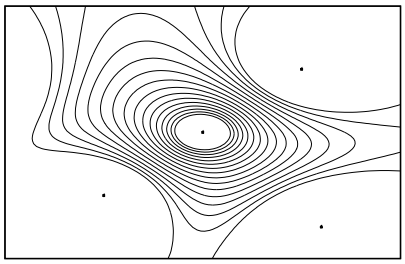

Figure 7.1: Lines of equal interference surrounding a site.

in data transmission over noisy channels or channels with interference. For low signal strength, more FEC bits are needed per information bit. In mobile cellular networks, the transmitted signal strength and hence the SIR depend on several factors but one of most dominant factors in their reduction is the path loss, which is produced by the spatial dispersion and depends on the distance between transmitter and receiver. Therefore, the SIR perceived by a MT, and thus also the MCS used, depend on the distance between the MT and the BS. In Figure. 7.1, the lines with equal SIR for a site surrounded by a given number of interfering sites are represented. We can see that the lines are cycles around the BS. As a consequence, the number of resources which are needed to send a certain number of bits towards a user depends on the position of the user.

In systems based on OFDMA like LTE, the frequency domain is divided into non-overlapping subchannels which occupy a bandwidth of 180kHz. The time domain is divided into slots of 1ms. These subdivisions in time and frequency, generally referred to as Resource Blocks (RBs), are the smallest time-frequency units that can be assigned to a user and correspond to a set of twelve adjacent subcarriers and seven OFDM symbols [NV08]. We assume that each of these RBs contains 84 symbols, resulting in a symbol rate

of 84000 Bd per subchannel. The number of RBs needed by a user depends on the position of the user, and thus, it changes as the user moves around. As pointed above, the total cell capacity turns out to be time-varying.

In this work, we assume an optimal use of the available resources. The RBs are only left unused if there is no data to transmit. Otherwise, if the users require more RBs than the available ones, the available RBs are divided among the users proportionally to the number of RBs that they require. The cell capacity is estimated based on the technique described in [JHJ05]. Let r_k be the required bitrate by user k and b_k be the number of bits that can be sent in a single RB. The average number of RB per second that are needed by user k, n_k, and the total number of resources needed to serve all users with their required bitrates, R, are given by:

$$n_k = \frac{r_k}{b_k}, \qquad R = \sum_k n_k = \sum_k \frac{r_k}{b_k}. \tag{7.1}$$

Suppose R_A is the available number of RBs per second in the cell. Then, the cell capacity C can be estimated as follows:

$$C = \begin{cases} \dfrac{R_A}{R} \displaystyle\sum_k r_k & \text{if} \quad R < R_A \\ \displaystyle\sum_k \frac{R_A}{R} b_k n_k = \frac{R_A}{R} \sum_k r_k & \text{if} \quad R \geq R_A \end{cases} \tag{7.2}$$

Note that if the required number of RBs per second is smaller than the available number of RBs per second ($R < R_A$), each user is given the bitrate it requires and the unused RBs are considered as if they would be proportionally used for all users by multiplying with the term $\frac{R_A}{R}$. If there are insufficient RB ($R \geq R_A$), since the available number of resources is divided among the users proportionally to the number of RBs they require, a user k will receive $\frac{n_k}{R} R_A$ RBs.

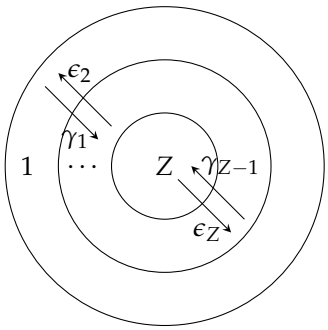

Figure 7.2: Zones with transition rates.

7.3 System description and analytical model

In order to model that time-varying cell capacity, we consider a single cell that is divided into Z concentric rings called zones, where zone 1 is the outermost zone (see Fig. 7.2). Only one, real-time, SC is considered. Although for our purposes it is not of great importance which traffic direction is considered, we consider only the downlink traffic direction. Denote by ρ_i the bits transmitted per RB to a user in zone i ($\rho_1 < \rho_2 < \cdots < \rho_Z$). The cell has R_A RB available per second and an active user requires a bitrate of $r_k = r$ bits per second to fulfill its QoS requirements. Therefore, the number of RB per second that a user needs to achieve its required bitrate r in zone i is:

$$s_i = \frac{r}{\rho_i}.$$

We consider that once a user is accepted in the system it cannot be expelled. If a user changes the zone where it is served, it will remain in the system even when more RBs than the available resources R_A are needed, i.e., we consider a non-preemptive AC policy. This issue will be discussed below in the subsections devoted to the AC policy.

For the sake of mathematical tractability we make the common assumptions that new sessions arrive according to a Poisson process with rate λ and

session durations are exponentially distributed with rate μ. Assuming uniformly distributed traffic, if A_T is the total area of the cell and A_i is the area corresponding to zone i (i.e. $A_T = \sum_i A_i$), we can consider that the arrival rate for new sessions in zone i is $\lambda_i = \frac{A_i}{A_T}\lambda$. The zone residence time, i.e the time that a user stays in a certain zone before entering another one, is also assumed to be exponentially distributed with rate $\epsilon_i + \gamma_i$, where ϵ_i is the rate for transitions from zone i to zone $i-1$, $i = 2, \ldots, Z$ and γ_i is the rate for transitions from zone i to zone $i+1$, $i = 1, \ldots, Z-1$ (see Fig. 7.2). We assume that users start their sessions inside the cell and do not leave the cell until their session is finished.

In order to evaluate the performance of the analytical model, two performance measurements are defined: the total blocking probability (P_T) and the low QoS probability (P_{QoS}). The total blocking probability is the probability that a session, which arrives at any zone of the cell, is blocked by the AC. The low QoS probability is the fraction of time that active users experience a low QoS, i.e. the bitrate they receive is lower than r due to the time-varying cell capacity.

We first present a static AC with fixed policy parameters and the analytical model used to evaluate its performance. Then, we present a dynamic AC which optimizes the policy parameters and an extension of the analytical model developed for the static AC policy, which is used to evaluate the performance of the dynamic AC policy.

7.3.1 Static AC policy

In order to guarantee an acceptable QoS (i.e., an acceptable bitrate) for the users which are in the system, the acceptance of new arrivals is controlled by an AC policy. Let $f \in [0,1]$ denote the AC threshold that determines which fraction of resources has to be available for new sessions in order to be accepted by the AC policy. A new session is accepted if after accepting the session there would still be more than $(1-f)R_A$ RBs available.

Remember that R is the total number of resources needed to serve all users with the required bitrate r defined in (7.1). Upon the arrival of a new session the system compares with fR_A the number of RBs that will be needed to serve all the users at bitrate r if its arrival is accepted. Therefore, to decide on the acceptance of a new session in zone i, the following decisions can be taken:

$$R + s_i \begin{cases} \leq fR_A & \text{accept} \\ > fR_A & \text{reject} \end{cases} \tag{7.3}$$

Once a user is admitted to the system we assume that it cannot be expelled before its session ends. If a user makes an outward zone-transition, it can happen that more RBs than the available are needed ($R > R_A$), since users need more RB in the destination zone than in the origin zone to maintain their bitrate. In this case, all users share the available RBs proportionally to the amount they requested. Thus, all users will receive a lower bitrate than the required r and hence, all users will be served with a lower QoS.

We model the proposed system with the static AC policy using a multidimensional CTMC, where the system state vector is described by the Z-tuple $\boldsymbol{x} = (x_1, \ldots, x_Z)$, where x_i represents the number of users in zone i. Let M be the maximum number of active users which can be present in the system. Since the highest number of bits which can be transmitted per RB corresponds to users in zone Z, M is determined by the maximum number of users accepted in this zone when there are no active users in the other zones:

$$M = \left\lfloor \frac{fR_A}{s_Z} \right\rfloor .$$

Thus, the set of feasible states is thus given by:

$$\mathcal{W}_s := \left\{ \boldsymbol{x} : x_i \in \mathbb{Z}; \quad \sum_{i=1}^{Z} x_i \leq M \right\}. \tag{7.4}$$

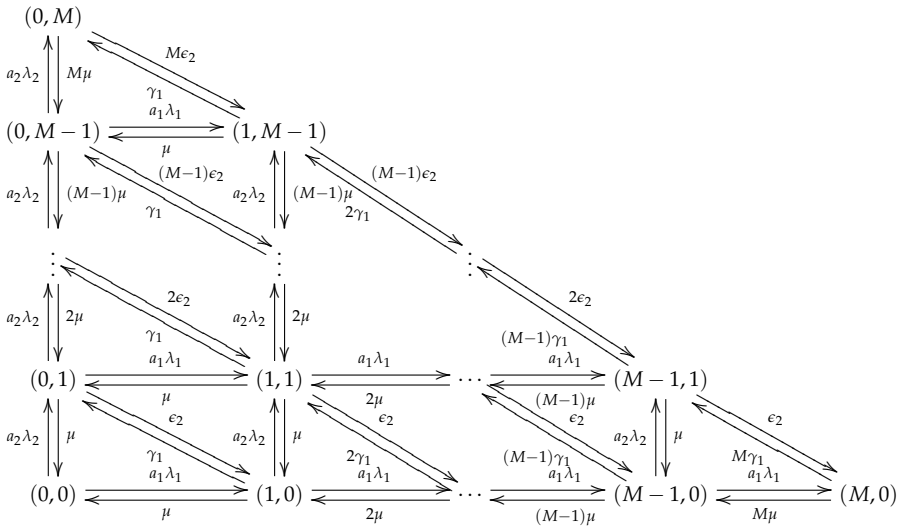

Figure 7.3: Transition diagram of the bi-dimensional model.

The total number of RBs that are needed per second to serve all users at the required bitrate r at state $\boldsymbol{x}$ is represented by

$$R(\boldsymbol{x}) = \sum_{i=1}^{Z} x_i s_i.$$

The function $a_i(\boldsymbol{x})$ denotes whether a new session which arrives in zone i when the system is in state $\boldsymbol{x}$ is accepted by the AC policy or not. The value $a_i(\boldsymbol{x}) = 1$ means that the session is accepted and $a_i(\boldsymbol{x}) = 0$ means that the session is blocked.

As an example, figure 7.3 shows a CTMC for $Z = 2$, i.e., for a system with two zones ($i = 1, 2$) and therefore one incoming rate γ_1 and one outgoing rate ϵ_2. For clarity, the notation has been simplified as $a_i(\boldsymbol{x}) = a_i$. If we define phases as the number of users in zone $i = 2$ and levels as the number of users in zone $i = 1$, we can study the model as a finite level-dependent QBD process [Neu81] with $M + 1$ levels, where level h ($h = 0, \ldots, M$) has

$M+1-h$ phases. Therefore, we can construct the transition rate matrix $\boldsymbol{Q}$ with a block-tridiagonal form, see (7.5). The first row of blocks corresponds to level $h = 0$, the second row of blocks to level $h = 1$, etc. Blocks $\boldsymbol{Q}_1^h$ correspond to transitions between phases in level h, blocks $\boldsymbol{Q}_0^h$ to transitions from level h to level $h+1$ and blocks $\boldsymbol{Q}_2^h$ to transitions from level h to level $h-1$.

$$\boldsymbol{Q} = \begin{bmatrix} \boldsymbol{Q}_1^0 & \boldsymbol{Q}_0^0 & \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots \\ \boldsymbol{Q}_2^1 & \boldsymbol{Q}_1^1 & \boldsymbol{Q}_0^1 & \mathbf{0} & \mathbf{0} & \cdots \\ \mathbf{0} & \boldsymbol{Q}_2^2 & \boldsymbol{Q}_1^2 & \boldsymbol{Q}_0^2 & \mathbf{0} & \cdots \\ & & \ddots & \ddots & \ddots & \\ \cdots & \mathbf{0} & \mathbf{0} & \boldsymbol{Q}_2^{M-1} & \boldsymbol{Q}_1^{M-1} & \boldsymbol{Q}_0^{M-1} \\ \cdots & \mathbf{0} & \mathbf{0} & \mathbf{0} & \boldsymbol{Q}_2^M & \boldsymbol{Q}_1^M \end{bmatrix} \tag{7.5}$$

Note that the different blocks of $\boldsymbol{Q}$ have different sizes for different levels h. The total size of the transition rate matrix $\boldsymbol{Q}$ for $Z = 2$ is:

$$M_T^{S,2} = \sum_{h=0}^{M} M+1-h = \frac{M^2+3M+2}{2}.$$

In the case of $Z = 3$ zones, we can define a layered-level structure composed of phases and 2 level layers. Phases are defined as the number of users in zone $i = 3$, low-levels as the number of users in zone $i = 2$ and high-levels as the number of users in zone $i = 1$. The model is a three-dimensional level-dependent finite QBD process where the transition rate matrix $\boldsymbol{Q}$ again follows the structure of (7.5). Moreover, the block matrices $\boldsymbol{Q}_0^h$, $\boldsymbol{Q}_1^h$ and $\boldsymbol{Q}_2^h$ are also constructed with a block-tridiagonal form, see (7.6), (7.7) and (7.8).

$$\boldsymbol{Q}_1^h = \begin{bmatrix} \boldsymbol{A}_1^{h,0} & \boldsymbol{A}_0^{h,0} & \mathbf{0} & \cdots \\ \boldsymbol{A}_2^{h,1} & \boldsymbol{A}_1^{h,1} & \boldsymbol{A}_0^{h,1} & \cdots \\ \ddots & \ddots & \ddots & \\ \cdots & \mathbf{0} & \boldsymbol{A}_2^{h,M-h} & \boldsymbol{A}_1^{h,M-h} \end{bmatrix} \tag{7.6}$$

$$Q_0^h = \begin{bmatrix} \boldsymbol{B}_1^{h,0} & \mathbf{0} & \mathbf{0} & \cdots \\ \boldsymbol{B}_2^{h,1} & \boldsymbol{B}_1^{h,1} & \mathbf{0} & \cdots \\ \ddots & \ddots & & \\ \cdots & \mathbf{0} & \mathbf{0} & \boldsymbol{B}_2^{h,M-h} \end{bmatrix} \tag{7.7}$$

$$Q_2^h = \begin{bmatrix} \boldsymbol{C}_1^{h,0} & \boldsymbol{C}_0^{h,0} & \mathbf{0} & \cdots \\ \mathbf{0} & \boldsymbol{C}_1^{h,1} & \boldsymbol{C}_0^{h,1} & \cdots \\ & \ddots & \ddots & \\ \cdots & \mathbf{0} & \boldsymbol{C}_1^{h,M-h} & \boldsymbol{C}_0^{h,M-h} \end{bmatrix} \tag{7.8}$$

Note that the blocks inside the matrices have different sizes for different levels h and l. Blocks $\boldsymbol{A}_1^{h,l}$ correspond to transitions between phases inside high-level h and low-level l, blocks $\boldsymbol{A}_0^{h,l}$ correspond to transitions from low-level l to low-level $l+1$ inside high-level h and blocks $\boldsymbol{A}_2^{h,l}$ correspond to transitions from low-level l to low-level $l-1$ inside high-level h. Blocks $\boldsymbol{B}_1^{h,l}$ correspond to transitions from high-level h to high-level $h+1$ with low-level l, blocks $\boldsymbol{B}_2^{h,l}$ correspond to transitions from high-level h to high-level $h+1$ and from low-level l to low-level $l-1$. Blocks $\boldsymbol{C}_1^{h,l}$ correspond to transitions from high-level h to high-level $h-1$ with the same low-level l and blocks $\boldsymbol{C}_0^{h,l}$ correspond to transitions from high-level h to high-level $h-1$ and from low-level l to low-level $l+1$. For more details see Appendix B.2.1. The total size of the transition rate matrix $\boldsymbol{Q}$ is:

$$\begin{aligned} M_T^{S,3} &= \sum_{h_1=0}^{M} \sum_{h_2=0}^{M-h_1} M - h_1 + 1 - h_2 = \\ &= \frac{2M^3 + 12M^2 + 22M + 12}{12}. \end{aligned}$$

This block design can be generalized to any number of zones Z by constructing matrix blocks inside matrix blocks with $Z-1$ different level layers.

To solve the level-dependent finite QBD Markov process and obtain the

stationary distribution $\pi(\boldsymbol{x})$ we use the Linear Level Reduction (LLR) algorithm [LR99]. Basically, this algorithm has two stages. First, the state space is reduced by removing one high-level at each step until there is a Markov process on the last high-level left. That Markov process is solved and the stationary distribution vector is constructed in the second stage by adding back one high-level at each step. More details are shown in Appendix C.2. Note that despite the transition rate matrices being large, their sparseness makes the computations feasible.

Let us denote by P_i the blocking probability for new arrivals in zone i and the total blocking probability in the system by P_T. Then:

$$P_i = \sum_{\boldsymbol{x} \in \mathcal{W}_s} (1 - a_i(\boldsymbol{x}))\pi(\boldsymbol{x}), \tag{7.9}$$

$$P_T = \frac{\sum_{i=1}^{Z} \lambda_i P_i}{\lambda}. \tag{7.10}$$

Let $I(\boldsymbol{x})$ denote the indicator function which takes the value 1 when $R(\boldsymbol{x}) > R_A$ and otherwise it takes the value 0. The low QoS probability is then given by:

$$P_{QoS} = \sum_{\boldsymbol{x} \in \mathcal{W}_s} I(\boldsymbol{x})\pi(\boldsymbol{x}). \tag{7.11}$$

7.3.2 Dynamic AC policy

We also propose a dynamic AC policy, which tunes the parameter f of the AC policy. The goal of this policy is to dynamically adapt the parameters of the AC policy to changes of the environment. When for instance the load is high, more resources will be reserved for ongoing sessions, whereas when the load is low, more new sessions will be admitted to the cell.

This dynamic AC policy, at certain time instances, checks the current load of the system and the parameter f will be updated based on this load. If at

the time instances the load is checked, the load is high, f will be lowered in order to reserve more resources for ongoing sessions while blocking more new sessions. If, on the other hand, the load is considerably low, meaning that only little resources are in use, f will be raised in order to allow more new sessions to the system and have a higher resource utilization.

The load is considered to be high when there are insufficient resources available to serve all sessions in the cell with their required bitrate r, i.e., when $R > R_A$. The load is considered to be considerably low when the total number of resources that are in use is less than a certain fraction, g, of the available resources for new sessions, i.e., when $R < fgR_A$. The fraction g is a system parameter with a predefined value. The raising and lowering of f is done in discrete, evenly sized steps denoted by the parameter Δf. The parameter f is also bounded by a lower and upper limit, denoted f_m and f_M respectively. The number of different discrete values that f can take is denoted by n_f. The upper limit of f is $f_M = f_m + (n_f - 1)\Delta f$.

The algorithm that decides whether f is raised or lowered is given by:

$$R \quad \begin{cases} > R_A & f \leftarrow \max(f - \Delta f, f_m) \\ < fgR_A & f \leftarrow \min(f + \Delta f, f_M) \\ \text{otherwise} & \text{leave } f \text{ unchanged.} \end{cases} \tag{7.12}$$

In order to decide on the acceptance of a new session in zone i, the same decisions than in (7.3) are taken, but considering that the parameter f can be different for different arrivals.

The system with the dynamic AC policy is also modeled using a multi-dimensional CTMC. In this case the stationary distribution is described by the $(Z+1)$-tuple $\boldsymbol{x} = (x_1, \ldots, x_Z, f)$. In this stationary distribution vector, x_i represents the number of users in zone i and f represents the value of the AC threshold which can take the values f_m, $f_m + \Delta f$, $f_m + 2\Delta f$, $\ldots$, f_M. The intervals between two optimizations, i.e., the intervals after which Eq. (7.12) is checked and the appropriate action is taken, are considered to be exponentially distributed with mean $1/\eta$.

As with the static AC policy, the highest number of bits that can be transmitted per RB is achieved in zone Z. Also, the maximum number of active sessions in the system, M, is determined by the maximum number of sessions accepted in zone Z when there are no active users in the other zones and the AC threshold is equal to its upper limit, i.e. $f = f_M$. The set of feasible states is thus given by:

$$\mathcal{W}_d := \Big\{ \boldsymbol{x} : \quad x_i \in \mathbb{Z}, \quad \sum_{i=1}^{Z} x_i s_Z \leq f_M R_A; \\ f \in \{f_m, f_m + \Delta f, \ldots, f_M\} \Big\}. \tag{7.13}$$

We use the same system parameters and make the same assumptions as when the static AC policy is used.

Again, we can study the model as a finite level-dependent QBD process. In the case of $Z = 3$ zones, the model is a four-dimensional level-dependent finite QBD process where the transition rate matrix $\boldsymbol{Q}$ follows the structure of (7.5). But in this case, we define a layered-level structure composed of phases and 3 level layers, where the phases are defined as the value of f, low-levels as the number of users in zone $i = 3$, medium-levels as the number of users in zone $i = 2$ and high-levels as the number of users in zone $i = 1$. Thus, the block matrices $\boldsymbol{Q}_0^h$, $\boldsymbol{Q}_1^h$ and $\boldsymbol{Q}_2^h$ also follow the block design defined in ((7.6)), ((7.7)) and ((7.8)). Moreover, the matrices $\boldsymbol{A}_0^{h,m}$, $\boldsymbol{A}_1^{h,m}$ and $\boldsymbol{A}_2^{h,m}$ follow again a block design, as defined in ((7.14)), ((7.15)) and ((7.16)) respectively, where $p = h + m$. For more details see Appendix B.2.2.

$$\boldsymbol{A}_1^{h,m} = \begin{bmatrix} \boldsymbol{D}_1^{h,m,0} & \boldsymbol{D}_0^{h,m,0} & \boldsymbol{0} & \cdots \\ \boldsymbol{D}_2^{h,m,1} & \boldsymbol{D}_1^{h,m,1} & \boldsymbol{D}_0^{h,m,1} & \cdots \\ \ddots & \ddots & \ddots & \\ \cdots & \boldsymbol{0} & \boldsymbol{D}_2^{h,m,M-p} & \boldsymbol{D}_1^{h,m,M-p} \end{bmatrix} \tag{7.14}$$

$$\boldsymbol{A}_0^{h,m} = \begin{bmatrix} \boldsymbol{E}_1^{h,m,0} & \boldsymbol{0} & \boldsymbol{0} & \cdots \\ \boldsymbol{E}_2^{h,m,1} & \boldsymbol{E}_1^{h,m,1} & \boldsymbol{0} & \cdots \\ \ddots & \ddots & & \\ \cdots & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{E}_2^{h,m,M-p} \end{bmatrix} \tag{7.15}$$

$$\boldsymbol{A}_2^{h,m} = \begin{bmatrix} \boldsymbol{F}_1^{h,m,0} & \boldsymbol{F}_0^{h,m,0} & \boldsymbol{0} & \cdots \\ \boldsymbol{0} & \boldsymbol{F}_1^{h,m,1} & \boldsymbol{F}_0^{h,m,1} & \cdots \\ & \ddots & \ddots & \\ \cdots & \boldsymbol{0} & \boldsymbol{F}_1^{h,m,M-p} & \boldsymbol{F}_0^{h,m,M-p} \end{bmatrix} \tag{7.16}$$

Note that $\boldsymbol{A}_0^{h,m}$ and $\boldsymbol{A}_2^{h,m}$ are not square matrices. The total size of the transition rate matrix $\boldsymbol{Q}$ is:

$$M_T^{A,3} = \sum_{h_1=0}^{M} \sum_{h_2=0}^{M-h_1} n_f(M - h_1 + 1 - h_2) = \\ = n_f \cdot \frac{2M^3 + 12M^2 + 22M + 12}{12}.$$

This block design can be generalized to any number of zones Z by constructing matrix blocks inside matrix blocks with Z different level layers.

To solve the level-dependent finite QBD Markov process and obtain the stationary distribution $\pi(\boldsymbol{x})$ we use again the LLR algorithm.

The blocking probability P_i for new arrivals in zone i, the total blocking probability in the system P_T and the low QoS probability P_{QoS} are again given by (7.9), (7.10) and (7.11).

7.4 Numerical evaluation

In this section we discuss the performance results of the AC policies presented in Sections 7.3.1 and 7.3.2. We also present the results of a compara-

tive study between results obtained with the analytical model and results obtained with a simulation model. This study is performed in order to evaluate the assumptions made in the analytical model of exponentially distributed service durations and transition rates between zones in comparison to more realistic modeling assumptions. In the simulation, users move around with a certain velocity and distance traveled in a single leg of the mobility model. Sessions are generated according to a Poisson process with arrival rate λ. The duration of a session is, unlike in the analytical model, chosen from a lognormal distribution as this distribution models the duration of sessions more realistically [GLZ07]. For more details about the simulation model see Appendix D.1.

7.4.1 Parameter setting

The parameters that are fed into the analytical and simulation models are based on the evaluation scenarios described in [NGM08]. The carrier frequency is chosen to be 2 GHz, the pathloss model that is associated with this frequency is

$$L = 37.6 \log_{10}(D) + 128.1,$$

where D is the distance between the BS and the MT. The operating bandwidth is 5 MHz which means that there are 25 subchannels of 180 kHz (plus guard band), resulting in 25000 RB per second.

The bitrate that can be achieved at a certain distance D from the BS can be calculated as follows. Using the attenuated Shannon bound [3GP10b], the minimum SIR that is needed to achieve the bitrate corresponding to a MCS can be calculated according to:

$$\mathrm{SIR}_i = 2^{\frac{\beta_i}{\alpha}} - 1, \tag{7.17}$$

where β_i is the bitrate per Hertz in zone i and α is the attenuation factor which is 0.6 [3GP10b].

Considering the pathloss model from [NGM08] mentioned above, not taking noise into account and assuming that for every direction the interference comes from a single source at a distance D_{s2s} from the BS with the same transmit power (P_{Tx}) as the BS, an expression for the SIR (in the logarithmic domain) at a given distance from the BS (D) can be constructed:

$$\begin{aligned} \text{SIR} &= P_{\text{Tx}} - L_S - (P_{\text{Tx}} - L_I) = L_I - L_S \\ &= 37.6 \log_{10}(D_{\text{s2s}} - D) + 128.1 \\ &\quad - (37.6 \log_{10}(D) + 128.1) \\ &= 37.6 \log_{10} \left(\frac{D_{\text{s2s}}}{D} - 1 \right) \end{aligned} \tag{7.18}$$

where L_S is the path loss of the signal and L_I the path loss of the interference. And therefore, from (7.18) the radius D_i of zone i, given the SIR can be calculated:

$$D_i = \frac{D_{\text{s2s}}}{10^{\frac{\text{SIR}_i}{37.6}} + 1}. \tag{7.19}$$

By combining (7.17) and (7.19) the radius of a zone can be calculated given the bitrate per Herzt of the MCS. In Table 7.1, the different MCSs, the corresponding efficiency, bitrate per Hertz achieved, SIR and radii are shown. The efficiency represents the number of information bits which can be sent per symbol. By multiplying the efficiency with the symbol rate per Hertz the theoretical maximum bitrate per Hertz achieved can be calculated. Remember that our system has a symbol rate of 84000 Bd per subchannel and one subchannel corresponds to 180 kHz.

Ideally, 15 zones should be considered, each corresponding to a single MCS. We consider that the distance between two BSs is 500 m, which is related to cells in urban environments. As the cell border itself coincides with the circle on which the SIR is 0 dB, the cell border is in the middle and therefore the radius of zone 1 is 250 m. Inside the cell, 10 different MCSs are distinguished. Since using all 10 different MCSs would produce too much computational overhead, the cell was instead divided in 3 zones with equal

Table 7.1: The different MCSs used by LTE [3GP10a].

	Modulation	Efficiency	bit/s/Hz	SIR	Radius (m)
1	QPSK	0.1523	0.0711	-10.6766	329
2	QPSK	0.2344	0.1094	-8.7063	315
3	QPSK	0.3770	0.1759	-6.4710	298
4	QPSK	0.6016	0.2807	-4.1668	281
5	QPSK	0.8770	0.4093	-2.1861	266
6	QPSK	1.1758	0.5487	-0.5309	254
7	16-QAM	1.4766	0.6890	0.8521	243
8	16-QAM	1.9141	0.8932	2.5682	230
9	16-QAM	2.4063	1.1229	4.2477	271
10	64-QAM	2.7305	1.2742	5.2610	210
11	64-QAM	3.3223	1.5504	6.9862	197
12	64-QAM	3.9023	1.8211	8.5716	186
13	64-QAM	4.5234	2.1109	10.1942	174
14	64-QAM	5.1152	2.3871	11.6918	164
15	64-QAM	5.5547	2.5922	12.7824	156

areas (i.e. the size of each zone is $\frac{1}{3}$ of the size of the cell), where the number of bits per RB of zone i denoted be ρ_i, $i = 1, 2, 3$, is obtained considering the different MCSs inside the zone, i.e. it is the weighted average number of bits per RB of the areas that overlap with the corresponding zone i.

We model the traffic as a fluid flow with a bitrate of 128 kbit/s for scenarios involving only the static AC policy and 256 kbit/s for scenarios involving the dynamic AC policy. Unless otherwise indicated, the speed of the users is set to $v = 30$ km/h, the distance traveled by users $d = 30$ m, the average duration of a session is $1/\mu = 300$ s, the average time between arrivals $1/\lambda = 3$ s and the static AC policy parameter $f = 1$. For the dynamic AC policy, the parameter Δf is set to 0.1 and f_m and f_M are set to 0.54 and 1 respectively. This means that the value of f can take the values 0.5, 0.6, 0.7, 0.8, 0.9 and 1. The mean optimization interval $1/\eta$ is set to 60 s. The value of the optimization threshold g is set to 0.8.

The rates ϵ_i and γ_i that are used in the analytical model only depend on

Table 7.2: Model parameter summary

Parameter	Symbol	Value
BS-to-BS distance	D_{s2s}	500 m
Available resources	R_A	2500 RB/s
Traffic source rate	r	128 kbit/s 256 kbit/s
Mean session duration	$1/\mu$	300 s
Mean session i/a time	$1/\lambda$	3 s
AC threshold	f	1
Optimisation threshold	g	0.8
Minimum AC threshold	f_m	0.5
Maximum AC threshold	f_M	1
AC threshold step	Δf	0.1
User velocity	v	30 km/h
Mobility distance	d	30 m
Mean optimisation interval	$1/\eta$	60 s
Radius zone 1	D_1	250 m
Relative area zone 1	a_1	33 %
Bits/RB zone 1	ρ_1	154.81
Radius zone 2	D_2	204.12 m
Relative area zone 2	a_2	33 %
Bits/RB zone 2	ρ_2	349.87
Radius zone 3	D_3	144.33 m
Relative area zone 3	a_3	33 %
Bits/RB zone 3	ρ_3	466.59

the velocity, the distance traveled in a single leg of the mobility model and the size of the zones. In order to determine these transition rates, simulations with only one user were executed. In these simulations a single user walks around without starting any sessions. Each time the user crosses the border of a zone, the event is recorded and at the end of the simulation, the mean transition rates are calculated. A summary of all model parameters is given in Table 7.2.

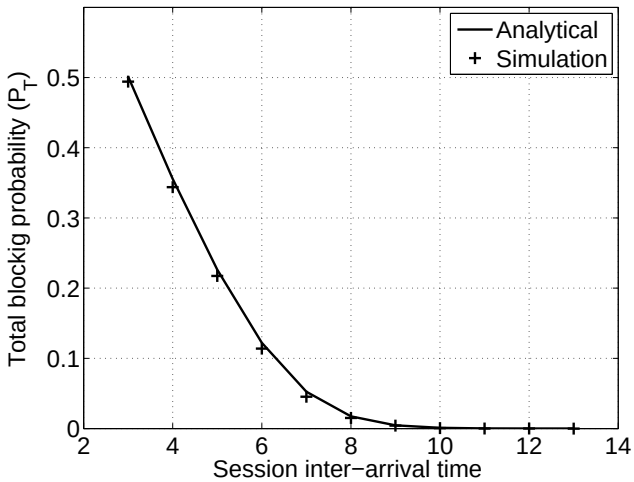

Figure 7.4: Blocking probability for various mean session inter-arrival times.

7.4.2 Numerical results

In this section, the analytical model is validated using simulations. Also, the numerical results obtained with both AC policies, the static and the dynamic, are analyzed and compared.

Validation of the analytical model

In this section, the analytical model is validated comparing the results obtained with the analytical model with the results obtained by simulations. In this subsection, we compare results using the static AC policy.

Results in Fig. 7.4 show the blocking probabilities P_T as a function of the mean session inter-arrival time ($1/\lambda$). The results obtained using the analytical model are represented using lines while the results of the simulations are represented using only markers. As expected, increasing the mean inter-arrival time causes the blocking probabilities P_T to decrease. That is because when the mean inter-arrival time is low, more sessions will be started in a

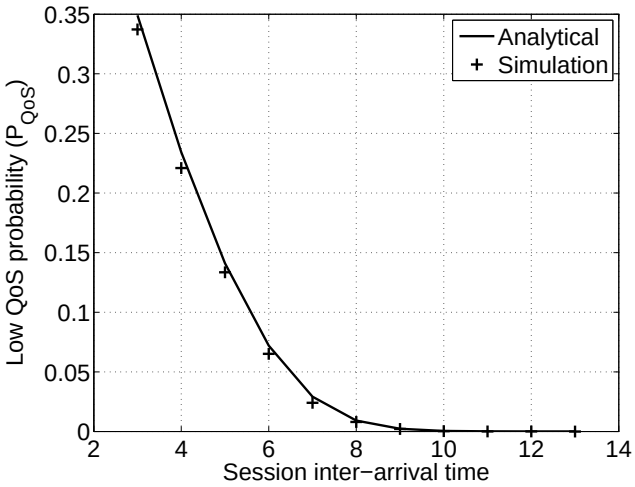

Figure 7.5: Low QoS probability for various mean session inter-arrival times.

shorter amount of time. As there are only a fixed number of resources available, this will cause more sessions to be blocked by the system resulting in a higher blocking probability. When the inter-arrival time is high, fewer sessions will be blocked. As can be seen, the results obtained with the analytical model and the simulations are very similar.

Also the low QoS probability P_{QoS} obtained with the analytical model and with the simulations are very similar as can be seen in Fig. 7.5. As with the blocking probabilities, the QoS is worse when the session inter-arrival time is low than when the session inter-arrival time is high. That is a consequence of the varying cell capacity. Remember that the AC threshold is set to 1 in this section, therefore the system will be filled up until all the RBs are occupied. A different number of resources is needed depending on the position of the users, users move around the cell and a user cannot be expelled out the system once accepted. If users tend to move to outermost zones, it can happen that more resources are needed than when the users were accepted and the system will no longer be able to fulfill the QoS requirements of all users.

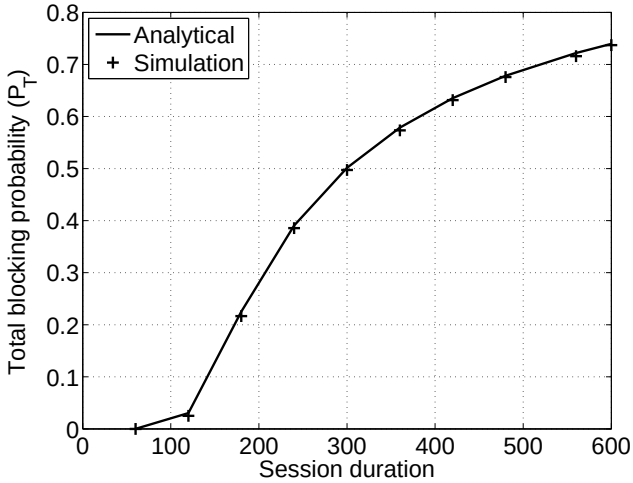

Figure 7.6: Blocking probability for various mean session durations.

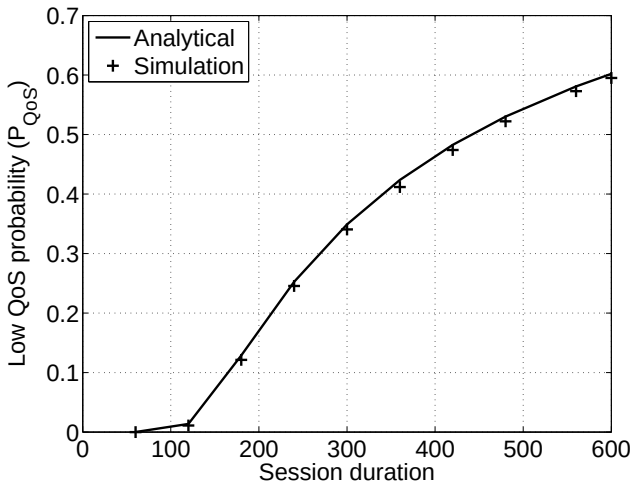

Figure 7.7: Low QoS probability for various mean session durations.

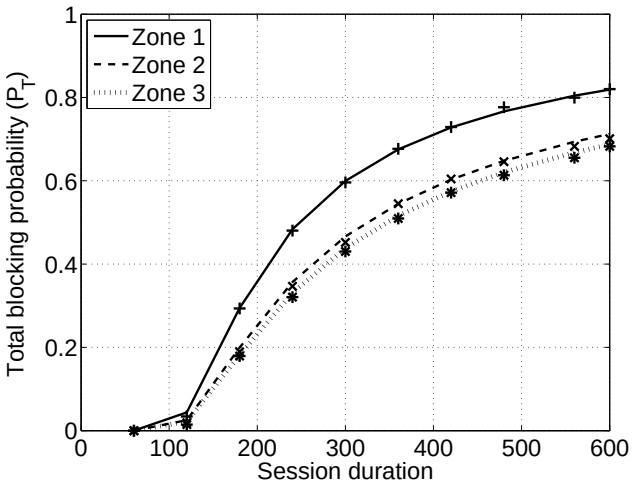

Figure 7.8: Blocking probability per zone for various mean session durations.

In Fig. 7.6 the total blocking probability when the mean session duration is varied is shown. When the session duration is longer, it means that user occupy resources longer times and hence, more sessions will be blocked. For the same reason the low QoS probability is higher when the session duration is longer as can be seen in Fig. 7.7. Here, the results of the analytical model and the simulations are again very similar in both figures.

In Fig 7.8, the blocking probabilities of the individual zones (P_1, P_2 and P_3) are shown. Again lines correspond to the results of the analytical model and markers to the results of the simulations. As can be seen in this figure the analytical and simulation results for the individual zones again match very accurately, meaning that the analytical model is also accurate for the individual zones and not only for the entire cell.

Although the results obtained with the analytical and simulation models fit very well, there are cases in which the results of the analytical model and the simulations can differ. Some parameters have been studied and it has been found that there are some deviations when the distances traveled by

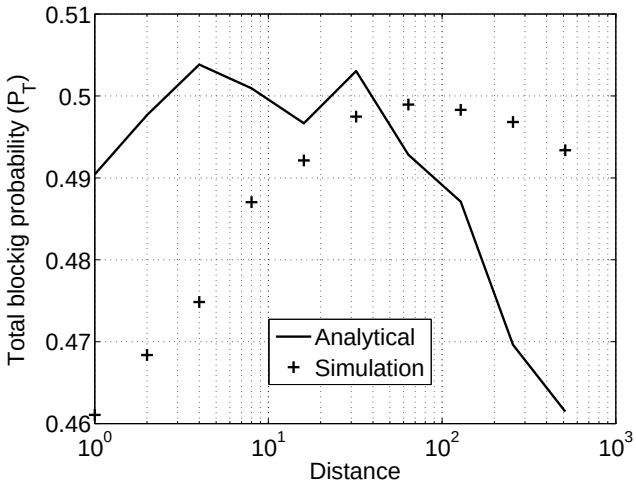

Figure 7.9: Blocking probability for various distances traveled by users in a single leg.

users in a single leg of the random walk mobility model (d) are either very short or very long in comparison to the radii of the zones. An example of this can be seen in Fig. 7.9, where the distance is varied between 1 m and 512 m (note that the cell radius is 250 m). As can be seen, the differences between the results obtained from the analytical model and the simulations are similar when the distance d lies between 10 m and 100 m. When d is either shorter or longer the deviations between both models become bigger.

In the case of short d, users remain at nearly the same location. Thus, the users which are close to the border of a zone will cross the border of the zone many times, while the users which are further away from the border will likely finish their session before they make a transition. In the case of long d, the distance that is traveled by users is bigger than the radii of the zones. Thus, users will cross the zones more than once before choosing a new direction. The time between entering a zone and leaving it again is bounded by the minimum and maximum distances that can be traveled in a zone in a straight line, divided by the velocity of the users. Therefore, the distance

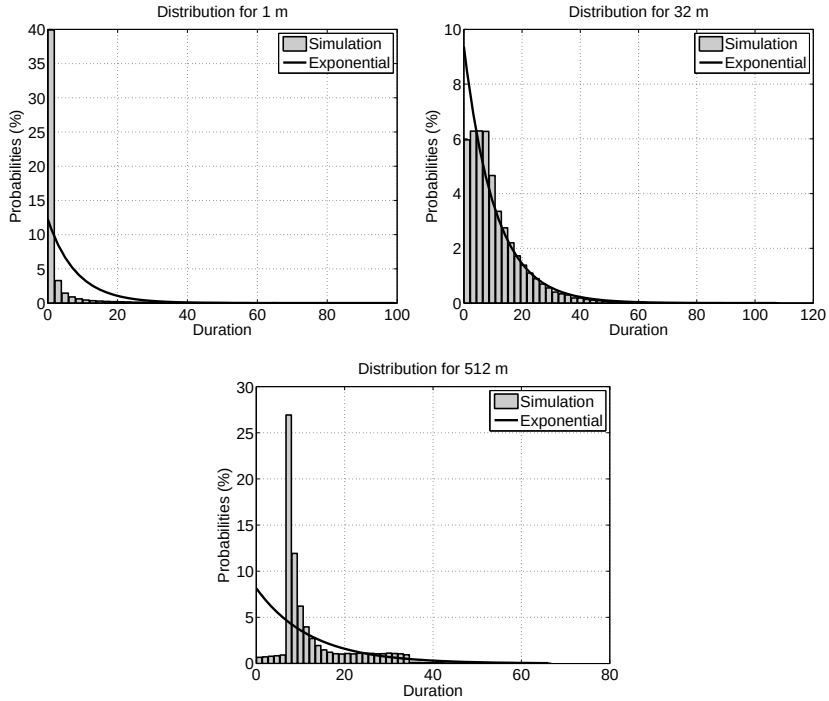

Figure 7.10: Distribution of time spent in zone 2 before going to zone 1.

traveled in a zone in a straight line will have an influence on the distribution of the transition times from one zone to another.

Results in Figure 7.10 show the distribution of the time that users spend in zone 2 before going to zone 1. The times are measured in simulations where the distance traveled in a single leg are respectively 1 m, 32 m and 512 m. The plots also contain the pdf of the exponential distribution that is used in the analytical model to model the time that a user stays in that zone, i.e., the pdf of an exponential distribution with mean $1/\epsilon_2$. As can be seen in the figures, the distribution of the times resembles the exponential distribution best when the traveled distance is around 30 m.

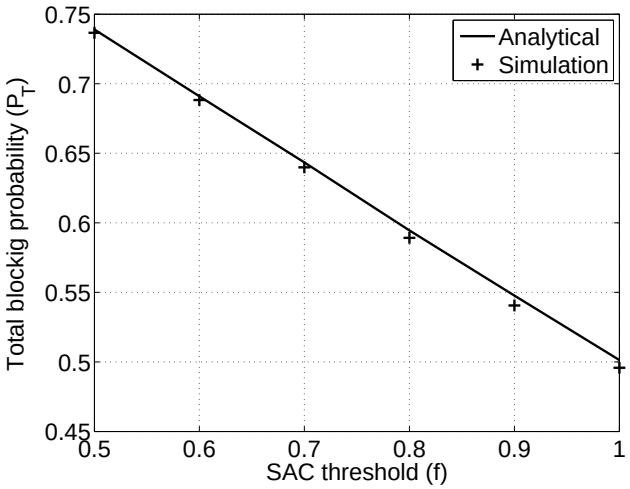

Figure 7.11: Blocking probability for various AC thresholds f.

Analysis of the static AC policy

In this section we briefly evaluate the static AC policy using the developed analytical model and simulations. In Fig. 7.11, the total blocking probability for various values of the AC threshold f is shown. As can be seen, the blocking probability decreases as the AC threshold increases as expected. It's clear that when the AC threshold increases, more new sessions will be allowed to the cell, causing the blocking probability to decrease.

In Fig. 7.12, the low QoS probability is shown as function of the AC threshold. The AC policy should prohibit the QoS from ever becoming bad, but as the cell capacity varies over time a certain amount of the cell capacity should be reserved as a buffer in order to avoid situations wherein the QoS becomes bad. As we can see in this figure, the QoS remains good until f reaches a value of more than 80 %. When f is higher than this value, the varying cell capacity causes that the cell capacity can drop below the required capacity resulting in a low QoS. This shows the effects of the time-varying cell capacity on the QoS experienced by the active users.

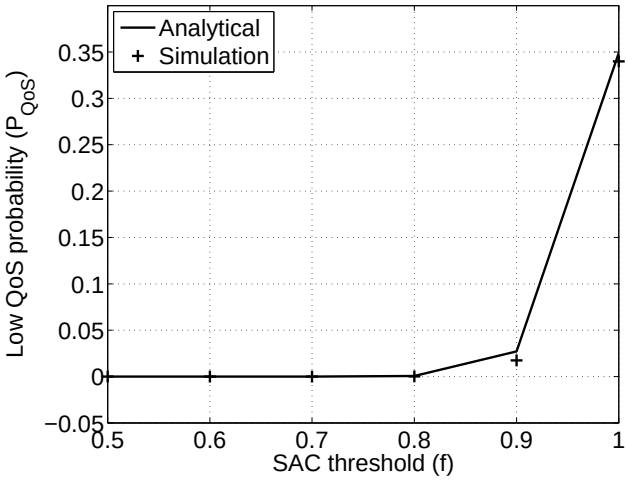

Figure 7.12: Low QoS probabilities for various AC thresholds f.

Analysis of the dynamic AC policy

In this section we use both the analytical and simulation models to examine the performance of the dynamic AC policy. The total blocking probability and the low QoS probability are plotted as function of the session duration and the inter-arrival time. The results obtained with the dynamic AC policy are compared with the results obtained with the static AC policy for various values of the AC threshold f in order to assess the benefits of the dynamic AC policy relative to a static one. The different values of f that are considered for the different static AC policies are $f \in \{0.5, 0.7, 0.9, 1\}$.

Results in Fig. 7.13 show the total blocking probability for both the dynamic AC policy and the static AC policy for various values of f. For short session durations the behavior of the dynamic AC policy tends to the static AC policy with a high AC threshold, while for long session durations, the dynamic AC policy tends to the static AC policy with low AC thresholds. The reason why that happens is because when the session duration is short the system is less loaded and more new sessions can be accepted in the sys-

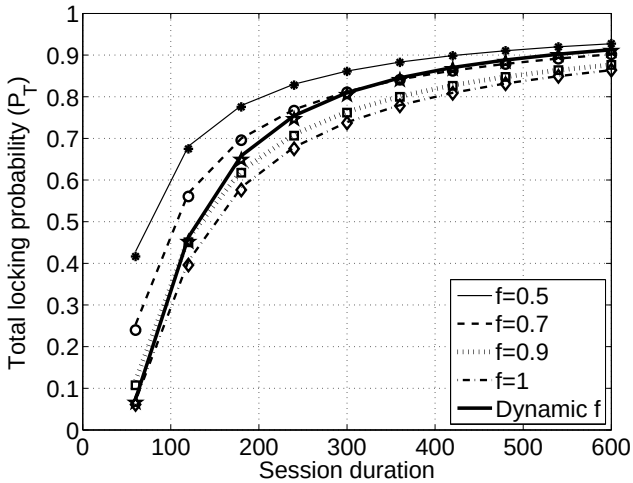

Figure 7.13: Total blocking probability for various session durations for the dynamic AC policy.

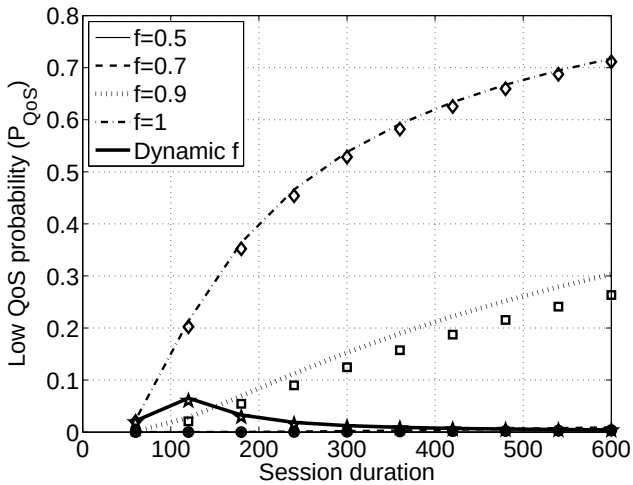

Figure 7.14: Low QoS probability for various session durations for the dynamic AC policy.

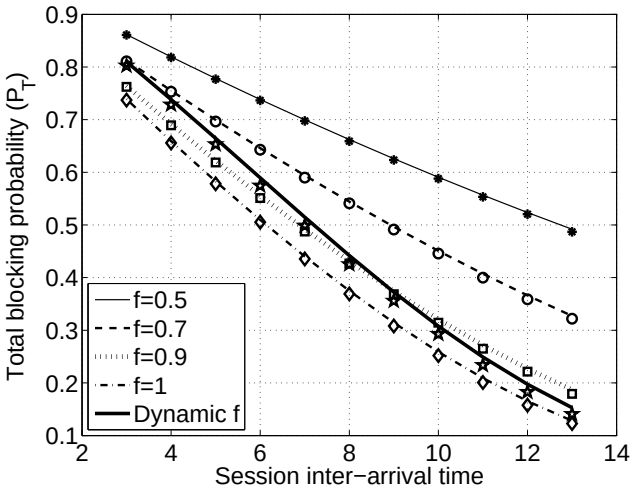

Figure 7.15: Total blocking probability for various inter-arrival times for the dynamic AC policy.

tem without being detrimental to active sessions, thus the AC threshold can be higher. When the session duration is longer, the system is more loaded and the AC threshold is more restrictive with new arrivals.

When looking at the low QoS probability shown in Fig. 7.14, the dynamic AC policy shows a even more optimized behaviour. When the static AC policy is implemented, the low QoS probability rises as the session duration becomes longer. However, the low QoS probability of the dynamic AC policy has a maximum at a mean session duration of 120 s and decreases again for longer session durations. That can be explained as follows: in case the session duration is relatively short, high loads are unlikely and the dynamic AC policy will raise the threshold f to a high value. In the case that the load does become too high (i.e. $R > R_A$), it will take the dynamic AC policy longer to react to this situation as the threshold f should be lowered starting from a high value, while if the load is higher, the threshold will already have a lower value and the algorithm will react swifter.

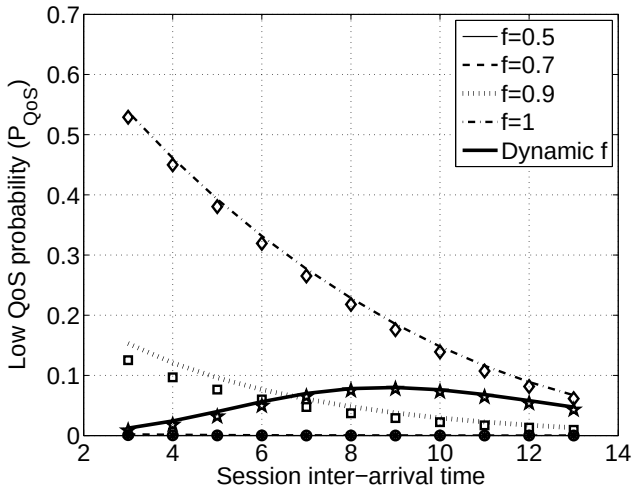

Figure 7.16: Low QoS probability for various inter-arrival times for the dynamic AC policy.

The total blocking probability and the low QoS probability are also shown as function of the session inter-arrival time in Figures 7.15 and 7.16. A similar kind of behavior can be observed. The trend in these plots is however reversed as long inter-arrival times mean that the load is lower, while long session durations mean that the load is higher. The benefits of using a dynamic AC policy can be clearly seen again in these figures. For instance, in Fig. 7.15, the total blocking probability of the static AC policy for long session inter-arrival times and low values of f is much higher than that of the dynamic AC policy. Meanwhile, in Fig. 7.16, the low QoS probability of the dynamic AC policy stays low for short session inter-arrival times in contrast to the static AC policy with $f = 0.9$ or $f = 1$.

Although the suitability of a dynamic AC policy depends on the properties which have to be optimized, like maximal resource utilization or high QoS for active users, using a dynamic AC policy clearly has benefits over a static one.

7.5 Conclusions

In this chapter, we present an analytical model which models the time-varying cell capacity of OFDMA based mobile cellular networks, such as LTE networks. The cell capacity in these systems changes over time as user movement causes the signal quality and thus the MCS to change over time. The time-varying cell capacity in this chapter is modeled by dividing a cell in multiple concentric zones in which a certain bitrate can be achieved when sending data to users in that zone. By assuming that the times between users changing from one zone to another are distributed according to an exponential distribution and the session duration is exponentially distributed, the system can be modeled using a CTMC. In order to verify these assumptions, results obtained with the analytical model in various scenarios were compared to simulation results which were obtained from a simulator that models the user mobility and the session duration more realistically. Results show that the analytical model captures the user mobility very accurately and that the assumption of exponentially distributed session durations yields also accurate resutls.

The developed models were also used to investigate the performance of a static AC policy and a dynamic AC policy which optimizes the parameter of this AC policy. The results obtained from this study show that using a policy that optimizes the parameters of the AC policy has better performance than using fixed parameters.

The analytical model developed here can be further used to evaluate other AC policies and study system design issues such as resource dimensioning. The model can also be extended with handovers to and from neighboring cells. This can be done by considering transitions to and from the outermost zones coming from and going to outside the cell.

Chapter 8

AC in femtocells

8.1 Introduction

As it has been pointed in previous sections, during the recent years, the high penetration of mobile-phone services into the society has lead to an unprecedented growth in the data-traffic volume. This trend will continue in the coming years, as mobile systems are expected to support a larger variety of multimedia services. Unfortunately, the current networks' features are not enough to face this development paradigm. Moreover, according to recent surveys [Man08], the traffic which is expected to produce the bulk of the network load will mainly occur indoor. In this context, the novel concept of femtocells [AGECR10, CAG08] has emerged to increase both network capacity and indoor coverage.

Femtocells are small coverage areas, created by low-power base stations called Femtocell Access Points (FAPs) for providing indoor services. They are owned and installed by the users. As a result, they benefit both users and operators. Users improve their QoS, while operators can manage the growth of traffic without the need to construct new network infrastructures. Moreover, the FAPs send the backhaul data over the Internet to the cellular

operator network, thus allowing operators to release resources from the BS for other macrocell users. However, the deployment of femtocells introduces several technical challenges [LPVdlRZ09].

From the perspective of the femtocell connectivity rights, two types of users are defined:

- *Femtocell Users* (FUs):

  Group of subscribed users registered in the femtocell, which can always connect to the femtocell, i.e., they are the rightful users. They are determined by the femtocell owner and usually belong to the femtocell owner, their family or friends.

- *Macrocell Users* (MUs):

  Non-subscribed users that are not registered in the femtocell. Depending on the type of AC implemented at the femtocell, they will be allowed to connect to the femtocell or not.

One of the performance-limiting factors in femtocell deployments is the cross-tier interference between the macrocell and the femtocell [Cla07]. This problem has been widely addressed in the literature and many approaches have been proposed to cope with it, which involve the use of power control [JMMY09, LQK09] or advanced spectrum management techniques [CA09, SHLK09]. Moreover, the radio interference can be managed by allowing strong macrocell interferers to connect and acquire some level of service in femtocells [dlRVLPZ10]. A key mechanism for operators to provide different levels of priority to FUs and MUs is the AC policy. Thereby, the femtocell has the ability to control which user can have access to it. For this, three AC modes exist [dlRVLPZ10, GMP09]:

- *Closed access mode*:

  Only a subscribed subset of users (FUs), defined by the femtocell owner, can access to the femtocell. A normal service is expected to this subset

of users. Any user that is not part of this subset (MUs) cannot connect to the femtocell, unless an emergency call is needed and no other acceptable cells are available. This model is referred to as Closed Subscriber Group (CSG) cell by the 3GPP.

- *Open access mode*:

  All users of the operator can make use of any femtocell. In that sense, when this AC policy is applied there are not differences between FUs and MUs. All users are treated equally from an access perspective and also from a charging perspective in the femtocell.

- *Hybrid access mode*:

  A limited number of the femtocell resources are available to all users, while the rest are only available to FUs. Users which are not subscribed to the femtocell can acquire some level of service on the femtocell, but the femtocell may also provide the ability to give preferential treatment to subscribed (FUs) over non-subscribed (MUs) users. Thereby, sessions from non-subscribed users may be preempted or rejected in favor of a subscribed user session. Alternatively, non-subscribed user sessions may be handed over to the macrocell. In addition to that, subscribed users may also get preferential charging in comparison with non-subscribed users to the femtocell. This model is referred to as hybrid cell by the 3GPP.

Several studies can be found in the literature which compare open and closed access modes for femtocell networks [NV, XCA10]. On the one hand, in the open access mode, the number of dropped sessions due to cross-tier interference between macrocells and femtocells can be reduced by allowing the most harmful interfering MUs to connect to the femtocell. On the other hand, the closed access mode does not entail security and sharing concerns, and it is more preferred by femtocell customers because they own and install the FAPs in their private environments. The hybrid access mode is proposed [DR09, LYS10] as a trade-off between open and closed access modes,

where the access control has to be carefully chosen depending on the scenario under study and the customer profile.

In this chapter, we develop an analytical model of the FU activity profile to study which and how many channels are the best to be operated in open access mode. Our model assumes that the FUs have priority over the MUs since the femtocell customers are the owners of the FAPs. In our study, if an MU is connected to the femtocell while an FU is in need of the resources used by the MU, the MU will vacate the channel. Hence, our work incorporates the fact that the MUs connect to the femtocell transparently to FUs. To the best of our knowledge, a preemptive and non-resume access control policy for MUs has not been considered in the existing works which study open, close and hybrid access modes.

The study of the hybrid access mode proposed in this chapter allows to identify which channels are the best option for MUs depending on the Signal to Interference Noise Ratio (SINR) experienced by users on each channel and the amount of time an FU is using these channels. The results motivate the need for novel resource management schemes which dynamically adapt the set of channels operating in open access mode depending on the network conditions. This work resulted into the publication in [BMPGEMB12].

This chapter is structured as follows. In the next section, we describe the system model to study the activity profile of FUs. In Section 8.3, we derive expressions for several performance parameters for MUs from the model in Section 8.2. In Section 8.4, we discuss and compare the numerical results. Finally, Section 8.5 concludes the chapter.

8.2 Femtocell user activity profile model

In this section, we present a model of the FU activity profile. We consider a single femtocell with C available channels, from which $C_m \leq C$ are operated in Open Access (OA) mode. Each channel experiences different signal and interference levels and therefore the data rate achieved in each channel is

different. The data rate on channel i is R_i bit/second. We consider that one specific channel has the same average radio characteristics, e.g. SINR, for all users (FUs and MUs) and these are static during the period of time under consideration.

Depending on if the traffic carried by user applications is streaming or elastic, the transmission with higher bitrates has a different impact on the user perception. In case of elastic traffic, the session duration depends on the data rate received, and high data rates entail shorter session durations. In case of streaming traffic, the session duration only depends on the user behavior, and high data rates entail better QoS perception. We consider that FUs generate streaming traffic, but the model could be extended to the case of elastic traffic.

8.2.1 System Model

We model the activity profile of FUs using a multidimensional CTMC. The system state vector $\boldsymbol{x}$ is described by the C-tuple $\boldsymbol{x} = (x_1, \ldots, x_C)$, where x_i represents the state of channel i, taking value 0 when channel i is idle and 1 when channel i is used by an FU, (we say it is busy). We consider that one FU session uses one channel, therefore the number of FUs connected to the femtocell at state $\boldsymbol{x}$ is represented by:

$$N(\boldsymbol{x}) = \sum_{i=1}^{i=C} x_i. \tag{8.1}$$

Due to the fact that the number of femtocell users is small, to consider infinite user population would not be accurate. We consider a finite user population with M FUs. The arrival rate λ at state $\boldsymbol{x}$ is thus given by

$$\lambda(\boldsymbol{x}) = (M - N(\boldsymbol{x}))\lambda_f, \tag{8.2}$$

where λ_f refers to the arrival rate for one idle FU.

Incoming FUs access the channels by following an order, namely, FUs access the channels by choosing the most preferred channel among all the available idle channels in the femtocell. The most preferred channel ($i = 1$) is the channel having the highest data rate, while the least preferred channel ($i = C$) is the channel having the lowest data rate. If there are no idle channels in the femtocell, i.e., all of them are occupied by FUs, an incoming FU is blocked out of the femtocell. For the sake of mathematical tractability, we consider exponentially distributed session durations for FUs with a mean $1/\mu$.

We consider that MUs generate packets with a fixed size $L = L_H + L_D$ bits, where L_H and L_D are, respectively, the header and payload lengths. The MUs use the channels operated in OA mode not used by FU traffic. We consider that the MUs are in saturation, i.e., there are always MUs waiting for free channels. Therefore, for this study, it is not relevant which channels the MUs choose first when they access the femtocell. Upon an FU arrival, MUs vacate the channel chosen by the FU and the MU packet that is being transmitted is interrupted and lost, i.e., we consider a preemptive and non-resume AC policy. One can think that a higher number of femtocell resources operated in OA mode entails a higher throughput achieved by the MUs, but when MUs access a higher number of channels, MU transmissions are more likely to be interrupted by FUs arrivals, and therefore the throughput achieved by MUs can be lower. It is under study in this work how many channels and which channels are assigned as OA mode to MUs.

Let $\boldsymbol{x}'$ represent the state achieved by the femtocell after a state transition and $q_{\boldsymbol{x}\boldsymbol{x}'}$ be the transition rate from $\boldsymbol{x}$ to $\boldsymbol{x}'$. The transition matrix $\boldsymbol{Q}$ when the states are lexicographically sorted can be easily obtained by using the transitions $q_{\boldsymbol{x}\boldsymbol{x}'}$.

The state transitions of the CTMC under study occur when a new FU session connects to the femtocell or when any FU session is finished. The state transitions are shown in Fig. 8.1, where $\boldsymbol{e_i}$ is a C-dimensional vector with a 1 on the i-th position and 0's elsewhere.

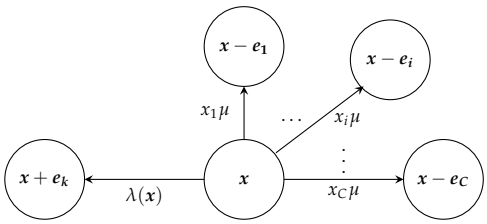

Figure 8.1: Transitions of the CTMC which models the FU activity profile.

- Arrival of an FU session:

  The FU session will connect to the kth channel where $k = \min\{i \mid x_i = 0\}$ and thus $\boldsymbol{x}' = \boldsymbol{x} + \boldsymbol{e_k}$. In this case, the transition rate is $q_{\boldsymbol{x}\boldsymbol{x}'} = \lambda(\boldsymbol{x})$ defined in (8.2). If $x_i = 1\ \forall i$ the arrival is blocked.

- Termination of an FU Session:

  When an FU session which is using the ith channel terminates, the state achieved by the femtocell is $\boldsymbol{x}' = \boldsymbol{x} - \boldsymbol{e_i}$. In this case the transition rate is $q_{\boldsymbol{x}\boldsymbol{x}'} = x_i\mu$.

Note that in state $\boldsymbol{x}$ only one transition can occur due to an arrival of an FU, while up to $N(\boldsymbol{x})$ different transitions can occur when an FU finishes its service. The size of the infinitesimal generator $\boldsymbol{Q}$ is 2^C, but since we are considering households, C is supposed not to take high values, thus, the problem is computationally tractable. The vector of the stationary distribution, denoted by $\boldsymbol{\pi}$, is obtained by solving the global balance equations and the normalization equation given by:

$$\pi(\boldsymbol{x}) \sum_{\forall \boldsymbol{x} \neq \boldsymbol{x}'} q_{\boldsymbol{x}\boldsymbol{x}'} = \sum_{\forall \boldsymbol{x}' \neq \boldsymbol{x}} q_{\boldsymbol{x}'\boldsymbol{x}} \pi(\boldsymbol{x}'),$$

$$\sum_{\forall \boldsymbol{x}} \pi(\boldsymbol{x}) = 1.$$

8.2.2 Characterization of Idle and Busy Periods

Our goal is to characterize the time intervals when an arbitrary channel is used by an FU (busy period, B_i) and the time intervals when an arbitrary channel is not used by any FU (idle period, I_i). Therefore, the busy period B_i corresponds to the CHT of FUs in channel i, which is exponentially distributed with rate μ for all channels. The idle period I_i corresponds to the period of time spent in the set of states with $x_i = 0$. Hence, the duration of an idle period I_i is composed of a number of phases with exponentially distributed duration. Hence, the idle period follows a PH distribution, which defines the time until absorption ($x_i \to 1$) in an AMP [Neu81], and it is represented by $PH(\boldsymbol{\alpha}, \boldsymbol{S})$. Remember $\boldsymbol{S}$ is the transition matrix which contains the transition rates between the states, and $\boldsymbol{\alpha}$ is the initial state probability vector.

For each channel i, a different AMP is denoted as $PH(\boldsymbol{\alpha}_i, \boldsymbol{S}_i)$. The AMP is initiated when channel i becomes idle and the absorption occurs when it becomes busy again. Therefore, the matrix $\boldsymbol{S}_i$ is obtained from $\boldsymbol{Q}$ by removing the rows and the columns corresponding to the states where channel i is busy. The probabilities $\boldsymbol{\alpha}_i$ are the normalized probabilities of initiating the process at each of the states where $x_i = 0$, given by:

$$\boldsymbol{\alpha_i} = \frac{1}{\sum_{\forall \boldsymbol{\pi}_{x_i=1}} \boldsymbol{\pi}_{x_i=1} \boldsymbol{Q}_{x_i=1,x_i'=0}} \boldsymbol{\pi}_{x_i=1} \boldsymbol{Q}_{x_i=1,x_i'=0} \tag{8.3}$$

where $\boldsymbol{\pi}_{x_i=1}$ is a row vector with the probabilities of the busy states and $\boldsymbol{Q}_{x_i=1,x_i'=0}$ is a matrix with transition rates from busy states with $x_i = 1$ to idle states with $x_i = 0$.

The distribution function corresponding to the idle period of channel i is the distribution function of $PH(\boldsymbol{\alpha_i}, \boldsymbol{S}_i)$ which is (see Appendix C.1.2):

$$F_{I_i}(t) = 1 - \boldsymbol{\alpha}_i e^{t\boldsymbol{S}_i} \mathbf{1}, \tag{8.4}$$

where $\mathbf{1}$ is a vector of ones.

The average time in which channel i is idle corresponds to the mean time until absorption in the $PH(\boldsymbol{\alpha_i}, \boldsymbol{S}_i)$ distribution, and it is given by

$$\overline{I_i} = E[PH(\boldsymbol{\alpha_i}, \boldsymbol{S}_i)] = -\boldsymbol{\alpha}_i \boldsymbol{S}_i^{-1} \mathbf{1}. \tag{8.5}$$

Once the FU activity profile is modeled, the expression for some MU performance parameters can be calculated as it is shown in the next section.

8.3 Performance Metrics for macrocell users

According to the model presented in Section 8.2, MUs see the femtocell as a set of channels with different radio characteristics and different FU activity profiles. In this section, we derive the analytical expressions for several performance parameters for MUs, such as the MU throughput, the interruption probability due to an FU arrival or the consumed energy per bit of data successfully transmitted, by starting from the model defined in Section 8.2.

From (8.4), the probability that at least n packets of length L, corresponding to nL bits, are transmitted during the idle period I_i of the channel i is:

$$p_i(n) = P\left(I_i \geq \frac{nL}{R_i}\right) = 1 - F_{I_i}\left(\frac{nL}{R_i}\right) = \boldsymbol{\alpha}_i e^{\frac{nL}{R_i}\boldsymbol{S}_i} \mathbf{1} \tag{8.6}$$

where R_i is the data rate on channel i.

The maximum achievable throughput for MUs in the channel i denoted by γ_i, is defined as the average successfully transmitted data bits during an idle period divided by the total average time of idle plus busy periods. The average number of successfully transmitted data bits in channel i denoted by $\overline{D}_i$, is given by

$$\begin{aligned} \overline{D}_i &= L_D \sum_{n=1}^{\infty} p_i(n) = \\ &= L_D \alpha_i \left(\sum_{n=0}^{\infty} \left(e^{\frac{L}{R_i}\boldsymbol{S}_i} \right)^n - 1 \right) \mathbf{1}, \end{aligned}$$

and finally,

$$\overline{D}_i = L_D \alpha_i e^{\frac{L}{R_i} S_i} \left(\boldsymbol{I} - e^{\frac{L}{R_i} S_i} \right)^{-1} \mathbf{1}, \tag{8.7}$$

where $\boldsymbol{I}$ is the identity matrix and L_D refers to the payload length. From (8.5), (8.7) and knowing that the busy period B_i is exponentially distributed with mean $1/\mu$, the throughput γ_i for the channel i is given by:

$$\gamma_i = \frac{\overline{D}_i}{\overline{I}_i + \overline{B}_i} = \frac{L_D \alpha_i e^{\frac{L}{R_i} S_i} \left(\boldsymbol{I} - e^{\frac{L}{R_i} S_i} \right)^{-1} \mathbf{1}}{-\boldsymbol{\alpha}_i \boldsymbol{S}_i^{-1} \mathbf{1} + 1/\mu}. \tag{8.8}$$

The total achievable throughput denoted by γ_T, is the sum of the achievable throughputs in the C_m channels operated in OA mode. If OA represents this set of channels, γ_T is:

$$\gamma_T = \sum_{\forall i \in OA} \gamma_i. \tag{8.9}$$

During an idle period of time there are $\Phi_i = D_i / L_D$ successfully transmitted packets on the channel i and one packet interrupted by an FU arrival. The probability that an MU packet is interrupted on the channel i, denoted by ξ_i is given by

$$\xi_i = \frac{1}{(\overline{\Phi}_i + 1)}. \tag{8.10}$$

The global MU interruption probability denoted by ξ_G, is obtained by dividing the sum of interrupted transmissions per time unit of each channel operated in OA mode by the total transmissions per time unit in the same channels. The number of transmitted packets per time unit on the channel i is $(\overline{\Phi}_i + 1)/(\overline{I}_i + \overline{B}_i)$. And, therefore the global MU interruption probability is given by

$$\xi_G = \frac{1}{\sum_{\forall j \in OA} \frac{\overline{\Phi}_j + 1}{\overline{I}_j + \overline{B}_j}} \sum_{\forall k \in OA} \frac{\overline{\Phi}_k + 1}{\overline{I}_k + \overline{B}_k} \xi_k,$$

From (8.8) and (8.10),

$$\xi_G = \frac{1}{\frac{\gamma_T}{L_D} + \sum_{\forall j \in OA} \frac{1}{\overline{I}_j + \overline{B}_j}} \sum_{\forall k \in OA} \frac{1}{\overline{I}_k + \overline{B}_k}.$$

And finally, from (8.5)

$$\xi_G = \frac{\sum_{\forall k \in OA} \frac{1}{-\boldsymbol{\alpha}_k \boldsymbol{S}_k^{-1} \mathbf{1} + 1/\mu}}{\frac{\gamma_T}{L_D} + \sum_{\forall j \in OA} \frac{1}{-\boldsymbol{\alpha}_j \boldsymbol{S}_j^{-1} \mathbf{1} + 1/\mu}}. \tag{8.11}$$

The consumed energy per successfully transmitted data bit on the channel i for MUs, denoted by Eb_i, is computed as the energy consumed by the MUs when they are transmitting plus the energy consumed due to the channel monitoring, when channel i is occupied by an FU. Thus, its average value is given by

$$\overline{Eb}_i = \frac{P_{TX}\overline{I}_i + P_s\overline{B}_i}{L_D \overline{\Phi}_i}, \tag{8.12}$$

where P_{TX} is the FAP average transmission power, P_s is the average power consumed to monitor which channels are occupied by FUs, and remember L_D stands for the payload length and Φ_i refers to the successfully transmitted packets in channel i.

The total average consumed energy per successfully transmitted data bit, denoted by $\overline{Eb}$, results from weighting the average energy consumed $\overline{Eb}_i$ per successfully transmitted data bit on each channel i operated in OA mode by the corresponding fractions of throughput in each channel i. This leads to:

$$\overline{Eb} = \sum_{\forall i \in OA} \frac{\gamma_i}{\gamma_T} \overline{Eb}_i. \tag{8.13}$$

8.4 Numerical evaluation

In this section, first the parameter setting of the system is defined. Then, the throughput and the interruption probability achieved considering different set of channels operated with OA mode are shown. Later, the throughput achieved considering different combinations of the SINR experienced by the channels are compared. Finally, the influence of the average session duration in the performance of the system and the consumed energy per successfully transmitted bit is studied.

8.4.1 Parameter Setting

In this subsection, we define the values of the parameters considered in the model. Remember that in commercial systems based on OFDMA, such as LTE, the frequency domain is divided into non-overlapping subchannels which occupy a bandwidth of 180kHz called Resource Blocks (RBs). The time domain is divided into slots of 1ms. These RBs, are the smallest time-frequency units that can be assigned to an user and correspond to a set of twelve adjacent subcarriers and seven OFDM symbols [NV08].

As previously pointed out, each channel experiences different SINR levels and therefore the data rate achieved for the MUs in each channel is considered to be different for each channel. In Table 8.1, the different data rates per RB are detailed depending on the experienced SINR levels [BFDL+09]. Here we assume that a channel corresponds to a RB. We consider a femtocell with $C = 8$ channels. Unless otherwise stated, the data rates achieved by each of the 8 channels are the eight highest rates in Table 8.1. Other combinations have been tried and the results are qualitatively the same.

Since we consider a system with finite population, the offered FU traffic (in Erlangs) to the system denoted by ρ_f, is given by:

$$\rho_f = \frac{\sum_{\boldsymbol{x}} (M - N(\boldsymbol{x}))\lambda_f \pi(\boldsymbol{x})}{\mu}. \tag{8.14}$$

Table 8.1: Data rates achieved per RB as function of the SINR [BFDL+09]

#	SINR (dB)	R_i (in kbit/s/RB)
1	SINR $\geq$ 22.05	792.00
2	19.91 $\leq$ SINR $<$ 22.05	715.96
3	17.78 $\leq$ SINR $<$ 19.91	640.30
4	15.64 $\leq$ SINR $<$ 17.78	565.27
5	13.50 $\leq$ SINR $<$ 15.64	491.22
6	11.37 $\leq$ SINR $<$ 13.50	418.75
7	9.23 $\leq$ SINR $<$ 11.37	348.69
8	7.09 $\leq$ SINR $<$ 9.23	282.26
9	4.96 $\leq$ SINR $<$ 7.09	221.00
10	2.82 $\leq$ SINR $<$ 4.96	166.64
11	0.68 $\leq$ SINR $<$ 2.82	120.73
12	$-1.45 \leq$ SINR $<$ 0.68	84.09
13	$-3.59 \leq$ SINR <-1.45	56.54
14	$-5.73 \leq$ SINR <-3.59	36.93
15	$-7.86 \leq$ SINR <-5.73	23.60
16	$-10 \leq$ SINR <-7.86	14.85
17	$SINR < -10$	0.00

The offered FU traffic can also be expressed as:

$$\rho_f = M\frac{1/\mu}{1/\lambda_f + 1/\mu} = M\frac{\lambda_f}{\lambda_f + \mu}. \tag{8.15}$$

Unless otherwise stated, the arrival rate per idle FU is chosen to be $\lambda_f = 12.5\ s^{-1}$, the average channel holding time is $1/\mu = 10$ ms, the packet header length is $L_H = 500$ bits and the total packet size is $L = 4$ kbits. The FU population is $M = 8$. The offered FU traffic as function of these values is $\rho_f = 0.5$. The FAP average transmission power is $P_{TX} = 10$ dBm and the average transmission power consumed to monitor which channels are occupied by FUs is $P_s = 0$ dBm.

8.4.2 Numerical Results

In this subsection, we compare the MU maximum achievable throughput and the interruption probability obtained when the channels operated in OA mode have the highest data rate, i.e. $i = 1, \ldots, C_m$ and when they have the lowest data rate, i.e. $i = C + 1 - C_m, \ldots, C$.

In Figures 8.2 and 8.3, the MU maximum achievable throughput γ_T from (8.9) for different sets of open channels operated in OA mode is shown as a function of the packet size L. We can see that for the same value C_m of channels operated in OA mode, higher throughputs are achieved when the OA channels have the lowest data rate (Fig. 8.2) than when they have the highest data rate (Fig. 8.3). This can be explained as follows. The FUs use first the channels with the highest data rate and therefore there are more interruptions which reduce the contribution of these channels to the total throughput, despite having higher data rates. When C_m is small and the OA channels are the channels with the lowest data rates (Fig. 8.2), having one more OA channel leads to higher gains. When C_m is high, the gain of having one more OA channel is smaller because this channel is used by an FU with a higher probability. This effect is more significant for high L as longer MU packets experience more interruptions. The opposite occurs when the OA channels have the highest data rate (Fig. 8.3). Regarding the influence of the packet size, the achievable throughput has a maximum for a given L. This is due to the fact that for a smaller packet size L, more header information is transmitted, and for longer packet size L, there are more interruptions. Notice that when C_m is higher, this maximum throughput is achieved for smaller L in Fig. 8.2 and for higher L in Fig. 8.3.

The interruption probability ϵ_G obtained from (8.11) is shown in Fig. 8.4 and 8.5. It can be clearly seen that when the OA channels are the channels with the highest data rate (Fig 8.5) the interruption probability ϵ_G is higher than when the channels with the lowest data rate are chosen (Fig. 8.4). This happens because FUs use first channels with the highest data rates and hence, these channels experience more interruptions.

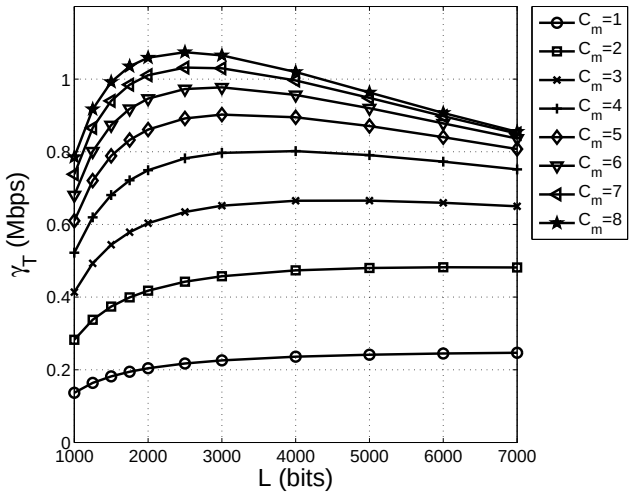

Figure 8.2: Maximum achievable throughput γ_T in Mbps. OA channels with the lowest data rates.

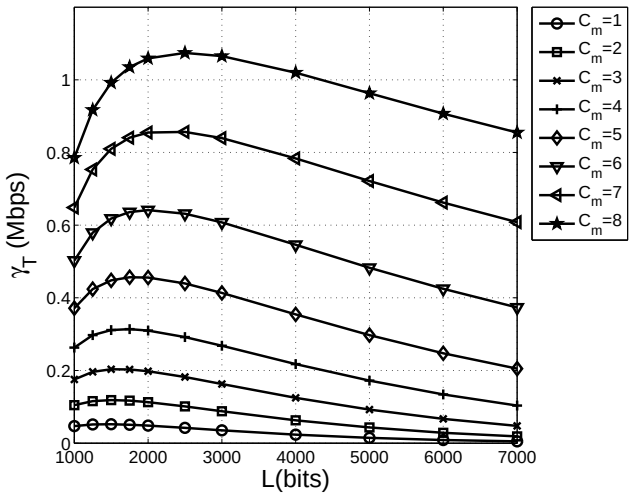

Figure 8.3: Maximum achievable throughput γ_T in Mbps. OA channels with the highest data rates.

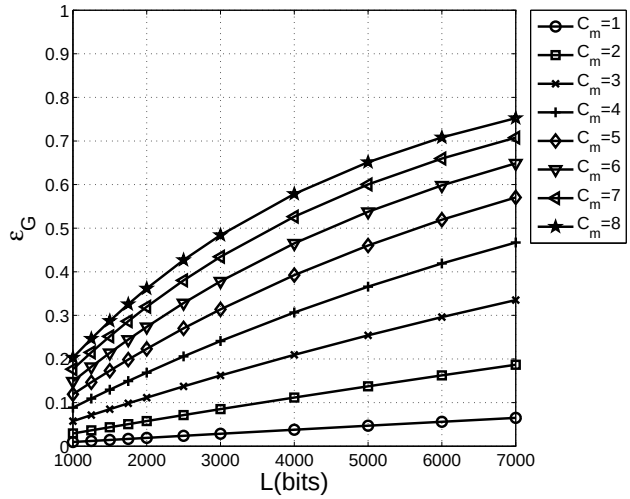

Figure 8.4: Interruption probability ϵ_G. OA channels with the lowest data rates.

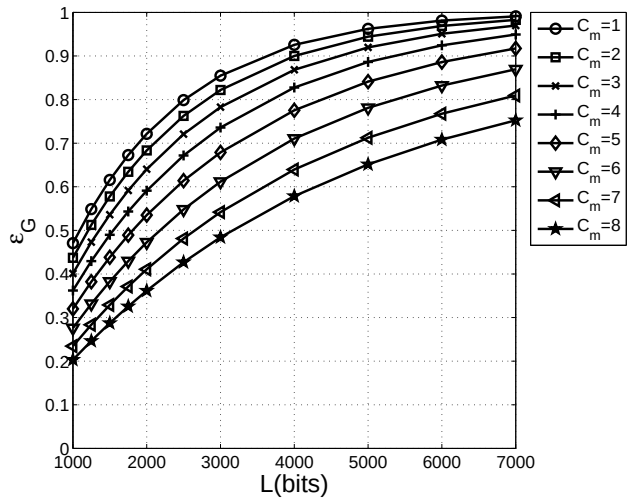

Figure 8.5: Interruption probability ϵ_G. OA channels with the highest data rates.

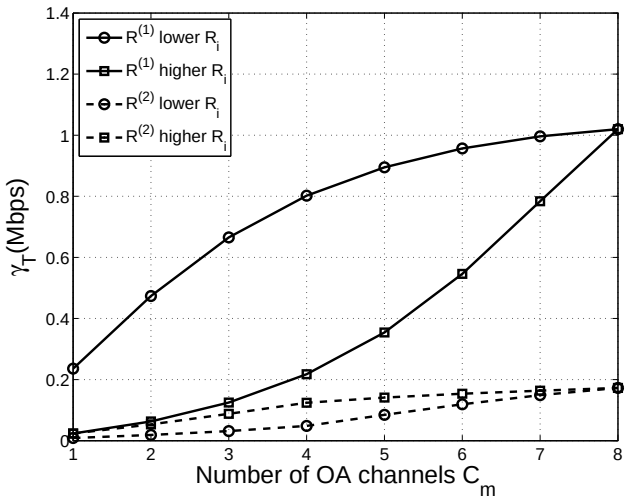

Figure 8.6: Maximum achievable throughput γ_T in Mbps for MU vs. C_m for different sets of data rates $R^{(1)}$ and $R^{(2)}$.

We can see that for this scenario, choosing the channels with the lowest data rate has better results but for other scenarios different results may be obtained. In Fig. 8.6, results considering different sets of data rates for each channel are shown. We consider $R^{(1)}$ as the set of data rates defined in Section 8.4.1 and $R^{(2)}$ are the set of data rates with values from Table 8.1 corresponding to rows #1, 3, 5, 7, 10, 12, 14 and 16 . We have $R_8^{(1)} = 0.356R_1^{(1)}$ and $R_8^{(2)} = 0.019R_1^{(2)}$. For $R^{(1)}$, it is better to operate in OA mode the channels with the lowest data rate. However, for $R^{(2)}$, the channels with the highest data rate yield better performance. This can be explained as follows. When the difference of data rates among channels is significant, the data rate achieved in the worst channels is too small, and it is better to access the best channels with higher data rates, despite having more interruptions.

When the set of channels operated in OA mode have the highest data rate, the performance only has better results when the difference of data rates among channels is very significant ($R^{(2)}$). Since common scenarios does not

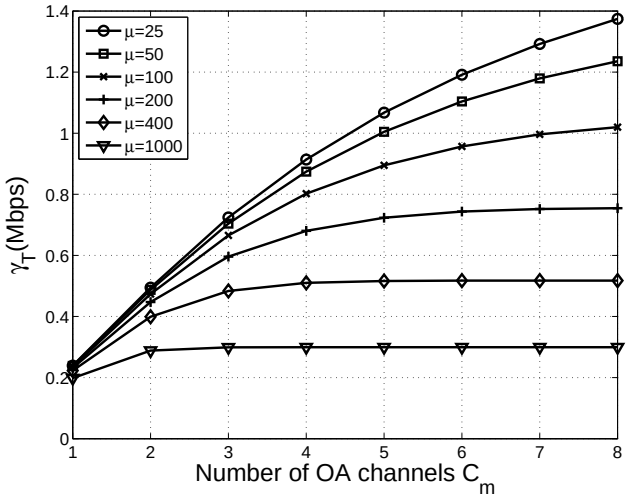

Figure 8.7: Maximum achievable throughput γ_T in Mbps for MU vs. C_m for different μ.

present these asymmetrical data rates, from now on $R^{(1)}$ is considered, and the set of channels operated in OA mode are considered to be the channels with the lowest data rate.

In Fig. 8.7, we show the maximum achievable throughput γ_T for MUs defined in (8.9) for different session durations while the offered traffic to the system ρ_f from (8.15) is kept constant. Even though ρ_f is kept constant, γ_T is higher when μ is small. For small μ the FUs are using the same channel for longer time. The system varies more slowly, there are less interruptions and therefore, the γ_T is higher. The opposite effect occurs for high μ. The FUs are using and releasing channels faster, the MUs experience more interruptions and therefore γ_T is lower. It can be seen that the number of C_m channels reaches a point at which considering one more channel operated in OA mode does not contribute to increase the throughput γ_T. This happens because the best channels are occupied and released continuously by the FUs, thus making these channels useless for MUs.

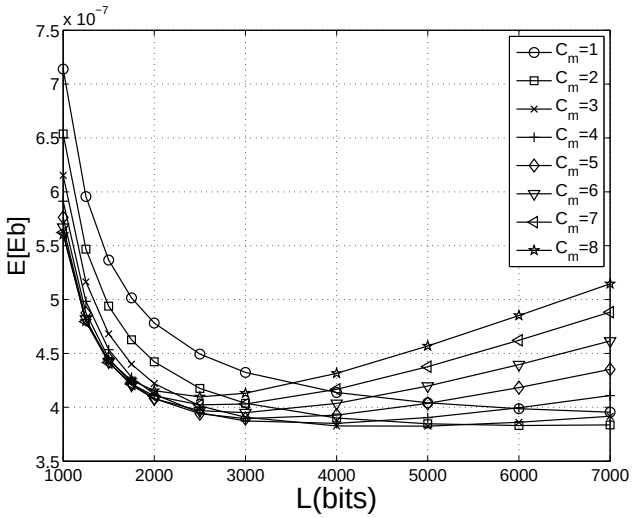

Figure 8.8: Consumed energy per successfully transmitted data bit $\overline{Eb}$ (J/bit) vs. L for different C_m.

In Fig. 8.8, the average consumed energy $\overline{Eb}$ per successfully transmitted data bit for MUs defined in (8.12) is shown. We can see that given a number of channels operated in OA mode C_m, there is a value of L which makes the $\overline{Eb}$ minimum. This happens because when small packet sizes L are considered, more energy is consumed by the header bits. On the other hand, when long L is considered, there are more interruptions and more energy is consumed by bits of packets that are not successfully transmitted. Note that the values of L which make $\overline{Eb}$ minimum, are close to the values of L for which the γ_T is maximum, as shown in Fig. 8.2. Regarding the influence of the number of channels operated in OA mode, C_m, given a value of L, the value of $\overline{Eb}$ first decreases with C_m, reaches a minimum and then increases again. This happens because for small C_m, the transmission of a bit takes longer since the OA channels have low data rates. For high C_m, more interruptions occur and more power is wasted, despite having high data rates.

8.5 Conclusions

In this chapter, we study a hybrid access control mode in femtocell networks. We consider a preemptive and non-resume access control policy for the MUs. Different data rates for each channel are considered depending on the SINR experienced by the users. We model the FU activity profile by a CTMC and we assess several performance parameters for MUs such as the maximum achievable throughput γ_T or the average consumed energy $\overline{Eb}$ per successfully transmitted data bit. We compute how many channels and which channels are the best channels to be operated in open access mode.

The results show that, if the SINR levels experienced by the users in each channel are comparable, and thus, the data rates achieved by the users in each channel are comparable too, the best channels to be operated in open access are the channels with the lowest data rates. Otherwise, if the data rates achieved by the best channels are significantly higher than the data rates achieved by the worst channels, it is better to operate the channels with the highest data rates in open access mode. In addition, we show that there is an optimal packet size for MU packets which maximizes the throughput γ_T and minimizes the average consumed energy $\overline{Eb}$ per successfully transmitted data bit. We also demonstrate that for short session durations, the number of channels operated in open access reaches a point at which having more channels operated in open access does not entail any gain to the MU throughput. These results motivate the need for novel resource management schemes which can dynamically adapt the set of open access channels to the channel conditions.

Chapter 9

AC in cognitive radio networks

9.1 Introduction

Many works have been devoted to AC policies and the mobility characteristics of terminals in the forthcoming 4G networks, but the scarcity of frequencies in the radio spectrum is still a hot topic. Today's cellular mobile networks are characterized by fixed spectrum assignment policies where the spectrum usage is concentrated on certain portions of the spectrum. Most of the spectrum which could be reasonably utilized for communications is licensed and these licenses are allocated for very long periods of time. As a result, a large portion of the assigned spectrum is used sporadically and that fact leads to an imbalance between the spectrum scarcity and spectrum underutilization. A study by the Federal Communication Commission (FCC) Spectrum Task Force [Com02] showed high temporal and geographic variations in the spectrum utilization, these variations range from 15% to 85% in the bands below 3 GHz. Although the fixed spectrum assignment policy generally served well in the past, the new paradigm of mobile cellular networks strains the effectiveness of the traditional spectrum policies and makes necessary to implement efficient methods to make use of the spectrum. The challenging task is how to efficiently use and share the radio spec-

trum, and thus the current research efforts in this field are devoted to the study of technologies that enable dynamic spectrum access. The *Cognitive Radio* (CR) [DAR03, MGM99, Hay05, C.-08, ALVM08, PPPMB09, MBPPP12] technology has been proposed as a concept which provides the ability to detect idle frequencies of a Primary Network (PN) which are not occupied by the Primary Users (PUs) and enables non-licensed users or Secondary Users (SUs) to use these idle bands in an opportunistic manner.

The CR technology improves the inefficient usage of the existing spectrum by allowing the PN to rent a specific number of channels to a Secondary Network (SN). The access of the SUs to the licensed bands must not interfere with the PUs. If an SU is using a licensed channel and a PU needs that channel, the SU undergoes a a spectrum handover, i.e., it vacates the channel and searches an idle channel among the channels which an SU can occupy. If there are not idle channels, the session of the SU is interrupted and aborted. Thus, the CR technology allows the PN obtain profit from its unused spectrum without comprising the services of the PUs.

This type of network imposes several challenges due to the interferences created between PUs and SUs, the diverse QoS requirements of applications and the necessity of guarantee seamless communications of the SUs regardless of the appearance of PUs. As a consequence new functionalities are required in order to manage the rented channels of the PN appropriately. Among others, an efficient channel sharing strategy, which decides which channels are rented and how the PUs and the SUs have access to them, and an AC policy for SUs should be in place.

In this chapter we consider a PN which rents only a set of channels and a SN which has a number of dedicated channels and can use opportunistically the rented set of channels of the PN. A similar scenario is proposed in [S.-09] but, among other assumptions, the PN rents all its channels and it does not consider any dedicated band for PUs. A SU can be blocked when it arrives at the system depending on the decision of the AC policy. If the SU is being served in a rented channel, it can be aborted when the PUs need their rented

channels. We propose different channel sharing strategies and we also obtain the optimal AC policy for each strategy depending on how much harmful is considered the forced termination of an ongoing SU session compared with blocking a new SU session. In order to find this optimal AC policy we use the theory of Markov Decision Processes (MDPs) [How60, Bel57, Ros70]. This work resulted into the publication in [BMPMB10a].

The rest of the chapter is organized as follows. In Section 9.2 the system model and three different channel sharing strategies between PUs and SUs are described. In Section 9.3, the MDP theory is described and applied to find an optimal AC policy under a given cost function, which is also proposed in this section. In Section 9.4 some numerical results are described and some concluding aspects are presented in Section 9.5.

9.2 System model and channel sharing strategies

In this section, the general system model of the CR network under study is presented. Next, three different channel sharing strategies are proposed and the specific analytical models used to evaluate their performance characteristics are presented.

9.2.1 System model

We consider a CR network with a PN and a SN. The PN has C_p channels which only PUs can occupy and C_r rented channels which can be occupied by SUs when they are idle. Thus, the total number of channels that the PUs have access to is $C_p + C_r$. Additionally, the SN has C_s dedicated channels that only can be occupied by SUs. Thus, the total number of channels that the SUs have access to is $N = C_s + C_r$. If an SU is using a rented channel and a PU needs that channel, the SU vacates the channel and searches an idle channel among the channels that an SU can occupy. If there are not idle channels among the channels the SU has access to, the session of the SU is aborted.

We make the common assumptions of Poisson arrival processes and exponentially distributed random variables for session durations. The arrival rate for primary (secondary) users is λ_p (λ_s) and the service rate of primary (secondary) users is exponentially distributed with rate μ_p (μ_s). The system state is described by the vector $\boldsymbol{x}$, where the definition of each vector dimension depends on the channel sharing strategy chosen. We consider that PUs have total priority over SUs which are using channels of the PN, and therefore a PU only is blocked when under its arrival all the channels of the PN are occupied by PUs. Under an arrival of an SU, its acceptance depends on the AC implemented for SUs. The function $a_s(\boldsymbol{x})$ denotes whether a new session of an SU is accepted by the AC policy or not when the system is in state $\boldsymbol{x}$. The value $a_s(\boldsymbol{x}) = 1$ means that the session is accepted and $a_s(\boldsymbol{x}) = 0$ means that the session is blocked. We use different finite QBD Markov processes [Neu81] for each channel sharing strategy to model the occupation of channels of PUs and SUs in the system. Next, we describe the three different channel sharing strategies proposed.

9.2.2 Channel sharing strategies

Strategy 1

The SN has a set of C_s dedicated channels which can be only used by SUs. The PN has a set of $C_p + C_r$ channels, of which C_r channels are rented and can be available to SUs. The channels available to SUs are not a fixed set, i.e., any channel of the set of the PN can be rented to SUs with the restriction that SUs cannot use more than C_r channels of the PN at the same time. This strategy is represented in Fig. 9.1. Incoming PUs choose the first channel not occupied by PUs, considering that channel 1 is the most preferred channel and channel $C_p + C_r$ is the least preferred channel. Incoming SUs occupy first their dedicated channels and if there are not available dedicated channels, SUs choose the first channel not occupied among the rented channels of the PN, choosing first the least preferred channels for PUs.

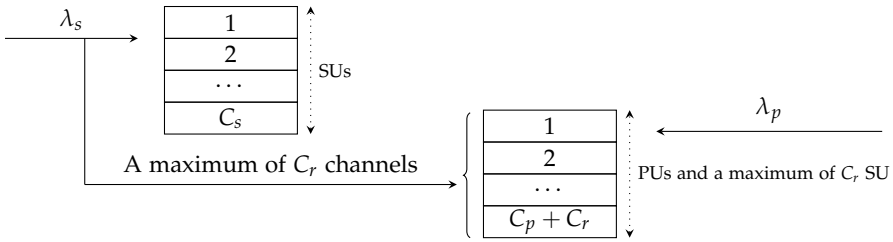

Figure 9.1: Strategy 1: Access of SUs to the PN

In this case, the system state vector is described by the 2-tuple $\boldsymbol{x} = (x_p, x_s)$, where x_p represents the total number of PUs in the PN and x_s represents the total number of SUs in the system, SN and PN. So that, the set of feasible states is thus given by

$$\begin{aligned} \mathcal{W}_1 := \{\boldsymbol{x} : \quad & x_p, x_s \in \mathbb{N}; \quad x_p + x_s \leq C_p + C_r + C_s; \\ & x_p \leq C_p + C_r; \quad x_s \leq C_s + C_r\}. \end{aligned} \tag{9.1}$$

Notice that when a PU arrives to the system, an SU session can be aborted due to the acceptance of the PU if the system is in the set of states given by

$$\mathcal{D}_1 := \{\boldsymbol{x} \in \mathcal{W}_1; \quad x_p + x_s = C_p + C_s + C_r; \quad x_s > C_s\}. \tag{9.2}$$

As an example, Figure 9.2 shows the transition diagram of the CTMC which models a system where $C_p = 2$, $C_s = 1$ and $C_r = 2$. For clarity, the notation has been simplified as $a_s(\boldsymbol{x}) = a_s$. The transition due to an aborted SU is represented for the diagonal transitions from state $(2,3)$ to $(3,2)$ and from state $(3,2)$ to $(4,1)$.

If we define levels as the number of PUs in the system, x_p, and phases as the number of SUs in the system, x_s, we can study the model as a finite level-dependent QBD process with $C_p + C_r + 1$ levels where levels with $x_p = 0, \cdots, C_p$ have $C_s + C_r + 1$ phases and levels with $x_p = C_p + i$ where $i = 1, \cdots, C_r$ have $C_s + C_r + 1 - i$ phases. Therefore we can construct the

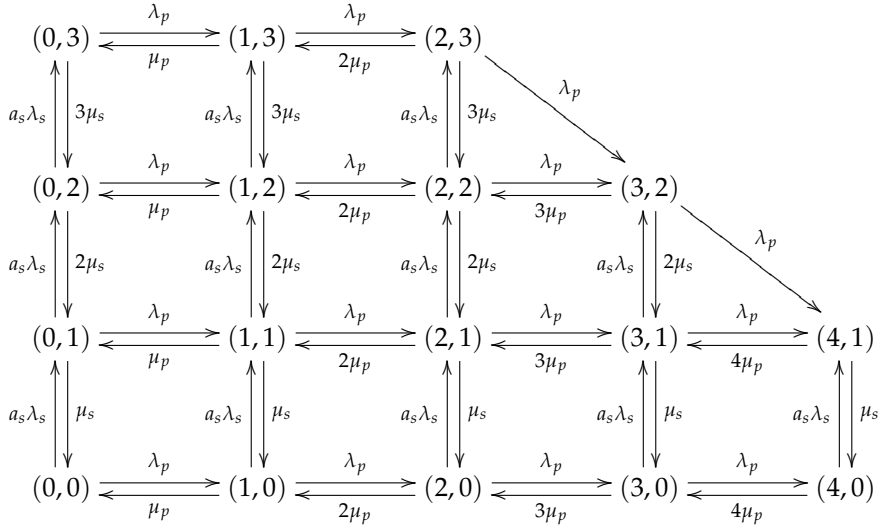

Figure 9.2: Transition diagram for strategy 1.

transition rate matrix Q_1 with a block-tridiagonal form, see (9.3). The first row of blocks corresponds to level $x_p = 0$, the second row of blocks to level $x_p = 1$, etc., where blocks $\boldsymbol{A}_1^{x_p}$ correspond to transitions among phases in level x_p, blocks $\boldsymbol{A}_0^{x_p}$ to transitions from level x_p to level $x_p + 1$ and blocks $\boldsymbol{A}_2^{x_p}$ to transitions from level x_p to level $x_p - 1$.

$$\boldsymbol{Q}_1 = \begin{bmatrix} \boldsymbol{A}_1^0 & \boldsymbol{A}_0^0 & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \cdots \\ \boldsymbol{A}_2^1 & \boldsymbol{A}_1^1 & \boldsymbol{A}_0^1 & \boldsymbol{0} & \boldsymbol{0} & \cdots \\ \boldsymbol{0} & \boldsymbol{A}_2^2 & \boldsymbol{A}_1^2 & \boldsymbol{A}_0^2 & \boldsymbol{0} & \cdots \\ & & \ddots & \ddots & \ddots & \\ \cdots & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{A}_2^{C_p+C_r-1} & \boldsymbol{A}_1^{C_p+C_r-1} & \boldsymbol{A}_0^{C_p+C_r-1} \\ \cdots & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{A}_2^{C_p+C_r} & \boldsymbol{A}_1^{C_p+C_r} \end{bmatrix} \tag{9.3}$$

Notice that the first $C_p + 1$ levels has the same phases, and thus:

$$\begin{aligned} \boldsymbol{A}_{\mathbf{1}}^{x_p} &= \boldsymbol{A}_{\mathbf{1}}^{\mathbf{0}} \qquad \text{where} \quad x_p = 1, \cdots, C_p\,, \\ \boldsymbol{A}_{\mathbf{0}}^{x_p} &= \boldsymbol{A}_{\mathbf{0}}^{\mathbf{0}} \qquad \text{where} \quad x_p = 1, \cdots, C_p - 1\,, \\ \boldsymbol{A}_{\mathbf{2}}^{x_p} &= x_p \cdot \boldsymbol{A}_{\mathbf{2}}^{\mathbf{1}} \quad \text{where} \quad x_p = 2, \cdots, C_p\,, \end{aligned} \tag{9.4}$$

where $\boldsymbol{A}_{\mathbf{1}}^{\mathbf{0}}$, $\boldsymbol{A}_{\mathbf{0}}^{\mathbf{0}}$ and $\boldsymbol{A}_{\mathbf{2}}^{\mathbf{1}}$ are $(C_s + C_r + 1)$ square matrices.

The matrices $\boldsymbol{A}_{\mathbf{1}}^{\boldsymbol{C_p+i}}$ are $(C_s + C_r + 1 - i)$ square matrices. The size of $\boldsymbol{A}_{\mathbf{0}}^{\boldsymbol{C_p-1+i}}$ is $(C_s + C_r + 2 - i) \times (C_s + C_r + 1 - i)$ and the size of $\boldsymbol{A}_{\mathbf{2}}^{\boldsymbol{C_p+i}}$ is $(C_s + C_r + 1 - i) \times (C_s + C_r + 2 - i)$. For more details, see Appendix B.3.

To solve the level-dependent finite QBD Markov process and obtain the stationary distribution $\pi(\boldsymbol{x})$, we use again the LLR algorithm defined in Appendix C.2. From the values of the stationary distribution and the state spaces $\mathcal{W}_1$ from (9.1) and $\mathcal{D}_1$ from (9.2), the blocking probability for SUs P_s^b, defined as the probability that an SU session which arrives to the system is blocked, and the dropping probability for SUs P_s^d, defined as the probability that a PU session which arrives to the system causes an SU session abortion, can be calculated as follows:

$$P_s^b = \sum_{\boldsymbol{x} \in \mathcal{W}_1} (1 - a_s(\boldsymbol{x}))\pi(\boldsymbol{x}), \tag{9.5}$$

$$P_s^d = \sum_{\boldsymbol{x} \in \mathcal{D}_1} \pi(\boldsymbol{x}). \tag{9.6}$$

Strategy 2

The SN has a set of C_s dedicated channels which can be only used by SUs. The PN has a set of C_p dedicated channels only available for PUs and a set of C_r rented channels which can also be used opportunistically by SUs, thus the total number of channels in the PN is $C_p + C_r$. The rented channels are a fixed set of channels. The access of SUs to the PN is represented in Fig. 9.3. Incoming PUs choose the first channel no occupied by PUs, considering that

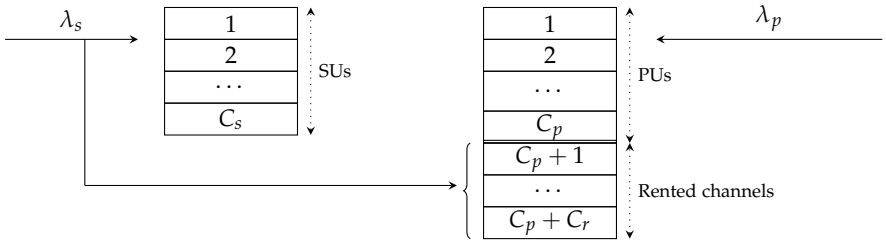

Figure 9.3: Strategies 2 and 3: Access of SUs to the PN

channel 1 is the most preferred channel and channel $C_p + C_r$ is the least preferred channel. Incoming SUs occupy first their dedicated channels and if there are not available dedicated channels, SUs choose the channels not occupied among the set of C_r rented channels of the PN, choosing first the least preferred channels for PUs. In this strategy we consider that there is no repacking for PUs, i.e. if a PU is using one of the C_r rented channels and one of the C_p dedicated channels of the PN becomes idle, the PU remains in the rented set of channels despite having idle channels in the dedicated set.

The system state vector is described by the 3-tuple $\boldsymbol{x} = (x_p, x_r, x_s)$, where x_p represents the number of PUs in the dedicated channels of the PN, x_r represents the number of PUs in the rented channels of the PN, and x_s represents the total number of SUs in the system, i.e. in the PN and the SN. So that, the set of feasible states is thus given by

$$\begin{aligned}\mathcal{W}_2 := \{\boldsymbol{x} : \quad & x_p, x_r, x_s \in \mathbb{N}; \quad x_p + x_r + x_s \leq C_p + C_r + C_s; \\ & x_p \leq C_p; \quad x_r \leq C_r; \quad x_r + x_s \leq C_s + C_r\}.\end{aligned} \tag{9.7}$$

Notice that when a new PU arrives to the system, an SU session can be aborted due to the acceptance of the PU if the system is in the set of states given by

$$\mathcal{D}_2 := \{\boldsymbol{x} \in \mathcal{W}_2; \quad x_p + x_r + x_s = C_p + C_s + C_r; \quad x_s > C_s\}. \tag{9.8}$$

A 3-dimensional finite QBD Markov process is used to model the system. We define high-levels as x_p, low-levels as x_r and phases as x_s. Then, the QBD process has $C_p + 1$ high-levels where all of them have $C_r + 1$ low-levels and each low-level x_r has $C_s + C_r + 1 - x_r$ phases. Therefore, we can construct the transition rate matrix $\boldsymbol{Q}_2$ with a block-tridiagonal form, see (9.9). The first row of blocks corresponds to high-level $x_p = 0$, the second row of blocks to high-level $x_p = 1$, etc., where blocks $\boldsymbol{B}_1^{x_p}$ correspond to transitions among low-levels in high-level x_p, blocks $\boldsymbol{B}_0^{x_p}$ to transitions from high-level x_p to high-level $x_p + 1$ and blocks $\boldsymbol{B}_2^{x_p}$ to transitions from high-level x_p to high-level $x_p - 1$.

$$\boldsymbol{Q}_2 = \begin{bmatrix} \boldsymbol{B}_1^0 & \boldsymbol{B}_0^0 & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \cdots \\ \boldsymbol{B}_2^1 & \boldsymbol{B}_1^1 & \boldsymbol{B}_0^1 & \boldsymbol{0} & \boldsymbol{0} & \cdots \\ \boldsymbol{0} & \boldsymbol{B}_2^2 & \boldsymbol{B}_1^2 & \boldsymbol{B}_0^2 & \boldsymbol{0} & \cdots \\ & & \ddots & \ddots & \ddots & \\ \cdots & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{B}_2^{C_p-1} & \boldsymbol{B}_1^{C_p-1} & \boldsymbol{B}_0^{C_p-1} \\ \cdots & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{B}_2^{C_p} & \boldsymbol{B}_1^{C_p} \end{bmatrix} \tag{9.9}$$

All the high-levels have the same state transitions but the last high-level $x_p = C_p$ due to SUs forced terminations, then:

$$\begin{aligned} \boldsymbol{B}_1^{x_p} &= \boldsymbol{B}_1^0 \qquad \text{where} \quad x_p = 1, \cdots, C_p - 1\,, \\ \boldsymbol{B}_0^{x_p} &= \boldsymbol{B}_0^0 \qquad \text{where} \quad x_p = 1, \cdots, C_p - 1\,, \\ \boldsymbol{B}_2^{x_p} &= x_p \cdot \boldsymbol{B}_2^1 \quad \text{where} \quad x_p = 2, \cdots, C_p\,. \end{aligned} \tag{9.10}$$

Likewise, the matrices $\boldsymbol{B}_1^0$ and $\boldsymbol{B}_1^{C_p}$ have a block-tridiagonal form, where the first row of blocks corresponds to low-level $x_r = 0$, the second row of blocks to low-level $x_r = 1$, etc., where blocks in the main diagonal correspond to transitions among phases, blocks in the upper diagonal to transitions to lower low-levels and blocks in the lower diagonal correspond to transitions to higher low-levels. For more details, see Appendix B.3.

To solve this finite QBD Markov process and obtain the stationary distribution $\pi(\boldsymbol{x})$ we use again the LLR algorithm. From the values of the stationary distribution and the state spaces $\mathcal{W}_2$ from (9.7) and $\mathcal{D}_2$ from (9.8), the blocking probability for SUs P_s^b and the dropping probability for SUs P_s^d can be calculated as follows:

$$P_s^b = \sum_{\boldsymbol{x}\in\mathcal{W}_2} (1 - a_s(\boldsymbol{x}))\pi(\boldsymbol{x}), \tag{9.11}$$

$$P_s^d = \sum_{\boldsymbol{x}\in\mathcal{D}_2} \pi(\boldsymbol{x}). \tag{9.12}$$

Strategy 3

This strategy is similar to strategy 2 but using repacking for PUs., i.e. if a PU is using one of the C_r rented channels and one of the C_p dedicated channels of the PN becomes idle, the PU leaves the channel that was using and occupies the channel in the dedicated set of channels which became idle. The access of SUs to the PN is also represented in Fig. 9.3.

The system state vector is described by the 3-tuple $\boldsymbol{x} = (x_p, x_r, x_s)$, where again, x_p represents the number of PUs in the dedicated channels of the PN, x_r represents the number of PUs in the rented channels of the PN and x_s represents the total number of SUs in the system. So that, the set of feasible states is thus given by:

$$\begin{aligned}\mathcal{W}_3 := \{\boldsymbol{x} : \quad & x_p, x_r, x_s \in \mathbb{N}; \quad x_p + x_r + x_s \leq C_p + C_r + C_s; \\ & x_p \leq C_p; \quad x_r + x_s \leq C_s + C_r; \\ & x_r \leq C_r; \quad x_r = 0 \Leftarrow x_p < C_p\}.\end{aligned} \tag{9.13}$$

When a new PU arrives to the system, an SU session can be aborted due to the acceptance of the PU if the system is in the set of states given by:

$$\mathcal{D}_3 := \{\boldsymbol{x} \in \mathcal{W}_3; \quad x_r + x_s = C_s + C_r; \quad x_s > C_s\}. \tag{9.14}$$

The repacking occurs if a PU session which is served in the dedicated channels terminates its service in the set of states:

$$\mathcal{R}_3 := \{\boldsymbol{x} \in \mathcal{W}_3; \quad x_p = C_p; \quad x_r \neq 0\}. \tag{9.15}$$

Again, a 3-dimensional finite QBD Markov process is used to model the system. We define high-levels as x_p, low-levels as x_r when $x_p = C_p$, otherwise there are not low-levels, and phases as x_s. Then, the QBD process has $C_p + 1$ high-levels where only the high-level $x_p = C_p$ has $C_r + 1$ low-levels each one with $C_s + C_r + 1 - x_r$ phases. The high-levels $x_p \neq C_p$ have $C_s + C_r + 1$ phases. Therefore we can construct the transition rate matrix $\boldsymbol{Q}_3$ with a block-tridiagonal form, see (9.16). The first row of blocks corresponds to high-level $x_p = 0$, the second row of blocks to high-level $x_p = 1$, etc., where blocks $\boldsymbol{C}_1^{x_p}$ correspond to transitions among low-levels in high-level $x_p = C_p$ and among phases in $x_p \neq C_p$, blocks $\boldsymbol{C}_0^{x_p}$ corresponds to transitions from high-level x_p to high-level $x_p + 1$ and blocks $\boldsymbol{C}_2^{x_p}$ to transitions from high-level x_p to high-level $x_p - 1$.

$$\boldsymbol{Q}_3 = \begin{bmatrix} \boldsymbol{C}_1^0 & \boldsymbol{C}_0^0 & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \cdots \\ \boldsymbol{C}_2^1 & \boldsymbol{C}_1^1 & \boldsymbol{C}_0^1 & \boldsymbol{0} & \boldsymbol{0} & \cdots \\ \boldsymbol{0} & \boldsymbol{C}_2^2 & \boldsymbol{C}_1^2 & \boldsymbol{C}_0^2 & \boldsymbol{0} & \cdots \\ & & \ddots & \ddots & \ddots & \\ \cdots & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{C}_2^{C_p-1} & \boldsymbol{C}_1^{C_p-1} & \boldsymbol{C}_0^{C_p-1} \\ \cdots & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{C}_2^{C_p} & \boldsymbol{C}_1^{C_p} \end{bmatrix} \tag{9.16}$$

Clearly, All the high-levels have the same state transitions but the last high-level $x_p = C_p$ due to SUs forced terminations, then:

$$\begin{aligned} \boldsymbol{C}_1^{x_p} &= \boldsymbol{C}_1^0 \qquad \text{where} \quad x_p = 1, \cdots, C_p - 1\,, \\ \boldsymbol{C}_0^{x_p} &= \boldsymbol{C}_0^0 \qquad \text{where} \quad x_p = 1, \cdots, C_p - 2\,, \\ \boldsymbol{C}_2^{x_p} &= x_p \cdot \boldsymbol{C}_2^1 \quad \text{where} \quad x_p = 2, \cdots, C_p - 1\,. \end{aligned} \tag{9.17}$$

Likewise, the matrix $\boldsymbol{C}_{\mathbf{1}}^{C_p}$ has a block-tridiagonal form, where the first row of blocks corresponds to low-level $x_r = 0$, the second row of blocks to level $x_r = 1$, etc., where blocks in the main diagonal correspond to transitions between phases, blocks in the upper diagonal to transitions to lower low-levels and blocks in the lower diagonal correspond to transitions to higher low-levels. For more details, see Appendix B.3.

In order to obtain the stationary distribution $\pi(\boldsymbol{x})$, we use again the LLR algorithm. From the values of the stationary distribution and the state spaces $\mathcal{W}_3$ from (9.13) and $\mathcal{D}_3$ from (9.14), the blocking probability for SUs P_s^b and the dropping probability for SUs P_s^d are given by

$$P_s^b = \sum_{\boldsymbol{x} \in \mathcal{W}_3} (1 - a_s(\boldsymbol{x}))\pi(\boldsymbol{x}), \tag{9.18}$$

$$P_s^d = \sum_{\boldsymbol{x} \in \mathcal{D}_3} \pi(\boldsymbol{x}). \tag{9.19}$$

In addition, the repacking rate denoted by κ is defined as the number of PU ongoing sessions per second which being served in the rented set of channels leave these channels and finish their service in the dedicated set of channels of the PN. Knowing the state space $\mathcal{R}_3$ from (9.15), κ is given by

$$\kappa = \sum_{\boldsymbol{x} \in \mathcal{R}_3} C_p \mu_p \pi(\boldsymbol{x}). \tag{9.20}$$

9.3 Markov decision processes and optimal AC policies

In this section, we describe in general terms the MDP theory applied in order to find an optimal AC policy given a cost function. Next, we present the specific optimization problem under study and the cost function considered to obtain the corresponding optimal AC policy.

9.3.1 Markov decision processes

The Markov process theory assumes that the system, the states and the transitions are given in advance. The problem is to find the stationary probabilities of the system and then deduce interesting performance parameters, such as blocking probabilities or dropping probabilities. However, in some scenarios the behavior of the system is not defined in advance. The possible actions made in a state such as accept or reject an arrival or the transitions from one state to another, are not defined in advanced and depend on some choices defined by the operation policy. This type of process is called Markov Decision Process (MDP) [Ros70] and the problem is to find an optimal policy that depends on a given objective function, such that the expected revenues are maximized or the expected costs are minimized, it does not matter which problem is formulated. From now on, we consider the problem of minimizing the expected costs given by the objective function.

The theory of MDPs studies decision problems when the stochastic behavior of the system can be defined as a Markov process. It combines Dynamic Programming [Bel57] and Markov process theory [How60]. In MDPs, when the system is in a state, a decision can be made, which may incur an immediate cost and, in addition, affects next transitions. Under Markovian assumptions, the action to be chosen in each state depends only on the state itself, and generally, a policy (optimal or not) defines for each state the action to be chosen. The action associated with each state determines the transition probabilities of the next transitions and these probabilities depend only on the state. Hence, each policy defines a different Markov process. The solution of the problem is to find the Markov process which has minimum average cost. The MDPs are classified in discrete time and continuous time decision processes. From now on, we consider continuous time MDPs.

When the system is in state x, an action α_x which belongs to the set A_x, $\alpha_x \in A_x$, has to be chosen among all possible actions in state x. The action chosen in each state x among the set of possible actions is defined by the policy a, then $\alpha_x = \alpha_x(a)$. Action $\alpha_x(a)$ incurs an immediate cost $\gamma_x(\alpha_x(a))$.

If the cost is stochastic, the value of $\gamma_x(\alpha_x(a))$ denotes its mean. At the next instant, the system moves into a new state y with a transition rate denoted by $q_{xy}(\alpha_x(a))$ which depends on the action chosen in state x. Since the transition rates do not depend on how the state x has been reached, time homogeneous systems are considered, where $\gamma_x(\alpha_x(a))$ and $q_{xy}(\alpha_x(a))$ do not depend on the time. The cost $\gamma_x(\alpha_x(a))$ and the transition rates $q_{xy}(\alpha_x(a))$ are functions of the policy a and of the state x. For brevity, we will denote them by $\gamma_x(a)$ and $q_{xy}(a)$.

Given the policy a, the transition rates $q_{xy}(a)$ are fixed and the Markov process has a stationary distribution $\pi_x(a)$. Then, the average cost $\gamma(a)$, i.e. the expected cost rate, is given by:

$$\gamma(a) = \sum_x \pi_x(a)\gamma_x(a). \tag{9.21}$$

Now, the objective is to find the optimal policy a^*, which minimizes the average cost and therefore:

$$\gamma(a^*) \leq \gamma(a), \quad \forall a. \tag{9.22}$$

Since the definition of a policy is discrete, a discrete optimization problem has to be solved. The average cost can be calculated for each possible policy, but the solution is not quite straightforward and some systematic approach is needed. Several approaches have been introduced in the literature, such as policy iteration, value iteration or linear programming approaches. We will focus on the policy iteration approach. Let us introduce first the Howard's equation which determines the problem to be solved by the policy iteration approach.

Howard's equation

The value of $\gamma(a)$ is the average cost rate under policy a. Let $V_x(t, a)$ be the expected cumulative cost in the interval $(0, t)$ (integral of the cost rate over

time), when the system starts from state x. Then, the relative value $v_x(a)$ of state x, given by

$$v_x(a) = \lim_{t\to\infty} \left(V_x(t,a) - t\cdot\gamma(a)\right), \tag{9.23}$$

denotes how much greater the expected cumulative cost over an infinite time horizon is in comparison with the average cumulative cost when the system starts from the initial state x.

The relative values of states $v_x(a)$, for a given policy a satisfy the Howard's equations [How60]:

$$\gamma_x(a) - \gamma(a) + \sum_{y\neq x} q_{xy}(a)(v_y(a) - v_x(a)) = 0, \quad \forall x \tag{9.24}$$

where the policy a explicitly determines the transition rates $q_{x,y}(a)$, and then $v_x(a)$ and $\gamma(a)$ can be determined by solving these equations.

Note that only the differences $v_y(a) - v_x(a)$ appear in the equation (9.24). If the same constant is added to the relative values of all states x, $v_x(a)$, the equation remains satisfied. The relative values will be determined up to an undetermined additive constant. Hence, we can arbitrarily set, for example, $v_1(a) = 0$. The number of unknown $v_x(a)$ is one less the number of equations, but the average cost rate $\gamma(a)$ is also unknown and thus, there are as many equations as unknown variables.

Note also that the solution $\gamma(a)$ obtained from the Howard's equation is the same as the average cost rate obtained by (9.21). This can be seen clearly by multiplying the Howard equation by $\pi_x(a)$ and summing all the Howard equations given by each state x. For simplicity when showing the equations, we omit the dependence on the policy a in the proof.

$$\sum_x \gamma_x \pi_x - \gamma \sum_x \pi_x + \sum_x \pi_x \left(\sum_{y\neq x} q_{xy} v_y - \overbrace{\sum_{y\neq x} q_{xy}}^{=-q_{xx}} v_x \right) = 0,$$

$$\sum_x \gamma_x \pi_x - \gamma \sum_x \pi_x + \sum_x \pi_x \left(\sum_y q_{xy} v_y \right) = 0,$$

$$\sum_x \gamma_x \pi_x - \gamma \underbrace{\sum_x \pi_x}_{=1} + \sum_y \underbrace{\left(\sum_x \pi_x q_{xy} \right)}_{=0} v_y = 0,$$

finally,

$$\sum_x \gamma_x \pi_x - \gamma = 0, \tag{9.25}$$

and (9.25) is the same as (9.21).

Policy iteration

The policy iteration approach can be divided in two stages.

- *Evaluation Stage* (ES)

  The policy iteration is started with an initial policy a and the relative values $v_x(a)$ and the average cost rate $\gamma(a)$ are calculated from the Howard's equation.

- *Improvement Stage* (IS)

  The initial policy can be improved by choosing the action α_x in each state as follows:

$$\alpha_x = \min_{\alpha} \left\{ \gamma_x(\alpha) - \gamma(a) + \sum_{y \neq x} q_{xy}(\alpha)(v_y(a) - v_x(a)) \right\}. \tag{9.26}$$

After the choices made in the improvement stage a new policy a' is defined. Then, new values of $v_x(a')$ and $\gamma(a')$ can be calculated from the Howard's equation in the next evaluation stage and a new policy can be determined in the improvement stage. This iteration is continued until nothing changes. The idea is that the decision made in state x minimizes the expected cost by considering the immediate cost of the action and its influence on the

next transition, but from that point on assuming that all the decisions are made considering the old policy. We can summarize the process as follows:

$$\begin{gathered} a_0 \overset{ES}{\rightarrow} v(a_0), \gamma(a_0) \overset{IS}{\rightarrow} a_1 \overset{ES}{\rightarrow} v(a_1), \gamma(a_1) \overset{IS}{\rightarrow} a_2 \cdots \\ \cdots \quad \text{until} \quad a_{n+1} = a_n. \end{gathered} \tag{9.27}$$

Finally, the optimal policy is obtained $a^* = a_n$. Generally, the police iteration converges quickly.

9.3.2 Optimal AC policy

The AC policy implemented for SUs has not been defined yet. Remember that the function $a_s(\boldsymbol{x})$ determines whether a new SU session is accepted or not when arrives to the system depending on the state of the system in that moment. When the stochastic behavior of the system can be described as a Markov process, the MDP theory can be applied to study sequential decision problems. Thus, for each channel sharing strategy described in the previous section, we can determine a different optimal AC policy for SUs under a given cost rate function modeling the systems as MDPs. We consider stationary deterministic Markovian policies, which define the next decision based only on the current state.

In the MDP theory when a decision is made (accepting or not a new SU) the system is penalized with some immediate cost. In our system, the optimization problem is formulated as the minimization of the average cost rate per time unit. If $a_s(\boldsymbol{x})$ is the AC policy, we denote the average cost rate by $\gamma(a_s)$ and consider the problem of finding the policy a_s^* that minimizes $\gamma(a_s)$, which we name the optimal policy. From the perception of users, dropping a session in progress is generally considered to be more harmful than blocking a new session. But blocking too many new SU sessions decreases the efficiency of the CR network since the benefits given by allowing SUs to rent channels from the PN are not exploited. The main goal of an efficient AC policy for SUs is then, to find a trade-off between these two conflicting re-

quirements. Thus, the cost structure has been chosen so that the cost rate represents a weighted sum of loss rates for new blocking SUs and for ongoing SUs whose service is aborted by the arrival of a PU that needs the rented channels used by the SU.

Let $I(\boldsymbol{x})$ denote the indicator function which takes the value 1 when all the channels of the PN are occupied, with at least one of these channels being occupied by an SU, and all the secondary dedicated channels are also occupied, otherwise it takes the value 0. In other words, $I(\boldsymbol{x})$ indicates in which states an SU, which is renting channels from the PN, is dropped from the system under an arrival of a PU due to the lack of other idle channels in the system. It can be written as:

$$I(\boldsymbol{x}) = \begin{cases} 1 & \text{if} \quad \boldsymbol{x} \in \mathcal{D} \\ 0 & \text{otherwise,} \end{cases} \tag{9.28}$$

where $\mathcal{D}$ is the state space which corresponds to $\mathcal{D}_1$, $\mathcal{D}_2$ or $\mathcal{D}_3$ defined in (9.2), (9.8) and (9.14) respectively, depending on the channel sharing strategy 1, 2 or 3 considered.

Considering all the facts pointed out above, the cost rate at state $\boldsymbol{x}$ under policy $a_s(\boldsymbol{x})$ can be defined as:

$$\gamma_x(a_s) = (1 - a_s(\boldsymbol{x}))\lambda_s + w\lambda_p I(\boldsymbol{x}), \tag{9.29}$$

where, remember, $a_s(\boldsymbol{x}) = 1$ when the new SU session is accepted and 0 otherwise, and w is a weight that determines how much harmful it is to abort an ongoing session of an SU compared with blocking a new session of an SU. The first part of the equation refers to the cost associated with blocking a new SU session and the second part is associated with dropping an ongoing SU session due to an arrival of a PU when there are not idle channels (states given by $I(\boldsymbol{x})$).

Once the cost rate function is defined, the AC optimization problem can be conducted applying the MDP theory. The relative values of state $\boldsymbol{x}$, $v_x(a_s)$,

and the average cost rate, $\gamma(a_s)$, can be calculated solving the system of equation defined by the Howard's equations (9.24), setting $v_{x_1}(a_s) = 0$, where x_1 is $(0,0)$ for strategy 1 and $(0,0,0)$ for strategies 2 and 3. Then, the policy iteration approach summarized in expression (9.27) can be applied. We choose as the initial AC policy the CS policy and from that point, the AC policy is improved step by step according to the cost rate function, until the AC policy remains the same in two consecutive steps.

9.4 Numerical evaluation

In this section we study the variation of the optimal AC policy for different values of the weight of the cost rate function, w, and the blocking and dropping probabilities for SUs (P_s^b and P_s^d respectively) as a function of different system parameters.

For the numerical examples we consider, unless otherwise indicated, a system where the PN has 11 channels, of which 6 are exclusively dedicated to PUs, $C_p = 6$, and 6 channels can be used opportunistically by SUs, $C_r = 5$. In addition, the SN has 2 dedicated channels for SUs, $C_s = 2$. For PUs, the arrival rate is $\lambda_p = 3.2$ and the service rate $\mu_p = 0.5$. For SUs, the arrival rate is $\lambda_s = 1.8$ and the service rate $\mu_s = 1.25$. The weight which defines the cost rate function is set to $w = 5$. A summary of all model parameters is given in Table 9.1.

Table 9.1: Definition of system parameters.

Parameter	Value	Parameter	Value
C_p	6	C_s	2
C_r	5	w	5
λ_p	3.2	μ_p	0.5
λ_s	1.8	μ_s	1.25

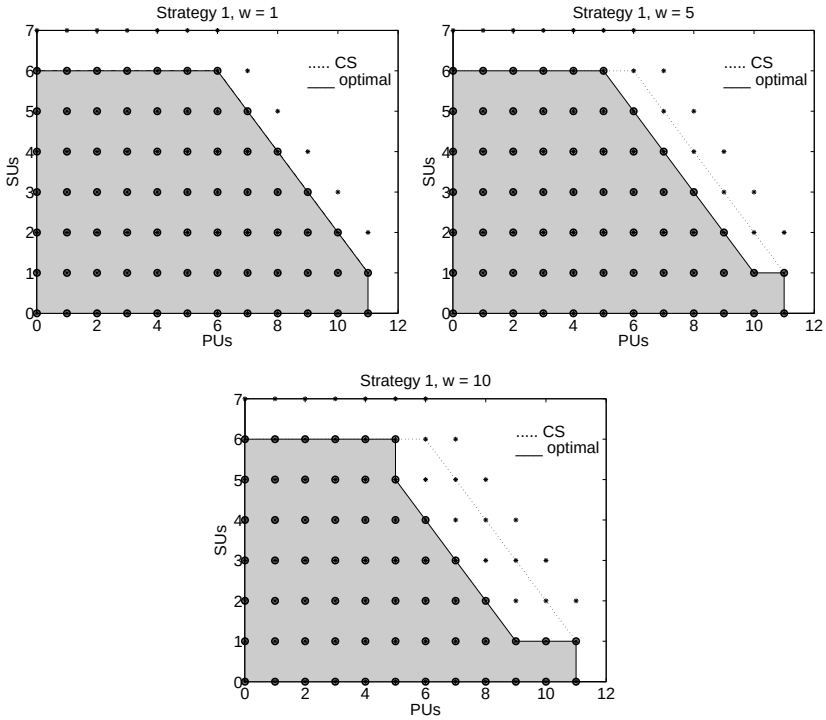

Figure 9.4: Optimal AC policy for strategy 1.

Results in Fig. 9.4 show the optimal AC policy for strategy 1 under the cost rate function defined in (9.29) for different values of the weight w, which determines the cost of dropping an ongoing SU session compared with blocking a new SU session. Each figure corresponds with one value of the weight, $w = 1$, $w = 5$ and $w = 10$. The possible states which the system can take are given by the number of PUs (x axis) and SUs (y axis) and they are indicated by markers. Under a new SU arrival, when the optimal AC policy is applied a new SU session is accepted if the system is in the states inside the shady area and indicated by circles. The CS policy is also indicated in this figure considering also strategy 1. When the CS policy is applied, a new SU session is accepted as long as there are available channels. The states where a new SU

session is accepted under CS policy are the states inside the area indicated by the dotted line. As expected, for higher values of w, a forced termination of an ongoing SU session is considered more harmful and thus, not accepting a new SU can be better for a given set of states despite being available channels for it. Therefore, the optimal AC policy is more restrictive with the new SUs and the shady area is smaller for higher w.

It is worth noting that the strategies 1 and 3 are equivalent when the blocking probability for SUs, P_s^b, and the dropping probability for SUs, P_s^d, are calculated. This is because SUs see the same system in terms of channel occupancy for both strategies due to the repacking mechanism implemented in strategy 3. In the figures plotted from now on, results for the three strategies are shown, but strategy 1 and strategy 3 overlap, and they are considered as one line. The differences between these two strategies lies in interference and management aspects more than in these performance results. The three-dimensional analytical model for strategy 3 is more complex than the two-dimensional analytical model of strategy 1, however, it is still interesting since it allows us to calculate exclusive parameters of strategy 3 like the repacking rate κ.

In Fig. 9.5 and 9.6, the blocking probability, P_s^b, and the dropping probability, P_s^d, respectively, for SUs are shown as a function of the weight w. The results are displayed for the optimal AC policy for strategies 1 and 3 in a solid line, the optimal AC policy for strategy 2 in a dashed line and the CS policy for strategy 1 in a dotted line. As expected, the results for the CS policy do not depend on w. The results obtained considering strategies 1, 2 and 3 have a staircase shape due to the dependency of the optimal AC policy with the value of the weight w. Also as expected, SU blocking and dropping probabilities have opposite behavior. For higher w, P_s^b is higher and P_s^d is lower. Moreover, when w is very low means that dropping a SU session has not a high cost and therefore the CS policy and the optimal AC policy, both for strategy 1, have the same results. Notice also that the differences between strategy 2 and the other two strategies is higher for lower values of w. From now on, we consider $w = 5$.

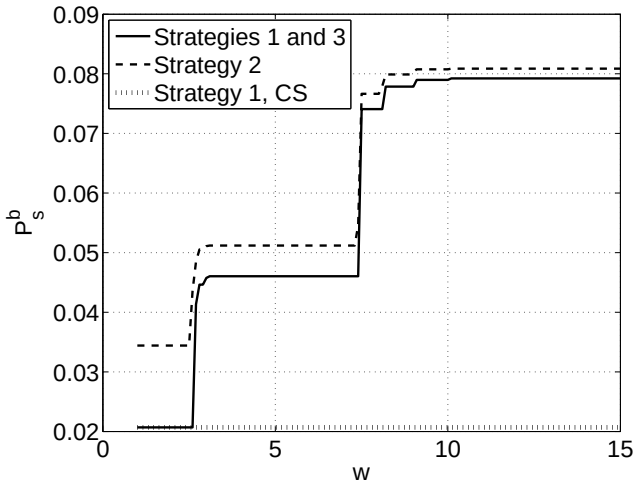

Figure 9.5: SU blocking probability as a function of w.

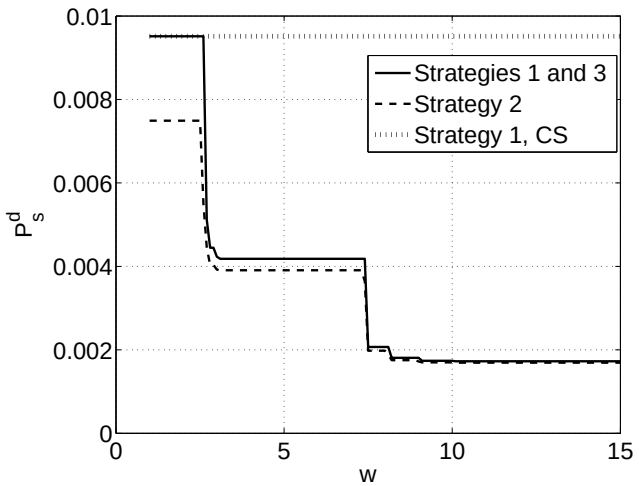

Figure 9.6: SU dropping probability as a function of w.

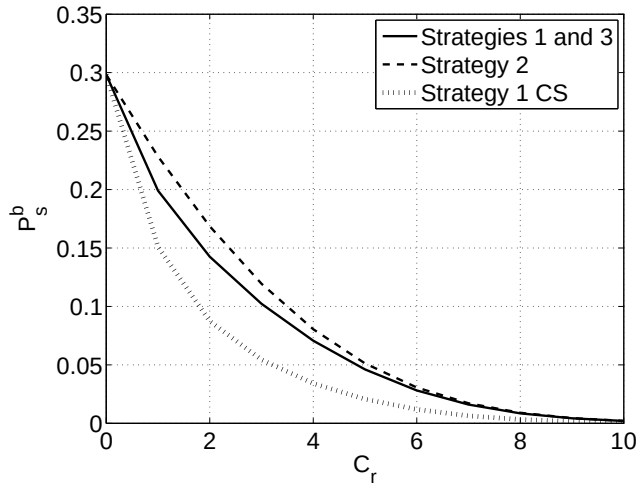

Figure 9.7: SU blocking probability as a function of C_r.

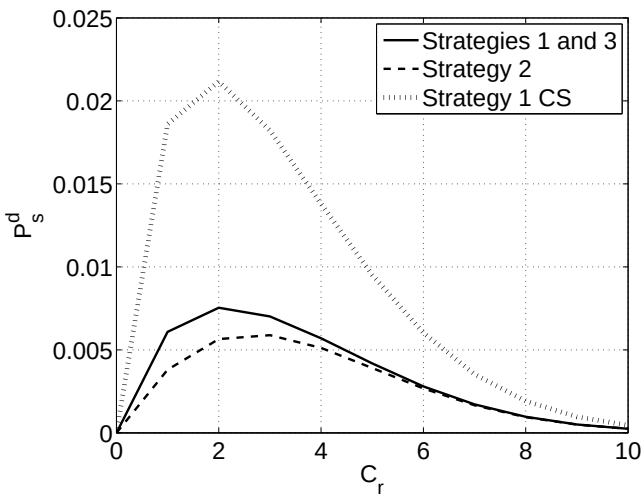

Figure 9.8: SU dropping probability as a function of C_r.

In Fig. 9.7 and 9.8, P_s^b and P_s^d, respectively, are shown as a function of the number of rented channels of the PN, C_r. Again, the results are displayed for the optimal AC policy for strategies 1 and 3 in a solid line, for strategy 2 in a dashed line and the CS policy for strategy 1 is shown in a dotted line. We can see that the CS policy for strategy 1 has lower P_s^b and higher P_s^d than the optimal AC for strategy 1. This is because the optimal AC policy takes into account the cost of dropping an ongoing SU session and considers that is more optimal to block more SU sessions in order to decrease P_s^d. We can also see that strategy 2 has lower P_s^b and higher P_s^d than the other strategies. This is logical since for strategy 2, clearly there are more SU blocked because the PUs are using the rented channels despite the fact that these PUs may have idle channels in their dedicated set of channels. Since there are more blocked SUs, there will be less SUs in the system and P_s^d is lower. Moreover, the difference between these strategies is lower for high values of C_r since the system is less loaded and the channel sharing strategy is not as crucial as for more loaded systems. Regarding the variation of the probabilities as a function of C_r, we can observe that it exists a value for which P_s^d is maximum. This can be explained as follows. When C_r is small the system has high load and having one more rented channel leads to higher P_s^d because the system accept more SUs and therefore, more SU sessions can be aborted. But at some point having one more rented channel leads to a system with lower load since PUs have the same arrival rate and less interruptions occur despite accepting more SUs.

In Fig. 9.9 and 9.10, P_s^b and P_s^d, respectively, are shown as a function of the arrival rate for PUs, λ_p. We can see that for strategy 1, 2 and 3 the lines have abrupt changes, this phenomenon appears because the optimal AC policy changes for different values of λ_p. The behavior of the system can be explained similarly to the behavior reflected in Figures 9.7 and 9.8. When λ_p is low, the system has low load and increasing λ_p leads to higher P_s^b and P_s^d because the system is longer in blocking states. When λ_p is very high, the system is very loaded and increasing λ_p leads to block so many new arrivals that P_s^d is lower because there are few SUs in the system.

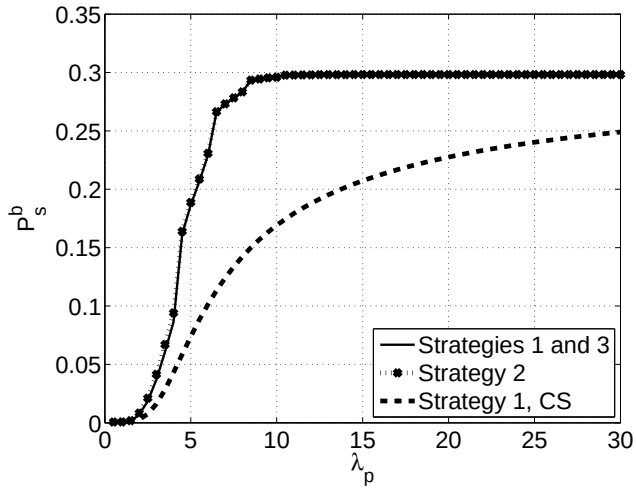

Figure 9.9: SU blocking probability as a function of λ_p.

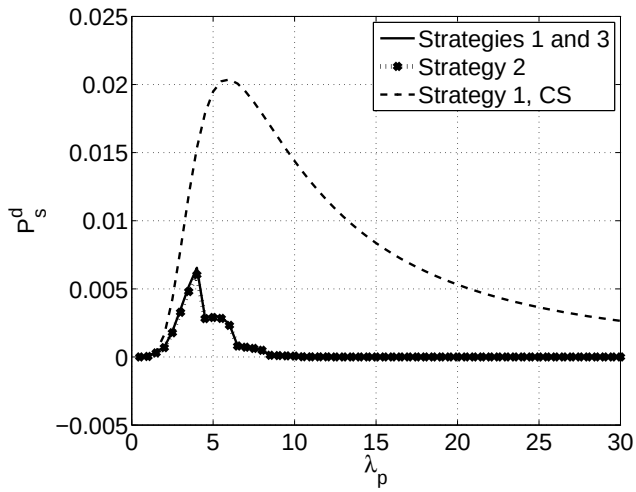

Figure 9.10: SU dropping probability as a function of λ_p.

Notice that when the optimal AC policy is applied, the value of P_s^d is 0 from a given value of λ_p. This is because the optimal AC policy decides not to accept SUs in the rented channels from this value of λ_p, as it is not worth renting channels because the number of aborted SUs is very high. However, the CS policy does not take this fact into account.

In Fig. 9.11 and 9.12, the probabilities P_s^b and P_s^d, respectively, are shown as a function of the arrival rate for SUs, λ_s. We can see that the plots have abrupt changes, especially the line for strategies 1 and 3. This is because the optimal AC policy varies for different λ_s. Strategies 1 and 3 have lower SU blocking probabilities and higher dropping probabilities than strategy 2 because in strategy 2 there are dedicated PU channels underutilized. Strategy 2 has dropping probabilities lower than strategies 1 and 3 because less SUs are accepted in the rented channels of the PN, the dedicated channels of the PN are less loaded and then, the system is shorter in the states where SUs have higher risk of being aborted.

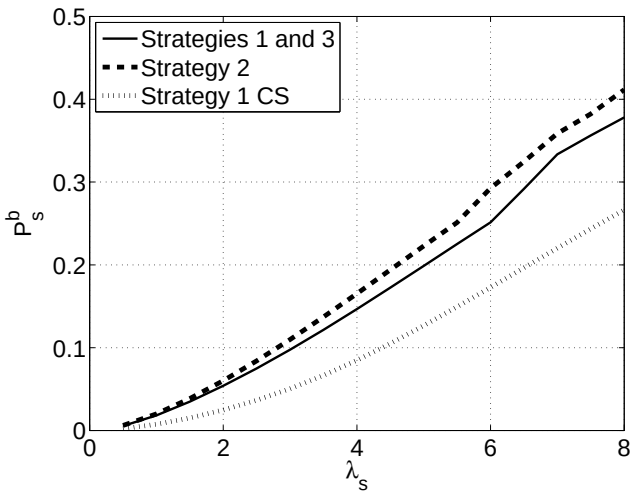

Figure 9.11: SU blocking probability as a function of λ_s.

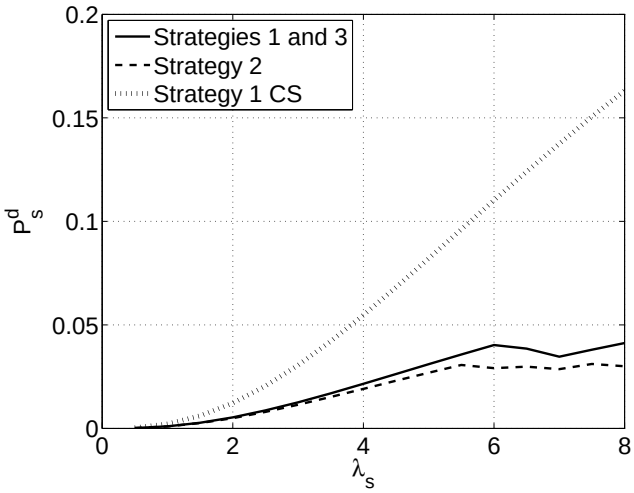

Figure 9.12: SU dropping probability as a function of λ_s.

As it was pointed out before, strategies 1 and 3 are equivalent in terms of idle channels since SUs find the same idle channels when they arrive in the system due to the repacking mechanism implemented in strategy 3. However, some parameters like the repacking rate κ of PUs defined in (9.20), can be studied only for strategy 3.

In Fig. 9.13, the repacking rate experienced by PUs when strategy 3 is applied, κ, is shown as a function of the service rate of PUs, μ_p. We can see that κ first increases with μ_p, reaches a maximum and then decreases again to 0. This can be explained as follows. If μ_p is low, PUs occupy channels longer and the system is more static. Then, less PUs which are using rented channels are reallocated to dedicated channels of the PN as the dedicated channels are not released very often. On the other hand, if μ_p is high, PUs are in the system shorter and the system is very dynamic. Then, less PUs which are using rented channels are reallocated to dedicated channels of the PN as the rented channels are released very quickly. Actually, for highly dynamic systems with very high μ_p the repacking rate κ tends to 0.

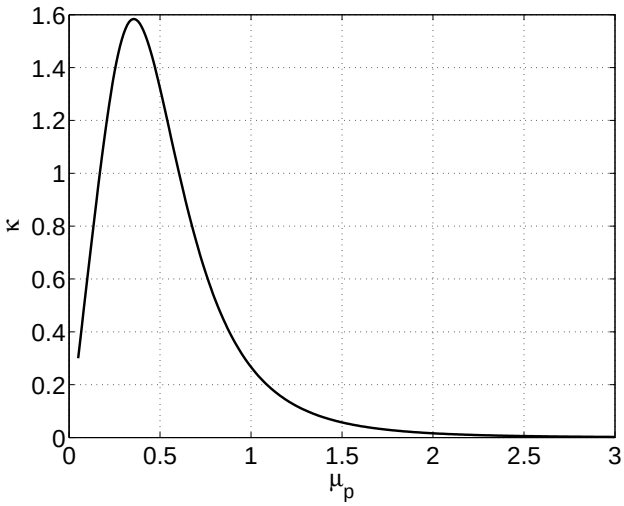

Figure 9.13: PU repacking rate for strategy 3 as a function of μ_p.

9.5 Conclusions

In this chapter, different channel sharing strategies between licensed or primary users and non-licensed or secondary users for CR networks are studied. Furthermore, a method to obtain an optimal AC policy for each strategy with respect to a cost function is presented.

The results show that strategies 1 ans 3 have equivalent results for blocking and dropping probabilities for SUs since SUs see the same system in terms of channel occupancy for both strategies. The differences between these two strategies lies in interference and management aspects more than in these performance results. Strategy 2 has higher blocking probabilities and lower dropping probabilities than strategies 1 and 3.

Regarding the optimal AC policy, it is more restrictive when the forced termination of an ongoing SU session is considered more harmful, i.e. when the weight w of the cost rate function is higher. In this case, strategies 1, 2 and 3 are more similar to each other. For low values of w the optimal AC policy

tends to a CS policy. When the optimal AC policy is applied and the primary network has a high traffic load of PUs, the optimal AC policy may decide not to accept SUs in the rented channels because the number of aborted SUs can be too high if they are accepted.

We also conclude that there is a value of rented channels for which the network becomes underloaded and higher values of C_r yield lower values of blocking and dropping probabilities for SUs. However, it will also entail a higher number of spectrum handovers for SUs, and thus a higher operational cost which means that renting channels of the primary network would have a higher economic cost for the secondary network. For future works, this fact can be studied in order to find a trade-off between the values of blocking and dropping probabilities and the operational cost produced by the spectral handovers.

Chapter 10

Conclusions

In the last decade, mobile cellular networks have experienced a major growth and progress due to a change in the way today's society creates, shares and consumes information. This fact has lead to an enormous increment of users and has opened the way to a wide technological market. 3G and forthcoming 4G networks have introduced a wide variety of services with different traffic characteristics and new applications are continuously appearing with higher QoS and bandwidth requirements. In addition, mobile cellular networks have to face strong bandwidth limitations due to the scarcity of frequencies in the radio spectrum. These new technologies have established new challenges in order to manage an increasing number of demanding services together with the scarcity of the spectrum. In this context, the radio resource management arises as a key mechanism to deal with that network characteristics. Specifically, the AC mechanism is a key aspect to efficiently use the available radio resources providing the required QoS guarantees.

In this thesis, the design and evaluation of AC policies was studied for current and forthcoming cellular networks. The first part of the thesis dealt with the implementation of AC policies in order to enhance the current mobile cellular networks. To this end, an appropriate traffic characterization was necessary. The characterization of the CRT, the CHT and the session

duration of streaming traffic has been widely studied in the literature, but this type of studies have not been carried out for elastic traffic so far. In Chapter 3, we proposed a model based on the phase-type distribution to characterize the flow duration and the handover probability of the elastic traffic. We found that the proposed model appropriately models the flow duration and the handover probability under general assumptions. Next, in Chapter 4, we compared several algorithms to design the parameter setting of the trunk reservation policy MFGC. Although the adaptive method achieves lower computational cost than the other algorithms studied, we can conclude that the computational cost necessary to design the conventional trunk reservation AC policies grows very quickly with the number of channels and SCs supported. Thus, determining parameters like the new and handover blocking probabilities, might become an unfeasible task. Moreover, after the design phase, the parameters of trunk reservation policies are static, which leads to AC policies with poor robustness. In Chapter 5, the robustness of AC policies under traffic overloads was studied. We proposed a new AC policy based on the VP policy for multiservice mobile cellular networks, which integrates streaming and elastic traffic. We found that this policy in addition to having a lower computational cost, it is also more robust against traffic overloads than conventional trunk reservation policies. Finally, in Chapter 6, we proved that trunk reservation policies do not lead to reversible and insensitive CTMC unless further restrictions are imposed and we proposed an AC policy, whose associate CTMC is reversible and insensitive to the CHT distribution.

The second part of this thesis dealt with proposing, designing and evaluating AC policies for the forthcoming mobile networks, such as the 4G networks. These networks introduce new technologies, such as the AMC technique, the femtocell concept or the CR technology. In all of them, the implementation of an appropriate AC policy is a hedging strategy in order to manage efficiently the available resources. In chapter 7, a mathematical model was presented to evaluate AC policies in OFDM based networks, which use the AMC technique. We validated this model by comparing its

results with simulations results. Moreover, we propose a dynamic AC policy, which optimizes its parameters. The dynamic AC policy was evaluated and compared to a static AC policy and we concluded that the dynamic policy outperforms a static policy. However, in order to manage the ever-increasing traffic load, it is also necessary to increase the network capacity. To this end, it is proposed the concept of femtocell. In chapter 8 an AC policy for femtocells was studied and designed. We showed that the most appropriate AC policy for users which are not subscribed to the femtocell depends on the SINR experienced by each channel and hence on the AMC used. Finally, resource management with dynamic spectrum access was also studied in Chapter 9. An optimal AC policy for secondary users is proposed. The results showed that renting channels from the primary network can be more or less convenient for the secondary network depending on the scenario and the traffic characteristics of the primary network.

Different extensions of the studies in this thesis can be identified, for example, the use of the proposed models to evaluate other AC policies. In addition to these extensions, we can point out some possible lines of future work based on the results obtained in this thesis and considering the evolution of mobile access networks. The new way that today's society consumes information leads to a cellular architecture increasingly complex. The traffic which is expected to produce the bulk of the network load will mainly occur indoor. Given that the current structure of urban areas is based on a vertical pattern, a vertical multi-layer architecture is getting more and more attractive. Moreover, the introduction of multimedia applications and smart-phones lead to an enormous economical relevance of mobile cellular networks, appearing an enormous variety of operators. In this context, the smallcell concept emerges as a more and more popular concept to combine the necessity for increasing the indoor coverage and develop a vertical multi-layer architecture operated by different companies. This new architecture establishes new challenges for the design of AC policies. The AC policy has to decide on the acceptance of a request and in which layer is served. Moreover, the acceptance decision should also be based on information such as the economical cost. Further-

more, the handover management is also more complex. The cells are smaller and, in addition to the normal horizontal pattern, they are also structured in a vertical multi-layer pattern. Therefore, the network undergoes more handovers which can be in the horizontal or the vertical dimension. These challenges motivate the need for novel resource management schemes in order to deal with an architecture with both horizontal and vertical dimensions and with a competitive economical market. Thus, the future work will be focused on extending the models developed in this thesis to networks with these characteristics.

Appendixes

Appendix A

Abbreviations and acronyms

3GPP	3rd Generation Partnership Project
AMC	Adaptive Modulation and Coding
AMP	Absorbing Markov Process
BGMP	Algorithm proposed in [BM07]
BS	Base Station
CHT	Channel Holding Time
CP	Complete Partitioning
CR	Cognitive Radio
CRT	Cell Residence Time
CTMC	Continuous-Time Markov Chain
CV	Coefficient of Variation
CR	Cognitive Radio
CS	Complete Sharing
CVO	Approximation based on K&R proposed in [CPVAOG04]
DCA	Dynamic Channel Allocation
FCA	Fixed Channel Allocation
FCC	Federal Communication Commission
FEC	Forward Error Correction
FGC	Fractional Guard Channel

FL	Fractional Limit
GC	Guard Channel
HCA	Hybrid Channel Allocation
IL	Integer Limit
K&R	Kaufman and Roberts recursion
LTE	Long Term Evolution
LLR	Linear Level Reduction
NGMN	Next Generation Mobile Networks
MCS	Modulation and Coding Scheme
MDP	Markov Decision Process
MGC	Multiple Guard Channel
MFGC	Multiple Fractional Guard Channel
MT	Mobile Terminal
OFDMA	Orthogonal Frequency-Division Multiple Access
PMC	Algorithm proposed in [PMCG05]
PH	Phase Type
PN	Primary Network
PU	Primary User
QAM	Quadrature Amplitude Modulation
QBD	Quasi Birth and Death Process
QoS	Quality of Service
QPSK	Guadrature Phase-Shift Keying
RRM	Radio Resource Management
RS	Ramdomized Stationary
RB	Resource Block
SAC	Session Admission Control
SC	Service Class
SIR	Signal to Interference Ratio
SMPD	Semi-Markov Decision Process
SON	Self-Optimizing Networks
SN	Secondary Network
SU	Secondary User

TDMA	Time Division Multiple Access
ULGM	Upper Limit and Guaranteed Minimum
VP	Virtual Partitioning
VPC	Virtual Partitioning for cellular networks
VPE	Virtual Partitioning for elastic traffic
VPS	Virtual Partitioning for streaming traffic
WiMAX	Worldwide Interoperability for Microwave Access

Appendix B

Algorithms and matrix definitions

B.1 PMC and BGMP algorithms

In this appendix, the steps followed by the PMC and BGMP algorithms are shown with more detail.

- **PMC Algorithm**

  The PMC algorithm is described in two stages called: *Algorithm* and *Procedure*. In *Algorithm* a initial value for λ^T is assigned, then in *Procedure*, it is checked if there is a set of values for $t_i^{n,h}$ that fulfill the QoS requirements. If this set exists the λ^T is increased and if not it is decreased, first with big steps and later with smaller steps. The *Procedure* is called again until a λ_{max}^T is found with a precision given by ϵ_1.

  Algorithm:

  $(\lambda_{max}^T, \boldsymbol{t}_{opt})$=**pmc**$(\boldsymbol{p_{max}}, \boldsymbol{f}, \boldsymbol{\mu^{d,s}}, \boldsymbol{\mu^{r,s}}, \boldsymbol{b}, C)$

  $\varepsilon_1 :=<$ precision $>$; $L := 0$; $U :=<$ high value $>$

  $(\text{ok}, \boldsymbol{t}) :=$ **sMFGCpmc**$(\boldsymbol{p_{max}}, U\boldsymbol{f}, \boldsymbol{\mu^{d,s}}, \boldsymbol{\mu^{r,s}}, \boldsymbol{b}, C)$

  atLeastOnce:=FALSE;

  while ok **do**

$L := U$; $\boldsymbol{t}_L := \boldsymbol{t}$; atLeastOnce:=TRUE ; $U := 2U$
$(\text{ok}, \boldsymbol{t}) :=$ **sMFGCpmc**$(\boldsymbol{p_{max}}, U\boldsymbol{f}, \boldsymbol{\mu^{d,s}}, \boldsymbol{\mu^{r,s}}, \boldsymbol{b}, C)$
end while /* it makes sure that $U > \lambda^T_{max}$ */
repeat
$\lambda := (L + U)/2$
$(\text{ok}, \boldsymbol{t}) :=$ **sMFGCpmc**$(\boldsymbol{p_{max}}, \lambda\boldsymbol{f}, \boldsymbol{\mu^{d,s}}, \boldsymbol{\mu^{r,s}}, \boldsymbol{b}, C)$
if ok **then** $L := \lambda$; $\boldsymbol{t}_L := \boldsymbol{t}$; atLeastOnce:=TRUE;
else $U := \lambda$
until $(U - L)/L \leq \varepsilon_1$ AND atLeastOnce
$\lambda^T_{max} := L$; $\boldsymbol{t} := \boldsymbol{t}_L$

In the *Procedure* the set of $t_i^{n,h}$ are initialized with small values and it is checked if there is a value of the parameters $t_i^{n,h}$ that fulfill QoS probabilities. If it exists, the optimal set is searched. The optimal set is found when all the blocking probabilities are lower than the QoS requirements but as close as possible to them. This proximity is given by the precision ϵ_2. The blocking probabilities are calculate by another procedure where the balance equations are solved.

Procedure:
(ok,$\boldsymbol{t}$)=**sMFGCpmc**$(\boldsymbol{p_{max}}, \boldsymbol{\lambda_n}, \boldsymbol{\mu^{d,s}}, \boldsymbol{\mu^{r,s}}, \boldsymbol{b}, C)$
$\varepsilon_2 :=<$ precision $>$; $\delta :=<$ small value $>$
$\boldsymbol{t} := (\delta, \delta, \ldots, \delta)$
$\boldsymbol{p} :=$ **MFGCpmc**$(\boldsymbol{t}, \boldsymbol{\lambda_n}, \boldsymbol{\mu^{d,s}}, \boldsymbol{\mu^{r,s}}, \boldsymbol{b}, C)$
repeat
canConverge:=TRUE; $i := 1$;
repeat
if $\boldsymbol{p}(i) > \boldsymbol{p_{max}}(i)$ **then**
$\boldsymbol{t}' := \boldsymbol{t}$; $\boldsymbol{t}'(i) := C$
$\boldsymbol{p}' :=$ **MFGCpmc**$(\boldsymbol{t}', \boldsymbol{\lambda_n}, \boldsymbol{\mu^{d,s}}, \boldsymbol{\mu^{r,s}}, \boldsymbol{b}, C)$
if $\boldsymbol{p}'(i) > \boldsymbol{p_{max}}(i)$ **then**
canConvege:=FALSE;

else
$L := \boldsymbol{t}(i); U := C$
repeat
$\boldsymbol{t}(i) := (L+U)/2$
$\boldsymbol{p} := \mathbf{MFGCpmc}(\boldsymbol{t}, \boldsymbol{\lambda}_n, \boldsymbol{\mu}^{d,s}, \boldsymbol{\mu}^{r,s}, \boldsymbol{b}, C)$
if $\boldsymbol{p}(i) > \boldsymbol{p}_{max}(i)$ **then** $L := \boldsymbol{t}(i)$
else $U := \boldsymbol{t}(i)$
until $(1-\epsilon_2)\boldsymbol{p}_{max}(i) \leq \boldsymbol{p}(i) \leq \boldsymbol{p}_{max}(i)$
end if
end if
$i := i+1$
until $(i > 2N)$ OR (NOT(canConverge))
if canConverge **then**
if $\boldsymbol{p}(i) \leq \boldsymbol{p}_{max}(i) \quad \forall i$ **then**
ok:=TRUE; exit:=TRUE;
else exit:=FALSE;
else ok:=FALSE; exit:=TRUE;
until exit

- **BGMP algorithm.**

The BGMP algorithm is described in three stages called: *Initialization*, *Algorithm* and *Procedure*. In *Initialization*, first the parameters $t_i^{n,h}$ and λ^T are obtained using the CVO approximation ($s = 1$) and then, the optimal parameters are obtained accurately solving the balance equations ($s = 2$).

Initialization:

$(\lambda_{max}^T, \boldsymbol{t}_{opt})$=**Initial**$(\boldsymbol{p}_{max}, \boldsymbol{f}, \boldsymbol{\mu}^{d,s}, \boldsymbol{\mu}^{r,s}, \boldsymbol{b}, C)$

$\lambda_0^T :=<$ high value $>$; $\delta :=<$ small value $>$
$\boldsymbol{t}_0 := (\delta, \delta, \ldots, \delta)$; $s = 1$;
$(\lambda_0^T, \boldsymbol{t}_0) := \mathbf{bgmp}(\lambda_0^T, \boldsymbol{t}_0, \boldsymbol{p}_{max}, \boldsymbol{f}, \boldsymbol{\mu}^{d,s}, \boldsymbol{\mu}^{r,s}, \boldsymbol{b}, C, s)$
$s = 2$;
$(\lambda_{max}^T, \boldsymbol{t}_{opt}) := \mathbf{bgmp}(\lambda_0^T, \boldsymbol{t}_0, \boldsymbol{p}_{max}, \boldsymbol{f}, \boldsymbol{\mu}^{d,s}, \boldsymbol{\mu}^{r,s}, \boldsymbol{b}, C, s)$

In *Algorithm*, as in the VPO algorithm, it is checked if a set of parameters $t_i^{n,h}$ that fulfill QoS objectives exists. The initial interval of λ^T is narrower than for the VPO algorithm, therefore the search of λ^T_{max} can be faster.

Algorithm:

$(\lambda^T_{max}, \boldsymbol{t}_{opt})$=**bgmp**$(\lambda^T_0, \boldsymbol{t}_0, \boldsymbol{p_{max}}, \boldsymbol{f}, \boldsymbol{\mu^{d,s}}, \boldsymbol{\mu^{r,s}}, \boldsymbol{b}, C, s)$

$\varepsilon_1 :=<$ precision $>$; $L := \lambda^T_{ini}$; $U := L$
$(\text{ok}, \boldsymbol{t}) :=$ **sMFGCbgmp**$(\boldsymbol{p_{max}}, U\boldsymbol{f}, \boldsymbol{t}_0, \boldsymbol{\mu^{d,s}}, \boldsymbol{\mu^{r,s}}, \boldsymbol{b}, C, s)$
atLeastOnce:=FALSE;
if ok **then**
  while ok **do**
    $L := U$; $\boldsymbol{t}_L := \boldsymbol{t}$; atLeastOnce:=TRUE ;
    if s==1 **then** $U := 2U$
    else $U := 1.1 * U$
    $(\text{ok}, \boldsymbol{t}) :=$ **sMFGCbgmp**$(\boldsymbol{p_{max}}, U\boldsymbol{f}, \boldsymbol{t}_0, \boldsymbol{\mu^{d,s}}, \boldsymbol{\mu^{r,s}}, \boldsymbol{b}, C, s)$
  end while /* it makes sure that $U > \lambda^T_{max}$ */
else
  while not(ok) **do**
    $U := L$; $\boldsymbol{t}_L := \boldsymbol{t}$; atLeastOnce:=TRUE ;
    $L := 0.9 * U$
    $(\text{ok}, \boldsymbol{t}) :=$ **sMFGCbgmp**$(\boldsymbol{p_{max}}, U\boldsymbol{f}, \boldsymbol{t}_0, \boldsymbol{\mu^{d,s}}, \boldsymbol{\mu^{r,s}}, \boldsymbol{b}, C, s)$
  end while /* it makes sure that $L < \lambda^T_{max}$ */
end if
repeat
  $\lambda := (L + U)/2$
  $(\text{ok}, \boldsymbol{t}) :=$ **sMFGCbgmp**$(\boldsymbol{p_{max}}, \lambda\boldsymbol{f}, \boldsymbol{t}0, \boldsymbol{\mu^{d,s}}, \boldsymbol{\mu^{r,s}}, \boldsymbol{b}, C, s)$
  if ok **then** $L := \lambda$; $\boldsymbol{t}_L := \boldsymbol{t}$; atLeastOnce:=TRUE;
  else $U := \lambda$
until $(U - L)/L \leq \varepsilon_1$ AND atLeastOnce
$\lambda^T_{max} := L$; $\boldsymbol{t} := \boldsymbol{t}_L$

In the *Procedure*, the initial values of $t_i^{n,h}$ will not be the same small initial values for each evaluation but they will be the calculated values in the previous evaluation. Note that depending on the value of s the **MFGCbgmp** will calculate the blocking probabilities by using the CVO approximation ($s = 1$) or by solving the balance equations ($s = 2$).

Procedure:

(ok,$\boldsymbol{t}$)=**sMFGCbgmp**($\boldsymbol{p_{max}}, \boldsymbol{\lambda_n}, \boldsymbol{t}^{n,h}, \boldsymbol{\mu}^{d,s}, \boldsymbol{\mu}^{r,s}, \boldsymbol{b}, C, s$)

$\varepsilon_2 :=<$ precision $>$; $\boldsymbol{t} := \boldsymbol{t}^{n,h}$
$\boldsymbol{p} :=$ **MFGCbgmp**$(\boldsymbol{t}, \boldsymbol{\lambda_n}, \boldsymbol{\mu}^{d,s}, \boldsymbol{\mu}^{r,s}, \boldsymbol{b}, C, s)$
repeat
  canConverge:=TRUE; $i := 1$;
  repeat
    if $\boldsymbol{p}(i) > \boldsymbol{p_{max}}(i)$ **then**
      $\boldsymbol{t}' := \boldsymbol{t}; \boldsymbol{t}'(i) := C$
      $\boldsymbol{p}' :=$ **MFGCbgmp**$(\boldsymbol{t}', \boldsymbol{\lambda_n}, \boldsymbol{\mu}^{d,s}, \boldsymbol{\mu}^{r,s}, \boldsymbol{b}, C, s)$
      if $\boldsymbol{p}'(i) > \boldsymbol{p_{max}}(i)$ **then**
        canConvege:=FALSE;
      else
        $L := \boldsymbol{t}(i); U := C$
        repeat
          $\boldsymbol{t}(i) := (L + U)/2$
          $\boldsymbol{p} :=$ **MFGCbgmp**$(\boldsymbol{t}, \boldsymbol{\lambda_n}, \boldsymbol{\mu}^{d,s}, \boldsymbol{\mu}^{r,s}, \boldsymbol{b}, C, s)$
          if $\boldsymbol{p}(i) > \boldsymbol{p_{max}}(i)$ **then** $L := \boldsymbol{t}(i)$
          else $U := \boldsymbol{t}(i)$
        until $(1 - \epsilon_2)\boldsymbol{p_{max}}(i) \leq \boldsymbol{p}(i) \leq \boldsymbol{p_{max}}(i)$
      end if
    end if
    if $\boldsymbol{p}(i) < 0.99\boldsymbol{p_{max}}(i)$ **then**
      $L := 0.9\boldsymbol{t}(i); U := \boldsymbol{t}(i)$
      repeat

```
            t(i) := (L + U)/2
            p := MFGCbgmp(t, λ_n, μ^{d,s}, μ^{r,s}, b, C, s)
            if p(i) < 0.99 p_max(i) then   U := t(i)
            else   L := t(i)
          until (1 − ε_2) p_max(i) ≤ p(i) ≤ p_max(i)
        end if
        i := i + 1
      until (i > 2N) OR ( NOT(canConverge))
      if canConverge then
        if p(i) ≤ p_max(i)   ∀i then
          ok:=TRUE; exit:=TRUE;
        else   exit:=FALSE;
      else   ok:=FALSE; exit:=TRUE;
    until exit
```

B.2 Matrix definitions for OFDMA based networks

B.2.1 Static AC policy

In this appendix the matrices of the system with the static AC policy are described for $Z = 2$ and $Z = 3$. The whole analytical model for this case is described in Section 7.3.1. Remember that the function $a_i(\boldsymbol{x})$ denotes whether a session that arrives in zone i when the system is in state $\boldsymbol{x}$ is accepted by the AC policy or not, $a_i(\boldsymbol{x}) = 1$ means that the session is accepted and $a_i(\boldsymbol{x}) = 0$ means that the session is blocked. For clarity, the notation has been simplified as $a_i(\boldsymbol{x}) = a_i$. The block matrices that were not described in Section 7.3.1 are listed below. In these matrices $p = h + l$ and the values of δ_i equal the opposite of the sum of the other elements of the same row to make the elements of each row of the transition rate matrix $\boldsymbol{Q}$ sum to 0.

- For $Z = 2$, the matrices are given by:

$$\boldsymbol{Q_1^h} = \begin{bmatrix} \delta_0 & a_2\lambda_2 & 0 & \cdots \\ \mu & \delta_1 & a_2\lambda_2 & \cdots \\ \ddots & \ddots & \ddots & \\ \cdots & 0 & (M-h)\mu & \delta_{M-h} \end{bmatrix} \tag{B.1}$$

where the size of $\boldsymbol{Q_1^h}$ is $(M+1-h) \times (M+1-h)$.

$$\boldsymbol{Q_0^h} = \begin{bmatrix} a_1\lambda_1 & 0 & 0 & \cdots \\ \epsilon_2 & a_1\lambda_1 & 0 & \cdots \\ \ddots & \ddots & & \\ \cdots & 0 & 0 & (M-h)\epsilon_2 \end{bmatrix} \tag{B.2}$$

where the size of $\boldsymbol{Q_0^h}$ is $(M+1-h) \times (M-h)$.

$$\boldsymbol{Q_2^h} = \begin{bmatrix} \mu & h\gamma_1 & 0 & \cdots \\ 0 & 2\mu & h\gamma_1 & \cdots \\ & \ddots & \ddots & \\ \cdots & 0 & (M-h)\mu & h\gamma_1 \end{bmatrix} \tag{B.3}$$

where the size of $\boldsymbol{Q_2^h}$ is $(M-h) \times (M+1-h)$.

- For $Z = 3$, the matrixes are given by:

$$\boldsymbol{A_1^{h,l}} = \begin{bmatrix} \delta_0 & a_3\lambda_3 & 0 & 0 & \cdots \\ \mu & \delta_1 & a_3\lambda_3 & 0 & \cdots \\ 0 & 2\mu & \delta_2 & a_3\lambda_3 & \cdots \\ & \ddots & \ddots & \ddots & \\ \cdots & 0 & 0 & (M-p)\mu & \delta_{M-p} \end{bmatrix}$$

where the size of $\boldsymbol{A_1^{h,l}}$ is $(M+1-p) \times (M+1-p)$.

$$\boldsymbol{A_0^{h,l}} = \begin{bmatrix} a_2\lambda_2 & 0 & 0 & \cdots \\ \epsilon_3 & a_2\lambda_2 & 0 & \cdots \\ 0 & 2\epsilon_3 & a_2\lambda_2 & \cdots \\ & \ddots & \ddots & \\ \cdots & 0 & 0 & (M-p)\epsilon_3 \end{bmatrix}$$

where the size of $\boldsymbol{A_0^{h,l}}$ is $(M+1-p) \times (M-p)$.

$$\boldsymbol{A_2^{h,l}} = \begin{bmatrix} l\mu & l\gamma_2 & 0 & 0 & \cdots \\ 0 & l\mu & l\gamma_2 & 0 & \cdots \\ & \ddots & \ddots & \ddots & \\ \cdots & 0 & 0 & l\mu & l\gamma_2 \end{bmatrix}$$

where the size of $\boldsymbol{A_2^{h,l}}$ is $(M+1-p) \times (M+2-p)$.

$$\boldsymbol{B_1^{h,l}} = \begin{bmatrix} a_1\lambda_1 & 0 & 0 & \cdots \\ & \ddots & \ddots & \\ \cdots & 0 & 0 & a_1\lambda_1 \\ \cdots & 0 & 0 & 0 \end{bmatrix} \tag{B.4}$$

where the size of $\boldsymbol{B_1^{h,l}}$ is $(M+1-p) \times (M-p)$.

The matrix $\boldsymbol{B_2^{h,l}}$ is a diagonal matrix where the values of the diagonal are equal to $l\epsilon_2$ and its size is $(M+1-p) \times (M+1-p)$.

$$\boldsymbol{C_1^{h,l}} = \begin{bmatrix} h\mu & 0 & 0 & \cdots \\ 0 & h\mu & 0 & \cdots \\ & \ddots & \ddots & \\ \cdots & 0 & h\mu & 0 \end{bmatrix} \tag{B.5}$$

where the size of $\boldsymbol{C_1^{h,l}}$ is $(M+1-p) \times (M+2-p)$.

The matrix $\boldsymbol{C_0^{h,l}}$ is a diagonal matrix where the values of the diagonal are equal to $h\gamma_1$ and its size is $(M+1-p) \times (M+1-p)$.

B.2.2 Dynamic AC policy

When the dynamic AC policy described in Section 7.3.2 is considered, the analytical model has one more level and hence one more block level than the analytical model with the same number of zones Z, for the static AC policy (see Section 7.3.2). The functions $t_1(\boldsymbol{x})$ and $t_2(\boldsymbol{x})$ denote whether the parameter f must be changed or not when the system is in state $\boldsymbol{x}$ according to the f-optimization algorithm in Eq. (7.12). When $t_1(\boldsymbol{x}) = 1$ and $t_2(\boldsymbol{x}) = 0$ means that the parameter f must be increased, $t_1(\boldsymbol{x}) = 0$ and $t_2(\boldsymbol{x}) = 1$ means that the parameter f must be decreased and $t_1(\boldsymbol{x}) = 0$ and $t_2(\boldsymbol{x}) = 0$ means that f does not change. Note that $t_1(\boldsymbol{x})$ and $t_2(\boldsymbol{x})$ cannot be 1 at the same time. For clarity, the notation has been simplified as $t_1(\boldsymbol{x}) = t_1$ and $t_2(\boldsymbol{x}) = t_2$. Remember that n_f is the number of the different discrete values that f can take and the intervals after which the optimization is performed are exponentially distributed with mean $1/\eta$. Also remember that $p = h + m$. The block matrices for $Z = 3$ that were not described in Section 7.3.2 are listed below. Again, the values of δ_i equal the opposite of the sum of the other elements on the same row to make the elements of each row of the transition rate matrix $\boldsymbol{Q}$ sum to 0.

$$\boldsymbol{D_1^{h,m,l}} = \begin{bmatrix} \delta_0 & t_1\eta & 0 & 0 & \cdots \\ t_2\eta & \delta_1 & t_1\eta & 0 & \cdots \\ 0 & t_2\eta & \delta_2 & t_1\eta & \cdots \\ & \ddots & \ddots & \ddots & \\ \cdots & 0 & 0 & t_2\eta & \delta_{n_f} \end{bmatrix}$$

where the size of $\boldsymbol{D_1^{h,m,l}}$ is $n_f \times n_f$. Note that $\boldsymbol{D_1^{h,m,l}}$ does not depend on the levels, i.e. it is equal for all the (h, m, l) levels.

The matrices $\boldsymbol{D_0^{h,m,l}}$, $\boldsymbol{D_2^{h,m,l}}$, $\boldsymbol{E_1^{h,m,l}}$, $\boldsymbol{E_2^{h,m,l}}$, $\boldsymbol{F_1^{h,m,l}}$ and $\boldsymbol{F_0^{h,m,l}}$ are diagonal matrices with size $n_f \times n_f$; the values on the diagonal are $a_3\lambda_3$, $l\mu$, $a_2\lambda_2$, $l\epsilon_3$, $m\mu$ and $m\gamma_2$ respectively. Then, the matrix $\boldsymbol{A_1^{h,m}}$ is a square matrix of size $n_f(M+1-p) \times n_f(M+1-p)$, the size of $\boldsymbol{A_0^{h,m}}$ is $n_f(M+1-p) \times n_f(M-p)$

and the size of $\boldsymbol{A_2^{h,m}}$ is $n_f(M+1-p) \times n_f(M+2-p)$.

$$\boldsymbol{B_1^{h,m}} = \begin{bmatrix} a_1\lambda_1 & 0 & 0 & \cdots \\ 0 & a_1\lambda_1 & 0 & \cdots \\ & \ddots & \ddots & \\ \cdots & 0 & 0 & a_1\lambda_1 \\ \cdots & 0 & 0 & 0 \\ \cdots & \vdots & \vdots & \vdots \\ \cdots & 0 & 0 & 0 \end{bmatrix}$$

where the number of rows with all elements 0 is n_f and the size of $\boldsymbol{B_1^{h,m}}$ is $(n_f(M+1-p)) \times (n_f(M-p))$.

The Matrix $\boldsymbol{B_2^{h,m}}$ is a diagonal matrix where the values of the diagonal are $m\epsilon_2$ and its size is $(n_f(M+1-p)) \times (n_f(M+1-p))$.

$$\boldsymbol{C_1^{h,m}} = \begin{bmatrix} h\mu & 0 & 0 & 0 & 0 & \cdots & 0 \\ 0 & h\mu & 0 & 0 & 0 & \cdots & 0 \\ & \ddots & \ddots & \ddots & \ddots & & \\ \cdots & 0 & h\mu & 0 & 0 & \cdots & 0 \\ \cdots & 0 & 0 & h\mu & 0 & \cdots & 0 \end{bmatrix}$$

where the number of columns with all elements 0 is n_f and the total size of $\boldsymbol{C_1^{h,m}}$ is $(n_f(M+1-p)) \times (n_f(M+2-p))$.

The matrix $\boldsymbol{C_0^{h,m}}$ is a diagonal matrix where the values of the diagonal are $h\gamma_1$ and its size is $(n_f(M+1-p)) \times (n_f(M+1-p))$.

B.3 Matrix definitions for CR technology

In this appendix the matrices of the system defined for the CR technology are described. The whole analytical model for this case is described in Section 9.2. Remember that the function $a_s(\boldsymbol{x})$ denotes whether an SU session that

arrives to the system when the system is in state $\boldsymbol{x}$ is accepted by the AC policy or not, $a_s(\boldsymbol{x}) = 1$ means that the session is accepted and $a_s(\boldsymbol{x}) = 0$ means that the session is blocked. For clarity, the notation has been simplified as $a_s(\boldsymbol{x}) = a_s$. The block matrices that were not described in Section 9.2 are listed below. In these matrices the values of δ_{row} equal the opposite of the sum of the other elements of the same row to make the elements of each row of the transition rate matrices $\boldsymbol{Q}_1$, $\boldsymbol{Q}_2$ and $\boldsymbol{Q}_3$ sum to 0.

- Strategy 1

  The matrices are given by:

$$\boldsymbol{A_1^0} = \begin{bmatrix} \delta_0 & a_s\lambda_s & 0 & \cdots \\ \mu_s & \delta_1 & a_s\lambda_s & \cdots \\ \ddots & \ddots & \ddots & \\ \cdots & 0 & (C_s + C_r)\mu_s & \delta_{C_s+C_r} \end{bmatrix}$$

  The matrix $\boldsymbol{A_0^0}$ is a diagonal matrix where the values of the diagonal are equal to λ_p.

  The matrix $\boldsymbol{A_2^1}$ is a diagonal matrix where the values of the diagonal are equal to μ_p.

$$\boldsymbol{A_1^{C_p+i}} = \begin{bmatrix} \delta_0 & a_s\lambda_s & 0 & \cdots \\ \mu_s & \delta_1 & a_s\lambda_s & \cdots \\ \ddots & \ddots & \ddots & \\ \cdots & 0 & (C_s + C_r - i)\mu_s & \delta_{C_s+C_r-i} \end{bmatrix}$$

$$\boldsymbol{A_0^{C_p-1+i}} = \begin{bmatrix} \lambda_p & 0 & 0 & \cdots \\ 0 & \lambda_p & 0 & \cdots \\ & \ddots & \ddots & \\ \cdots & 0 & 0 & \lambda_p \\ \cdots & 0 & 0 & 0 \end{bmatrix}$$

$$\boldsymbol{A_2^{C_p+i}} = (C_p + i) \begin{bmatrix} \mu_p & 0 & \cdots & 0 & 0 \\ 0 & \mu_p & \cdots & 0 & 0 \\ & \ddots & \ddots & & \vdots \\ 0 & 0 & \cdots & \mu_p & 0 \end{bmatrix}$$

- Strategy 2

The matrices are given by:

$$\boldsymbol{B_1^0} = \begin{bmatrix} \boldsymbol{B_{1,1}^0} & 0 & 0 & \cdots \\ \boldsymbol{B_{1,2}^1} & \boldsymbol{B_{1,1}^1} & 0 & \cdots \\ \ddots & \ddots & \ddots & \\ \cdots & 0 & \boldsymbol{B_{1,2}^{C_r}} & \boldsymbol{B_{1,1}^{C_r}} \end{bmatrix}$$

$$\boldsymbol{B_1^{C_p}} = \begin{bmatrix} \boldsymbol{B_{1,1}^0} & \boldsymbol{B_{1,0}^0} & 0 & \cdots \\ \boldsymbol{B_{1,2}^1} & \boldsymbol{B_{1,1}^1} & \boldsymbol{B_{1,0}^1} & \cdots \\ \ddots & \ddots & \ddots & \\ \cdots & 0 & \boldsymbol{B_{1,2}^{C_r}} & \boldsymbol{B_{1,1}^{C_r}} \end{bmatrix}$$

where

$$\boldsymbol{B_{1,1}^{x_r}} = \begin{bmatrix} \delta_0 & a_s\lambda_s & 0 & \cdots \\ \mu_s & \delta_1 & a_s\lambda_s & \cdots \\ \ddots & \ddots & \ddots & \\ \cdots & 0 & (C_s + C_r - x_r)\mu_s & \delta_{C_s+C_r-x_r} \end{bmatrix}$$

$$\boldsymbol{B_{1,2}^{x_r}} = x_r \begin{bmatrix} \mu_s & 0 & \cdots & 0 & 0 \\ 0 & \mu_s & \cdots & 0 & 0 \\ & \ddots & \ddots & & \vdots \\ 0 & 0 & \cdots & \mu_s & 0 \end{bmatrix}$$

$$\boldsymbol{B}_{\mathbf{1,0}}^{\boldsymbol{x_r}} = \begin{bmatrix} \lambda_p & 0 & 0 & \cdots \\ 0 & \lambda_p & 0 & \cdots \\ & \ddots & \ddots & \\ \cdots & 0 & 0 & \lambda_p \\ \cdots & 0 & 0 & \lambda_p \end{bmatrix}$$

The matrices $\boldsymbol{B}_{\mathbf{1,1}}^{\boldsymbol{x_r}}$ are square matrices with size $(C_s + C_r - x_r + 1) \times (C_s + C_r - x_r + 1)$. The sizes of matrices $\boldsymbol{B}_{\mathbf{1,2}}^{\boldsymbol{x_r}}$ are $(C_s + C_r - x_r + 1) \times (C_s + C_r - x_r + 2)$ and the sizes of matrices $\boldsymbol{B}_{\mathbf{1,0}}^{\boldsymbol{x_r}}$ are $(C_s + C_r - x_r + 1) \times (C_s + C_r - x_r)$.

The matrix $\boldsymbol{B}_{\mathbf{0}}^{\mathbf{0}}$ is a diagonal matrix where the values of the diagonal are equal to λ_p.

The matrix $\boldsymbol{B}_{\mathbf{2}}^{\mathbf{1}}$ is a diagonal matrix where the values of the diagonal are equal to μ_p.

- Strategy 3

The matrices are given by:

$$\boldsymbol{C}_{\mathbf{1}}^{\mathbf{0}} = \begin{bmatrix} \delta_0 & a_s\lambda_s & 0 & \cdots \\ \mu_s & \delta_1 & a_s\lambda_s & \cdots \\ \ddots & \ddots & \ddots & \\ \cdots & 0 & (C_s + C_r)\mu_s & \delta_{C_s+C_r} \end{bmatrix}$$

The matrix $\boldsymbol{C}_{\mathbf{0}}^{\mathbf{0}}$ is a diagonal matrix where the values of the diagonal are equal to λ_p.

The matrix $\boldsymbol{C}_{\mathbf{2}}^{\mathbf{1}}$ is a diagonal matrix where the values of the diagonal are equal to μ_p.

The sizes of $\boldsymbol{C}_{\mathbf{1}}^{\mathbf{0}}$, $\boldsymbol{C}_{\mathbf{0}}^{\mathbf{0}}$ and $\boldsymbol{C}_{\mathbf{2}}^{\mathbf{1}}$ are $(C_s + C_r + 1) \times (C_s + C_r + 1)$.

Regarding the matrix $\boldsymbol{C}_{\mathbf{1}}^{\boldsymbol{C_p}}$ of high-level C_p , we have that $\boldsymbol{C}_{\mathbf{1}}^{\boldsymbol{C_p}} = \boldsymbol{B}_{\mathbf{1}}^{\boldsymbol{C_p}}$.

And finally,

$$\boldsymbol{C_0^{C_p-1}} = \begin{bmatrix} \lambda_p & 0 & 0 & 0 & 0 & \cdots & 0 \\ 0 & \lambda_p & 0 & 0 & 0 & \cdots & 0 \\ & \ddots & \ddots & \ddots & \ddots & & \\ \cdots & 0 & \lambda_p & 0 & 0 & \cdots & 0 \\ \cdots & 0 & 0 & \lambda_p & 0 & \cdots & 0 \end{bmatrix}$$

$$\boldsymbol{C_2^{C_p}} = C_p \begin{bmatrix} \mu_p & 0 & 0 & \cdots \\ 0 & \mu_p & 0 & \cdots \\ & \ddots & \ddots & \\ \cdots & 0 & 0 & \mu_p \\ \cdots & 0 & 0 & 0 \\ \cdots & \vdots & \vdots & \vdots \\ \cdots & 0 & 0 & 0 \end{bmatrix}.$$

Appendix C

Mathematical tools

C.1 Random variable distributions

C.1.1 General distributions

Below you will find a brief review of the distributions considered in this work.

- *Lognormal distribution*

  If the parameters denoted μ_n and σ_n are respectively the mean and standard deviation of the variable's natural logarithm, the distribution function of the lognormal distribution is given by:

  $$F(x) = \frac{1}{2} + \frac{1}{2}\text{erf}\left[\frac{ln(x) - \mu_n}{\sqrt{2\sigma_n^2}}\right], \tag{C.1}$$

  where erf is the *error function*:

  $$\text{erf}(x) = \frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}dt.$$

The mean and the standard deviation of the lognormal variable are respectively given by:

$$m = e^{\left(\mu_n + \frac{\sigma_n^2}{2}\right)} \qquad \text{and} \qquad \sigma = \sqrt{\left(e^{\sigma_n^2} - 1\right) e^{2\mu_n + \sigma_n^2}}. \tag{C.2}$$

The Coefficient of Variation, CV, is given by:

$$CV = \frac{\sigma}{m} = \sqrt{e^{\sigma_n^2} - 1}. \tag{C.3}$$

- *Hyper-exponential distribution.*

  The hyper-exponential distribution consists of two exponential distributions with rates $a\nu$ and ν/a where $a > 1$. The probability that the random variable takes on the form of each exponential distribution is respectively $a/(1-a)$ and $1/(1+a)$. The distribution function of the hyper-exponential distribution is given by:

$$F(x) = \frac{a}{1-a}\left(1 - e^{-a\nu x}\right) + \frac{1}{1+a}\left(1 - e^{-\frac{\nu}{a}x}\right). \tag{C.4}$$

  The mean and the standard deviation are respectively given by:

$$m = \frac{1}{\nu} \qquad \text{and} \qquad \sigma = \frac{1}{\nu}\sqrt{2\left(a + \frac{1}{a} - \frac{3}{2}\right)}. \tag{C.5}$$

  The CV is given by:

$$CV = \frac{\sigma}{m} = \sqrt{2\left(a + \frac{1}{a} - \frac{3}{2}\right)} > 1. \tag{C.6}$$

- *Erlang distribution*

  The Erlang distribution consists of the sum of $n \in \mathbb{N}^+$ independent exponential variables with rate $n\nu$ each one. The distribution function

of the Erlang distribution is given by:

$$F(x) = \frac{\int_0^{n\nu x} t^{n-1} e^{-t}}{(n-1)!}. \tag{C.7}$$

The mean and the standard deviation are respectively given by:

$$m = \frac{1}{\nu} \qquad \text{and} \qquad \sigma = \frac{1}{\nu\sqrt{n}}. \tag{C.8}$$

The CV is given by:

$$CV = \frac{\sigma}{m} = \frac{1}{\sqrt{n}} < 1. \tag{C.9}$$

- *Pareto distribution*

  The Pareto distribution is defined by two parameters, the minimum value $x_m > 0$ and the shape $k > 0$. The distribution function of the Pareto distribution is given by:

$$F(x) = 1 - \left(\frac{x_m}{x}\right)^k \qquad \text{for} \quad x \geq x_m. \tag{C.10}$$

  The mean an the standard deviation of the Pareto variable are respectively given by:

$$m = \frac{k x_m}{k-1} \qquad \text{and} \qquad \sigma = \sqrt{\frac{x_m^2 k}{(k-1)^2(k-2)}}, \tag{C.11}$$

  where if $k \leq 1$, the mean does not exist and if $k \leq 2$, the standard deviation does not exist. The CV is given by:

$$CV = \frac{\sigma}{m} = \frac{1}{\sqrt{k(k-2)}}, \tag{C.12}$$

  where $k > 2$.

- *Bounded Pareto distribution*

  The bounded Pareto distribution is defined by three parameters, the shape $k > 0$ as in the standard Pareto distribution, the minimal value, L, and the maximum value, H. The distribution function of the bounded Pareto distribution is given by:

$$F(x) = \frac{1 - L^k x^{-k}}{1 - \left(\frac{L}{H}\right)^k}. \tag{C.13}$$

  The mean an the standard deviation of the bounded Pareto variable are respectively given by:

$$m = \frac{L^k}{1 - \frac{L}{H}^k} \cdot \left(\frac{k}{k-1}\right) \cdot \left(\frac{1}{L^{k-1}} - \frac{1}{H^{k-1}}\right) \qquad k \neq 1, \tag{C.14}$$

  and

$$\sigma = \sqrt{\frac{L^k}{1 - \frac{L}{H}^k} \cdot \left(\frac{k}{k-2}\right) \cdot \left(\frac{1}{L^{k-2}} - \frac{1}{H^{k-2}}\right)} \qquad k \neq 2, \tag{C.15}$$

C.1.2 Phase type distributions

Phase type (PH) distributions provide a versatile framework to extend many simple results on exponential distributions to more complex models, being these models still computationally tractable. The idea is to model random time intervals based on the method of states [Neu81], i.e., these distributions are composed by a number of exponentially distributed phases. Thus, the resulting Markovian structure can be exploited to simplify the analytical analysis. If a general case is considered, a PH distribution defines the time until absorption in a Markov process with an absorbing state, i.e., an Absorbing Markov Process (AMP) [Neu81]. An *absorbing* state is defined as a state which is impossible to leave. Then, a Markov process is absorbing if it

has at least one absorbing state. In a AMP, a state which is not absorbing is called *transient* and from every transient state, is must be possible to reach an absorbing state, not necessarily in one step. Absorption occurs when the absorbing state is reached.

From now on, we consider that the process only has one absorbing state. If the first states are transient and the last state is the absorbing state, the generator or transition matrix, $\boldsymbol{G}$, associated to the AMP with n transient states has the following canonical form

$$\boldsymbol{G} = \left[\begin{array}{c|c} \boldsymbol{S} & \boldsymbol{\tau} \\ \hline \mathbf{0} & 0 \end{array}\right],$$

where the matrix $\boldsymbol{S}$ is an $n \times n$ matrix with the transition rates among the transient states, the vector $\boldsymbol{\tau}$ is a column vector of size n with the transition rates from the transient states to the absorbing state and the vector $\mathbf{0}$ is a column vector of 0s with size n.

As $\boldsymbol{G}$ is the transition matrix of a Markov Process, it satisfies

$$G_{ii} < 0, \qquad \tau_i \geq 0, \qquad G_{ij}, \geq 0 \qquad\qquad 1 \leq i \neq j \leq \ n$$

and clearly, the vector $\boldsymbol{\tau}$ satisfies

$$\boldsymbol{\tau} = -\boldsymbol{S}\mathbf{1}, \tag{C.16}$$

where $\mathbf{1}$ is a column vector of 1s.

In order to define appropriately an AMP, it is also necessary to know the initial probability vector, $\boldsymbol{\alpha}$, which represents the probabilities that the process starts in any of the transient states. Hence, the vector $\boldsymbol{\alpha}$ is a row vector of size n. If α_0 is a scalar and corresponds to the probability that the system starts in the absorbing state, we have

$$\alpha_0 + \boldsymbol{\alpha}\mathbf{1} = 1.$$

It is usually assumed that α_0 is 0 and a PH distribution is normally represented by the pair $(\boldsymbol{\alpha}, \boldsymbol{S})$.

Distribution and moments [Neu81]

The cumulative distribution of a random variable X that is PH$(\boldsymbol{\alpha}, \boldsymbol{S})$ is

$$F_s(x) = 1 - \boldsymbol{\alpha} e^{x\boldsymbol{S}} \mathbf{1} \qquad x \geq 0, \tag{C.17}$$

and the probability density function is

$$f_s(x) = \boldsymbol{\alpha} e^{x\boldsymbol{S}} \boldsymbol{\tau} \qquad x > 0, \tag{C.18}$$

where the matrix exponential e^X is a $n \times n$ matrix given by the power series

$$e^X = \sum_{n \geq 0} \frac{1}{n!} X^n.$$

Finally, the k-th moment of the PH$(\boldsymbol{\alpha}, \boldsymbol{S})$ is given by

$$E[X^k] = k! \boldsymbol{\alpha} (-\boldsymbol{S}^{-1})^k \mathbf{1} \qquad k \geq 1. \tag{C.19}$$

Therefore, the mean of a PH$(\boldsymbol{\alpha}, \boldsymbol{S})$ distribution is

$$E[X] = \boldsymbol{\alpha} (-\boldsymbol{S}^{-1}) \mathbf{1} \tag{C.20}$$

C.2 Level-dependent finite QBDs: LLR algorithm

A QBD process is a CTMC where the state space can be divided into levels, and levels can be divided into phases. The process is restricted in level jumps only to its nearest neighbors. From one state, the process only can jump to states with one more level or one less level. Inside the same level

the jumps are unrestricted, i.e., the jumps in the phase dimension is not restricted. When the transitions of the QBD process are dependent of the level, it is called *level-dependent* or *inhomogeneous* QBD process. Moreover, when the number of levels is finite, the QBD process is called *finite* QBD process. Thus, a level-dependent finite QBD process with M levels has a transition rate matrix, $\boldsymbol{Q}$, of the form:

$$\boldsymbol{Q} = \begin{bmatrix} \boldsymbol{Q_1^0} & \boldsymbol{Q_0^0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \cdots \\ \boldsymbol{Q_2^1} & \boldsymbol{Q_1^1} & \boldsymbol{Q_0^1} & \boldsymbol{0} & \boldsymbol{0} & \cdots \\ \boldsymbol{0} & \boldsymbol{Q_2^2} & \boldsymbol{Q_1^2} & \boldsymbol{Q_0^2} & \boldsymbol{0} & \cdots \\ & & \ddots & \ddots & \ddots & \\ \cdots & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{Q_2^{M-1}} & \boldsymbol{Q_1^{M-1}} & \boldsymbol{Q_0^{M-1}} \\ \cdots & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{Q_2^M} & \boldsymbol{Q_1^M} \end{bmatrix} \tag{C.21}$$

In order to determine the stationary distribution, several algorithms may be followed. We have chosen the Level Linear Level Reduction (LLR) algorithm because its simplicity, its stability and its applicability to a large number of cases.

The LLR algorithm is adapted from Gaver, Jacobs and Latouche [GJL84] method and it consists of the two following stages:

1. Stage 1

   In the first stage the state space is reduced progressively by removing one level at each step until the Markov process on the last M level is left. The matrices $\boldsymbol{R}$ and $\boldsymbol{U}$ play an important role in this method. These matrices are a generalization of the counterpart matrices for the discrete-time case described in [Neu81]. In our continuous-time case, the matrices $\boldsymbol{U_k}$ are defined as follows:

$$\boldsymbol{U_0} = \boldsymbol{Q_1^0}$$

and

$$\boldsymbol{U_k} = \boldsymbol{Q_1^k} + \boldsymbol{Q_2^k}(-\boldsymbol{U_{k-1}})^{-1}\boldsymbol{Q_0^{k-1}}, \qquad 1 \leq k \leq M.$$

The matrices $\boldsymbol{R_k}$ are described as follows:

$$\boldsymbol{R_k} = \boldsymbol{Q_2^k}(-\boldsymbol{U_{k-1}})^{-1}, \qquad 1 \leq k \leq M.$$

The first stage of the LLR algorithm obtain these matrices using successively iterations:

1: $\boldsymbol{U} \leftarrow \boldsymbol{Q_1^0}$
2: **for** $l = 1, 2, \ldots, M$
3: $\boldsymbol{R^l} \leftarrow \boldsymbol{Q_2^l}(-\boldsymbol{U})^{-1}$
4: $\boldsymbol{U} \leftarrow \boldsymbol{Q_1^l} + \boldsymbol{R^l}\boldsymbol{Q_0^{l-1}}$
5: **end for**

2. Stage 2

In this second stage, first, the Markov process that corresponds to the M level is solved. Next, the stationary vector is constructed by adding back one level at each step. Finally, the stationary vector is normalized. If the vector $\mathbf{1}$ is a column vector of 1's, the second stage of the algorithm is as follows:

6: **solve** $\boldsymbol{\pi^M}$ **from** $\boldsymbol{\pi^M U = 0}$; $\boldsymbol{\pi^M 1} = 1$
7: **for** $l = M - 1, \ldots, 0$
8: $\boldsymbol{\pi^l} \leftarrow \boldsymbol{\pi^{l+1} R^l}$
9: **end for**
10: $\boldsymbol{\pi} \leftarrow 1/(\boldsymbol{\pi 1})\boldsymbol{\pi}$

Appendix D

Simulations tools

D.1 OPNET discrete-event simulator

In order to verify the mobility and session duration modeling assumptions made in the analytical model in Chapter 7 and to verify the results obtained with this analytical model, simulations that model the mobility and the session duration of the users more realistically are performed. In the simulations, sessions are generated according to a Poisson process with the same arrival rate that the analytical model. The duration of a session is, unlike in the analytical model, chosen from a lognormal distribution as this distribution more realistically models the duration of sessions [GLZ07]. The lognormal parameters are chosen such that this lognormal distribution has the same mean and variance as the exponential distribution considered in the analytical model.

When a session is generated, it is placed uniformly in the cell and is subjected to the AC policy. If it is admitted to the cell, it starts moving around. Users move around according to a random walk mobility model [CBD02]. This means that when a session is started, it chooses a direction ϕ (in radians) uniformly distributed in the interval $[0, 2\pi[$ and starts moving in the

chosen direction at a fixed velocity v. After the user has traveled over a fixed distance d, it again chooses a direction and starts moving in the newly chosen direction. This is repeated until the session finishes and the user is removed from the system. When a user reaches the boundary of the cell, it bounces against the circular edge and continues its path in the reflection direction. Thus, the residence time in each zone is not modeled using random distributions. Instead, transitions between zones occur when a user crosses the border of a zone. The session blocking probability is calculated by counting the total number of generated sessions and the number of sessions that are dropped by the AC and dividing the latter by the former. The low QoS probability is calculated by recording the time that the system has a low QoS and dividing it by the total simulation time.

In the simulation model of the dynamic AC policy, the optimization algorithm, when enabled, will check the load at regular time instances which are multiples of the mean of the exponential distribution that models the time between two optimizations. At these time instances, the test of the algorithm that decides whether the AC threshold is raised or lowered will be performed and, if necessary, the appropriate action will be taken. In contrast to the analytical model, the simulation model uses fixed optimization intervals; this is because fixed-length intervals are more commonly used in reality than exponentially distributed ones.

D.2 C++ discrete-event simulator

In order to verify the results obtained with the analytical models for multiservice mobile cellular networks presented in this work, simulations are performed to model more realistically the random variable distributions used in these analytical models. The simulation model is implemented using a C++ discrete event simulation environment. The model mimics the real system behaviour and therefore it is completely independent from the analytical model.

In these simulations, the results are obtained by considering a multi-cell scenario with a central cell and two outer rings of cells, which make a total of 19 cells. Upon cell residence time termination, terminals select a neighbor cell with equal probability, i.e. each one with probability 1/6. We consider wraparound to avoid abnormal terminations at the edges. In the simulations, sessions are generated according to a Poisson process with the same arrival rate that in the analytical model. The parameters of the distributions which model the random variable distributions under study, such as the cell residence time, are chosen such that this distribution has the same mean as the exponential distribution considered in the analytical model and different coefficient of variation, CV, defined in the corresponding section.

The blocking probabilities for new (handover) sessions are calculated as the number of new (handover) sessions initiated in the cell which are not accepted divided by the total number of new (handover) sessions initiated as new (handover) in a cell. The first handover probability is computed as the number of flows initiated as new in a cell that execute a handover, divided by the total number of flows initiated as new in a cell. The probability of handover beyond the first one is computed as the number of flows initiated as handover in a cell that execute another handover, divided by the total number of flows initiated as handover in a cell. The probability that an user performs n handovers before finishing its service in the system is computed as the number of flows that execute exactly n handovers divided by the total number of flows initiated in the system, i.e, the fraction of flows that complete successfully or abandon that execute exactly n handovers. Note that we do not count those flows that are forced to terminate upon a handover failure, nor those blocked at initiation time.

In order to generate a random variable which follows the residual life distribution of the cell residence time, a random variable has to be generated knowing its distribution function. There are several methods to generate a random variable from its distribution function. We use the method known as *acceptance-rejection method*, which is explained below.

Acceptance-rejection method

We want to generate a random variable $\hat{T}_r$ from its distribution function $\hat{F}_r(t)$, which has a probability density function $\hat{f}_r(t)$. This method can be applied if there is another probability density function $g(t)$ so that the ratio $\hat{f}_r(t)/g(t)$ is bounded by a constant $c > 0$, that is:

$$\hat{f}_r(t) < cg(t) \quad \forall t.$$

If $g(t)$ exists, this method follows these steps:

1. Generate x with density $g(x)$.
2. Generate u uniformly distributed in $[0, cg(x)]$.
3. If $u \leq f(x)$, set $t = x$ ("accept"). Otherwise go back to step 1, ("reject").

In our problem, the probability density function is:

$$\hat{f}_r(t) = \frac{1}{m}\left[1 - F_r(t)\right],$$

where $m = E[T_r]$ is the mean of the CRT. Clearly, this function is bounded by the constant $1/m$. Therefore, $\hat{f}_r(t)$ can be bounded by a rectangle with height $1/m$ and $g(t) = 1/(c \cdot m)$ where we choose a value of $c = 25$.

Appendix E

Publications

E.1 Related with this thesis

E.1.1 Journal

1. Elena Bernal-Mor, Vicent Pla, and Jorge Martinez-Bauset.
Handover Performance for Elastic Flows in Mobile Cellular Networks, IEEE Communication Letters, Vol. 16, no. 10, pp. 1632-1635. 2012.

2. Bart Sas, Elena Bernal-Mor, Kathleen Spaey, Vicent Pla, Chris Blondia and Jorge Martinez Bauset.
Modelling the time-varying cell capacity in LTE networks, Telecommunication Systems (accepted), Springer. Vol. , pp. .

3. Jorge Martinez-Bauset, Vicent Pla and Elena Bernal-Mor.
Insensitive Call Admission Control for Wireless Multiservice Networks, IEEE Communication Letters, Vol. 15, no. 9, pp. 989-991. 2011.

E.1.2 International conferences

1. Elena Bernal-Mor, Vicent Pla, David M. Gutierrez-Estevez and Jorge Martinez-Bauset.
   Resource Management for Macrocell Users in Hybrid Access Femtocells, In Proceeding GLOBECOM'12 - 2012 global communications conference, Anaheim, California, USA, 3-7 December, 2012, (Accepted).

2. Bart Sas, Elena Bernal-Mor, Kathleen Spaey, Vicent Pla, Chris Blondia and Jorge Martinez Bauset.
   An analytical model to study the impact of time-varying cell capacity in LTE networks, In Proceeding WMNC'11 - 2011 fourth joint IFIP wireless and mobile networking conference, Toulouse, France, 26-28 October, 2011, pp. 1-8.

3. Elena Bernal-Mor, Vicent Pla and Jorge Martínez-Bauset.
   Analysis of different channel sharing strategies in cognitive radio networks, In Proceeding MACOM'10 - The Third international conference on multiple access communications, Barcelona, Spain. Lecture Notes in Computer Science (LNCS) Springer-Verlag, Vol. 6235, pp. 70-73, September 2010.

4. Elena Bernal-Mor, Vicent Pla and Jorge Martínez-Bauset.
   Robust Admission control for streaming and elastic services in cellular networks, In Proceedings ISCC'10 - 2010 IEEE Symposium on Computers and Communications. Riccione, Italy. 22-25 June 2010, pp. 372-374.

5. Elena Bernal-Mor, David Garcia-Roger, Jorge Martinez-Bauset, and Vicent Pla.

Optimal design of Multiple Fractional Guard Channel Policy in multiservice cellular networks. In Proceedings UBICOMM'08 - The Second international conference on Mobile Ubiquitous Computing, Systems, Services and Technologies. Valencia, Spain, Sept. 29 2008-Oct. 4 2008, pp. 476-481

E.2 Other publications

E.2.1 International conferences

1. Vicent Pla, Stijn De Vuysty, Koen De Turcky, Elena Bernal-Mor, Jorge Martinez-Bauset and Sabine Wittevronge.
   Saturation Throughput in a Heterogeneous Multi-channel Cognitive Radio Network, In Proceedings ICC'11 - 2011 IEEE international conference on communications, Kyoto, Japon. 5-9 June 2011, pp. 1-5.

Appendix F

Research projects related to this thesis

This work has been developed in the framework of the following national and international projects:

- *Control de admisión en redes róviles heterogéneas*, CARMHet (Ministerio de Educación y Ciencia, TSI2005-07520-C03-03)
- *Admission control in mobile networks with rate-adaptive streams and hierarchical architecture*, ADMINISTRA (Ministerio de Ciencia e Innovación, TIN2008-06739-C04-02/TSI)
- *Security, services, networking and performance of next generation IP-based multimedia wireless Networks*, S2EunNet (European Commission under the FP7/People/IRSES action)

Likewise, we also thank the support from the European Commission through the European Network of Excellence *Anticipating the Network of the Future - From Theory to Design*, Euro-NF.

Finally, we thank the support from the Spanish Government through the scholarship *Formación de Personal Investigador* (FPI) with reference BES-2007-15030.

Bibliography

[3GP10a] 3GPP, *3gpp tr 36.213: Evolved universal terrestrial radio access (e-utra); radio resource control (rrc); physical layer procedures*, June 2010.

[3GP10b] 3GPP, *3gpp tr 36.942: Evolved universal terrestrial radio access (e-utra); radio resource control (rrc); radio frequency (rf) system scenarios*, September 2010.

[AAFS04] I.F. Akyildiz, Y. Altunbasak, F. Fekri, and R. Sivakumar, *AdaptNet: adaptive protocol suite for next generation wireless internet*, IEEE Communications Magazine **42** (2004), 128–138.

[AGECR10] I. F. Akyildiz, D. M. Gutierrez-Estevez, and E. Chavarria-Reyes, *The evolution to 4G cellular systems: LTE-Advanced*, Physical Communications (Elsevier) Journal **3** (2010), no. 4, 217–244.

[AJ09] Al-Rawi, M. and Jäntti, R., *Call admission control with active link protection for opportunistic wireless networks*, Telecommunication System **41** (2009), no. 1, 13–23.

[ALVM08] I. F. Akyildiz, W. Y. Lee, M. C. Vuran, and S. Mohanty, *A survey on spectrum management in cognitive radio networks*, IEEE Communications **46** (2008), no. 4, 40–48.

[Bar01] N. Bartolini, *Handoff and optimal channel assignment in wireless networks*, Mobile Networks and Applications (MONET) **6** (2001), no. 6, 511–524.

[BB97] F. Barceló and S. Bueno, *Idle and inter-arrival time statistics in public access mobile radio (PAMR) systems*, Proceedings of IEEE GLOBECOM, 1997, pp. 126–130.

[BBP01] A.-L. Beylot, S. Boumerdassi, and G. Pujolle, *NACR: A new adaptive channel reservation in cellular communication systems*, Telecommunication Systems **17** (2001), 233–241.

[BC02] N. Bartolini and I. Chlamtac, *Call admission control in wireless multimedia networks*, Proceedings of IEEE PIMRC, 2002.

[Bel57] R. Bellman, *Dynamic programming*, Princeton University Press, 1957.

[BF01] C. C. Beard and V. S. Frost, *Prioritized resource allocation for stressed networks*, IEEE/ACM Transactions on Networking **9** (2001), no. 5, 618–633.

[BFDL+09] H. B., I. Fernandez-Diaz, R. Litjens, K. Spaey, and E. U. Warriach, *Self-optimisation methods for stand-alone functionalities in wireless access networks: Packet Scheduling parameter optimisation*, Tech. report, INFSO-ICT-216284 SOCRATES,D3.1B, 2009.

[BJ00] F. Barceló and J. Jordán, *Channel holding time distribution in public telephony systems (PAMR and PCS)*, IEEE Transactions on Vehicular Technology **49** (2000), no. 5, 1615–1625.

[BM98] S.C. Borst and D. Mitra, *Virtual partitioning for robust resource sharing: computational techniques for heterogeneous traffic*, IEEE Journal on Selected Areas in Communications **16** (1998), no. 5, 668 – 678.

[BM07] E. Bernal-Mor, *Diseño óptimo de politicas de control de acceso del tipo multiple fractional guard channel para redes moviles celulares*, Tesina final de Master en Tecnologías, Sistemas y Redes de Comunicaciones, UPV (2007).

[BMGRPMB08] E. Bernal-Mor, D. Garcia-Roger, V. Pla, and J. Mart?nez-Bauset., *Optimal design of multiple fractional guard channel policy in multiservice cellular networks*, Proceedigs of UBICOMM'08, 2008, pp. 476–481.

[BMPGEMB12] E. Bernal-Mor, V. Pla, D. M. Gutierrez-Estevez, and J. Martinez-Bauset, *Resource management for macrocell users in hybrid access femtocells*, Proceedigs of GLOBECOM'12, 2012.

[BMPMB10a] E. Bernal-Mor, V. Pla, and J. Mart?nez-Bauset., *Analysis of different channel sharing strategies in cognitive radio networks*, In Proceeding MACOM'10, Lecture Notes in Computer Science (LCNS), vol. 6235, Springer-Verlag, 2010, pp. 70–73.

[BMPMB10b] _____, *Robust admission control for streaming and elastic services in cellular networks*, Proceedigs of ISCC'10, 2010, pp. 372–374.

[BMPMB12] E. Bernal-Mor, V. Pla, and J. Martinez-Bauset, *Handover performance for elastic flows in mobile cellular networks*, IEEE Communications Letters **16** (2012), no. 10, 1632–1635.

[Bon06] T. Bonald, *Insensitive queueing models for communication networks*, Valuetools '06: Proceedings of the 1st international conference on Performance evaluation methodolgies and tools, ACM Press, 2006, p. 57.

[BR03] T. Bonald and J.W. Roberts, *Congestion at flow level and the impact of user behaviour*, Computer Networks **42** (2003), 521–536.

[BS97] S.K. Biswas and B. Sengupta, *Call admissibility for multirate traffic in wireless ATM networks*, Proceedings of IEEE INFOCOM, vol. 2, 1997, pp. 649–657.

[C.-08] C.-X. Wang and H.-H. Chen and X. Hong and M. Guizani, *Cognitive radio network management*, Vehicular Technology Magazine, IEEE **3** (2008), no. 1, 28 – 35.

[CA09] V. Chandrasekhar and J. Andrews, *Spectrum allocation in tiered cellular networks*, IEEE Transactions on Communications **57** (2009), no. 10, 3059–3068.

[CAG08] V. Chandrasekhar, J. Andrews, and A. Gatherer, *Femtocell networks: A survey*, IEEE Communications Magazine **46** (2008), no. 9, 59–67.

[CBD02] T. Camp, J. Boleng, and V. Davies, *A survey of mobility models for ad hoc network research*, Wireless Communications and Mobile Computing **2** (2002), no. 5, 483–502.

[CC97] C.-C. Chao and Wai C., *Connection admission control for mobile multiple-class personal communications networks*, IEEE Journal

on Selected Areas in Communications **15** (1997), no. 8, 1618–1626.

[CHS08] H. Claussen, L. Ho, and L. Samuel, *An overview of the femtocell concept*, Bell Labs Technical Journal **13** (2008), no. 1, 221–245.

[CJP+11] Y. Choi, H. W. Ji, J.-Y. Park, H.-C. Kim, and J.A. Silvester, *A 3W network strategy for mobile data traffic offloading*, IEEE Communications Magazine **49** (2011), no. 10, 118–123.

[CL95] E. Chlebus and W. Ludwin, *Is handoff traffic really poissonian ?*, Proceedings of ICUPC'95, 1995, pp. 348–353.

[Cla07] H. Claussen, *Performance of macro-and co-channel femtocell in a hierarchical cell structure*, IEEE 18th international Symposium on Personal, Indoor and Mobile Rdio Communications, 2007, pp. 1–5.

[CLW95] G.L. Choudhury, K.K. Leung, and W. Whitt, *Efficiently providing multiple grades of service with protection agains overload in shared resources*, AT&T Technical Journal (1995), no. 74, (4):50–63.

[CNI04] T. K. Christensen, B. F. Nielsen, and V. B. Iversen, *Phase-type models on channel-holding times in cellular communication systems*, IEEE Transactions on Vehicular Technology **53** (2004), no. 3, 725–733.

[Com02] Federal Communication Commission, *Spectrum policy task force report*, Tech. report, FCC 02-155, 2002.

[CPVAOG04] F.A. Cruz-Perez, J.L. Vazquez-Avila, and L. Ortigoza-Guerrero, *Recurrent formulas for the multiple fractional channel reservation strategy in multi-service mobile cellular networks*, IEEE Communications Letters **8** (2004), no. 10, 629–631.

[DAR03] DARPA XG WG, *The XG architectural framework V1.0*, 2003.

[DGRP05] J. Martínez-Bauset D. Garcia-Roger, M. J. Domenech-Benlloch and V. Pla, *Adaptive trunk reservation policies in multiservice mobile wireless networks*, Lecture Notes in Computer Science **Management of Integrated Multimedia Services, J.Dalmau and G.Hasegawa (eds.)** (2005), no. 3754, 47–58.

[dlRVLPZ10] G. de la Roche, A. Valcarce, D. Lopez-Perez, and J. Zhang, *Access control mechanisms for femtocells*, IEEE Communications Magazine **48** (2010), no. 1, 33–39.

[DR09] D. Das and V. Ramaswamy, *Co-channel femtocell-macrocell deployments - access control*, IEEE 70th Vehicular Technology Conference Fall (VTC 2009-Fall), 2009, pp. 1–6.

[DS02] D. J. Daley and L. D. Servi, *Loss probabilities of hand-in traffic under various protocols. I. models and algebraic results*, Telecommunication Systems **19** (2002), no. 2, 209–226.

[e3] *End-to-End Efficiency project, E3*, ict-e3.eu.

[EC05] S.-E. Elayoubi and T. Chahed, *Admission control in the downlink of WCDMA/UMTS*, Mobile and Wireless Systems **LNCS 3427** (2005), no. 5, 136–151.

[Fan03] Y. Fang, *Thinning schemes for call admission control in wireless networks*, IEEE Transactions on Computers (2003), 685–687.

[FC02] Y. Fang and I. Chlamtac, *Analytical generalized results for handoff probability in wireless networks*, IEEE Transactions on Communications **50** (2002), no. 3, 396–399.

[GG09] J. M. Giménez-Guzmán, *Modelos para el análisis y optimización del control de admisión en redes celulare*, Ph.D. thesis, Universitat Politècnica de València, 2009.

[GGMBP07] J.M. Gutiérrez-Guzmán, J. Martínez-Bauset, and V. Pla, *A reinforcement learning approach for admission control in mobile multimedia networks with predictive information*, IEICE Transactions on Communications **E90-B** (2007), no. 7, 1663–1673.

[GJL84] D.P. Gaver, P.A. Jacobs, and G. Latouche, *Finite birth-and-death models in randomly changing environments*, Advances in Applied Probability **16** (1984), 715–731.

[GLZ07] J. Guo, F. Liu, and Z. Zhu, *Estimate the call duration distribution parameters in GSM system based on K-L divergence method*, International Conference on Wireless Communications, Networking and Mobile Computing, Shangai, China, September 2007, pp. 2988 – 2991.

[GMF08] Y. Pan G. Min and P. Fan, *Advances in wireless networks: Performance modelling, analysis and enhancement*, Nova Science Publishers, 2008.

[GMP09] A. Golaop, M. Mustapha, and L. B. Patanapongpibul, *Femtocell access control strategy in UMTS and LTE*, IEEE Communications Magazine **47** (2009), no. 9, 117–123.

[GRDBMBP05] D. Garcia-Roger, M. J. Domenech-Benlloch, J. Martinez-Bauset, and V. Pla, *Comparative evaluation of adaptive trunk reservation schemes for mobile cellular networks*, Proceedings of 3rd International Working Conference on Performance Modelling and Evaluation of Heterogeneous Networks (HET-NETs '05), July 2005.

[GRDBMBP07] _____, *Adaptative admission control in mobile multimedia networks with streaming and elastic traffic*, In Proceedings of the 20th International Teletraffic Congress (ITC-20),Ottawa,Canada, Juny 2007.

[GRMBP04] D. Garcia-Roger, J. Martínez-Bauset, and V. Pla, *Comparative evaluation of admission control policies in cellular multiservice networks*, Proceedings of the 16th International Conference on Wireless Communications (Wireless 2004), July 2004, pp. 517–531.

[GRMBP05] _____, *Admission control policies in multiservice cellular networks: Optimum configuration and sensitivity*, Mobile and Wireless Systems, LNCS 3427. Lecture Note in Computer Science (LCNS), vol. 3427, SPRINGER-VERLAG Berlin Heidelberg, 2005, pp. 121–135.

[Hay05] S. Haykin, *Cognitive radio: Brain-empowered wireless communications*, sac **23** (2005), no. 2, 201–20.

[HF01] J. Hou and Y. Fang, *Mobility-based call admission control schemes for wireless mobile networks*, Wireless Communications and Mobile Computing **1** (2001), no. 3, 269–282.

[HHS04] M. Hossain, M Hassan, and H. R. Sirisena, *Adaptive resource management in mobile wireless networks using feedback control theory*, Telecommunication Systems **24** (2004), no. 3-4, 401–415.

[HLL04] H.-N. Hung, P.-C. Lee, and Y.-B. Lin, *Random number generation for residual life of mobile phone movement*, Proceedigs of the IEEE ICNSC, 2004, pp. 30–33.

[How60] R. Howard, *Dynamic programming and markov processes*, MIT Press, 1960.

[HR86] D. Hong and S. S. Rappaport, *Traffic model and performance analysis for cellular mobile radio telephone systems with prioritized and nonprioritized handoff procedures*, IEEE Transactions on Vehicular Technology **VT-35** (1986), no. 3, 77–92.

[HR97] J.M. Hernandez-Rábanos, *Comunicaciones móviles*, Centro de Estudios Ramón Areces, Madrid, 1997.

[HSSK01] H. Hidaka, K. Saitoh, N. Shinagawa, and T. Kobayashi, *Teletraffic characteristics of cellular communication for different types of vehicle motion*, IEICE Transactions on Communications **E84-B** (2001), no. 3, 558–565.

[HSSK02] _____, *Self-similarity in cell dwell time caused by terminal motion and its effects on teletraffic of cellular communication networks*, IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences **E85-A** (2002), no. 7, 1445–1453.

[HUCPOG03a] H. Heredia-Ureta, F. A. Cruz-Pérez, and L. Ortigoza-Guerrero, *Capacity optimization in multiservice mobile wireless networks with multiple fractional channel reservation*, IEEE Transactions on Vehicular Technology **52** (2003), no. 6, 1519 – 1539.

[HUCPOG03b] _____, *Multiple fractional channel reservation for optimum system capcity in multi-service cellular networks*, Electronics Letters **39** (2003), no. 1, 133–134.

[Ive87] V.B. Iversen, *The exact evaluation of multi-service loss systems with access control*, Proceedings of the Teleteknik and Seventh Nordic Teletraffic Seminar (NTS-7), vol. 31, August 1987, pp. 56–61.

[JHJ05] S. S. Jeong, J. A. Han, and W. S. Jeon, *Adaptive connection admission control scheme for high data rate mobile networks*, Vehic-

ular Technology Conference, 2005. VTC-2005-Fall. 2005 IEEE 62nd, vol. 4, 2005, pp. 2607 – 2611.

[JL96] C. Jedrzycki and V.C.M. Leung, *Probability distributions of channel holding time in cellular telephony systems*, Proceedings of VTC'96, May 1996, pp. 247–251.

[JMMY09] H.-S. Jo, C. Mun, J. Moon, and J.-G. Yook, *Interference mitigation using uplink power control for two-tier femtocell networks*, IEEE Transactions on Wireless Communications **8** (2009), no. 10, 4906–4910.

[KGG10] D. K. Kim, D. Griffith, and N. Golmie, *A novel ring-based performance analysis for call admission control in wireless networks*, IEEE Communications Letters **14** (2010), no. 4, 324–326.

[Kol36] A.N. Kolmogorov, *Zur theorie der markoffschen ketten*, Mathematische Annalen **112** (1936), 155–160.

[LA95] C.-T. Lea and A. Alyatama, *Bandwidth quantization and states reduction in the broadband isdn*, IEEE/ACM Transactions on Networking **3** (1995), no. 3, 352–360.

[LLC98] B. L., C. Lin, and S. T. Chanson, *Analysis of a hybrid cutoff priority scheme for multiple classes of traffic in multimedia wireless networks*, Wireless Networks Journal (WINET) **4** (1998), no. 4, 279–290.

[LMN94] Y.-B. Lin, S. Mohan, and A. Noerpel, *Queueing priority channel assignment strategies for PCS hand-off and initial access*, IEEE Transactions on Vehicular Technology **43** (1994), no. 3, 704–712.

[LNH96] Y.-B. Lin, A. Noerpel, and D. Harasty, *The sub-rating channel assignment strategy for PCS hand-offs*, IEEE Transactions on Vehicular Technology **45** (1996), no. 1.

[LPVdlRZ09] D. Lopez-Perez, A. Valcarce, G. de la Roche, and J. Zhang, *OFDMA femtocells: A roadmap on interference avoidance*, IEEE Communications Magazine **47** (2009), no. 9, 41–48.

[LQK09] X. Li, L. Qian, and D. Kataria, *Downlink power control in co-channel macrocell femtocell overlay*, Proceeding of the 43rd annual conference on Information Sciences and Systems, 2009, pp. 383–388.

[LR99] G. Latouche and V. Ramaswami, *Introduction to matrix analytic methods in stochastic modeling*, ASA-SIAM, 1999.

[LYS10] Y.-Y. Li, L. Yen, and E.S. Sousa, *Hybrid user access control in HSDPA femtocells*, IEEE GLOBECOM 2010, 2010, pp. 679–683.

[Mac05] F. Machihara, *Mobile telecommunication systems and generalized Erlang loss formula*, IEICE Transactions on Communications **E88-B** (2005), no. 1, 183–189.

[Man08] G. Mansfield, *Femtocells in the US market-business drivers and consumer proposition*, Tech. report, FemtoCells Europe, ATT, 2008.

[MBGGP08] J. Martínez-Bauset, J. M. Giménez-Guzmán, and V. Pla, *Optimal admission control in multimedia mobile networks with handover prediction*, IEEE Wireless Communications **15** (2008), no. 5, 38–44.

[MBGRDB+09] J. Martínez-Bauset, D. Garcia-Roger, M. J. Domenech-Benlloch, , and V. Pla, *Maximizing the capacity of mobile cellular networks with heterogeneous traffic*, Computer Networks **53** (2009), no. 7, 973–988.

[MBPBM11] J. Martinez-Bauset, V. Pla, and E. Bernal-Mor., *Insensitive call admission control for wireless multiservice networks*, IEEE Communications Letters **15** (2011), no. 9, 989–991.

[MBPPP12] J. Martinez-Bauset, A. Popescu, V. Pla, and A. Popescu, *Cognitive radio networks with elastic traffic*, 8th Euro-NF conference on Next Generation Internet, 2012.

[MGM99] J. Mitola and JR. G. Maguire, *Cognitive radio: making software radios more personal*, IEEE Pers. Commun. **6** (1999), no. 6, 13–18.

[mon] *MONOTAS*, http://www.macltd.com/monotas.

[MRW98] D. Mitra, M.I. Reiman, and J. Wang, *Robust dynamic admission control for unified cell and call QoS in statistical multiplexers*, IEEE Journal on Selected Areas in Communications **16** (1998), no. 5, 692 – 707.

[MZ96] D. Mitra and I. Ziedins, *Virtual partitioning by dynamic priorities: Fair and efficient sharing by several services*, Broadband Communications, Proc. 1996 Int. Zurich Seminar Digital Commun. (1996), 173–185.

[Nel95] R. Nelson, *Probability, stochastic processes and queueing theory*, Springer-Verlag, 1995.

[Neu81] M. Neuts, *Matrix-geometric solutions in stochastic models: An algorithmic approach*, The Johns Hopkins University Press, 1981.

[NGM08] NGMN, *NGMN radio access performance evaluation methodology*, January 2008.

[NV] Nortel and Vodafone, *Open and closed access for home nodeBs*, Tech. report, 3GPP TSG-RAN WG 4(Radio).

[NV08] _____, *TS 36.211: Physical channels and modulation (release 8)*, Tech. report, 3rd Generation Partnership Project, 2008.

[OR01] P. V. Orlik and S. S. Rappaport, *On the handoff arrival process in cellular communications*, Wireless Networks Journal (WINET) **7** (2001), no. 2, 147–157.

[PB05] C. Prehofer and C. Bettstetter, *Self-organization in communication networks: principles and design paradigms*, Communications Magazine, IEEE **43** (2005), no. 7, 78 – 85.

[PCG03] V. Pla and V. Casares-Giner, *Optimal admission control policies in multiservice cellular networks*, Proceedings of the International Network Optimization Conference (INOC), October 2003, pp. 466–471.

[PCG05] _____, *Analysis of priority channel assignment schemes in mobile cellular communication systems: a spectral theory approach*, Performance Evaluation **59** (2005), no. 2-3, 199–224.

[PG85] E. C. Posner and R. Guérin, *Traffic policies in cellular radio that minimize blocking of handoff calls*, Proceedings of ITC 11, 1985.

[PGGMCG04] V. Pla, J. M. Giménez-Guzmán, J. Martínez, and V. Casares-Giner, *Optimal admission control using handover prediction in mobile cellular networks*, Proceedings of the 2nd International

Working Conference on Performance Modelling and Evaluation of Heterogeneous Networks (HET-NETs'04), July 2004, pp. 44/1–10.

[PMCG05] V. Pla, J. Martínez, and V. Casares-Giner, *Algorithmic computation of optimal capacity in multiservice mobile wireless networks*, IEICE Transactions on Communications **E88-B** (2005), no. 2, 797–799.

[PPPMB09] D. Pacheco-Paramo, V. Pla, and J. Martinez-Bauset, *Optimal admission control in cognitive radio networks*, CROWNCOM 2009, 2009.

[PRSADG05] J. Pérez-Romero, O. Sallent, R. Agustí, and M.A. Díaz-Guerra, *Radio resource management strategies in UMTS*, John Wiley, 2005.

[Ros70] S. M. Ross, *Applied probability models with optimization applications*, Holden-Day, 1970.

[Ros85] _____, *Introduction to probability models*, Academic Press, Inc., 1985.

[Ros95] K. W. Ross, *Multiservice loss models for broadband telecommunication networks*, Springer Verlag, 1995.

[RT01] M. Rajaratnam and F. Takawira, *Handoff traffic characterization in cellular networks under nonclassical arrivals and service time distributions*, IEEE Transactions on Vehicular Technology **50** (2001), no. 4, 954–970.

[RTN97] R. Ramjee, D. Towsley, and R. Nagarajan, *On optimal call admission control in cellular networks*, Wireless Networks Journal (WINET) **3** (1997), no. 1, 29–41.

[S.-09] S.-S. Tzeng, *Call admission control policies in cellular wireless networks with spectrum renting*, Computer Communitations **32** (2009), no. 18, 1905 – 1913.

[SBMS+11] B. Sas, E. Bernal-Mor, K. Spaey, V. Pla, C. Blondia, and J. Martinez Bauset, *An analytical model to study the impact of time-varying cell capacity in LTE networks*, Proceedigs of WMNC'11, 2011, pp. 1–8.

[SBMS+ed] _____, *An analytical model to study the impact of time-varying cell capacity in lte networks*, Telecommunication System (Accepted).

[SG00] M. Stasiak and M. Glabowski, *A simple approximation of the link model with reservation by a one-dimensional markov chain*, Performance Evaluation **41** (2000), 195–208.

[SHLK09] N. Saquib, E. Hossain, L. B. Le, and D. I. Kim, *Interference management in OFDMA femtocell networks: Issues and approaches*, IEEE Wireless Communications **57** (2009), no. 10, 3059–3068.

[SK04] W.-S. Soh and H. S. Kim, *Dynamic bandwidth reservation in cellular networks using road topology based mobility prediction*, Proceedings of IEEE INFOCOM, 2004.

[SNBH00] A.G. Spilling, A.R. Nix, M.A. Beach, and T.J. Harrold, *Self-organisation in future mobile communications*, Electronics & Communication Engineering Journal **12** (2000), 133–147.

[SNW08] E. Stevens-Navarro and V.W.S. Wong, *Virtual partitioning for connection admission control in cellular/WLAN interworking*, WCNC IEEE Conference, April 2008, pp. 2039–2044.

[soc] *SOCRATES*, www.fp7-socrates.eu.

[SR01] J. Siwko and I. Rubin, *Call admission control for capacity-varying networks*, Telecommunication Systems **16** (2001), no. 1-2, 15–40.

[SS97] M. Sidi and D. Starobinski, *New call blocking versus handoff blocking in cellular networks*, Wireless Networks Journal (WINET) **3** (1997), no. 1, 15–27.

[SSB10] K. Spaey, B. Sas, and C. Blondia, *Self-optimising call admission control for LTE downlink*, COST 2100 TD(10)10056, Joint Workshop COST 2100 SWG 3.1 & FP7-ICT-SOCRATES, Athens, Greece, February 2010.

[STKC09] D. Stratogiannis, G. Tsiropoulos, J. Kanellopoulos, and P. Cottis, *Probabilistic call admission control in wireless multiservice networks*, IEEE Communications Letters **13** (2009), no. 10, 746–748.

[XCA10] P. Xia, V. Chandrasekhar, and J. G. Andrews, *Femtocell access control in the TDMA/OFDMA uplink*, IEEE GLOBECOM 2010, 2010, pp. 1–5.

[XT03] A. E. Xhafa and O. K. Tonguz, *Does mixed lognormal channel holding time affect the handover performance of guard channel scheme ?*, Proceedings of IEEE GLOBECOM, 2003, pp. 3452–3456.

[YL02] F. Yu and V. Leung, *Mobility-based predictive call admission control and bandwidth reservation in wireless cellular networks*, Computer Networks **38** (2002), no. 5, 577–589.

[YMW+04] J. Yao, J.W. Mark, T. C. Wong, Y. H. Chew, K. M. Lye, and K.-C. Chua, *Virtual partitioning resource allocation for multiclass traffic in cellular systems with QoS constraints*, IEEE Transactions on Vehicular Technology **53** (2004), no. 3, 847 – 864.

[YR97] A. Yener and C. Rose, *Genetic algorithms applied to cellular call admission: local policies*, IEEE Transactions on Vehicular Technology **46** (1997), no. 1, 72–79.

[ZBA09] E. Zola and F. Barcelo-Arroyo, *Impact of mobility models on the cell residence time in WLAN networks*, Proceedigs of the IEEE SARNOFF'09, 2009, pp. 1–5.

[ZD97] M. M. Zonoozi and P. Dassanayake, *User mobility modeling and characterization of mobility patterns*, IEEE Journal on Selected Areas in Communications **15** (1997), no. 7, 1239–1252.

[Zha10] Y. Zhang, *Handoff performance in wireless mobile networks with unreliable fading channel*, IEEE Transactions on Mobile Computing **9** (2010), no. 2, 188 – 200.

[ZSXX11] G.-F. Zhao, Q. Shan, S. Xiao, and C. Xu, *Modeling web browsing on mobile internet*, IEEE Communications Letters **15** (2011), no. 10, 1081–1083.