Universidad de Málaga

Escuela Técnica Superior de Ingeniería de Telecomunicación

TESIS DOCTORAL

Modelling and Optimisation of GSM and UMTS Radio Access Networks

Autor:

Salvador Luna Ramírez

Directores:

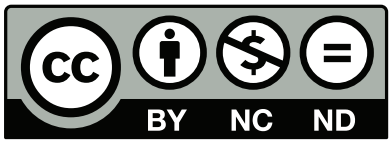
Matías Toril Genovés

Mariano Fernández Navarro

UNIVERSIDAD DE MÁLAGA
SPICUM
servicio de publicaciones

AUTOR: Salvador Luna Ramírez

EDITA: Servicio de Publicaciones de la Universidad de Málaga



Dr.D. Matías Toril Genovés y **Dr.D. Mariano Fernández Navarro**, profesores doctores del Departamento de Ingeniería de Comunicaciones de la Universidad de Málaga

CERTIFICAN:

Que **D. Salvador Luna Ramírez**, Ingeniero de Telecomunicación, ha realizado en el Departamento de Ingeniería de Comunicaciones de la Universidad de Málaga bajo su dirección, el trabajo de investigación correspondiente a su TESIS DOCTORAL titulada:

Modelling and Optimisation of GSM and UMTS Radio Access Networks

En dicho trabajo se han expuesto diversas aportaciones originales, entre las que cabe destacar un modelo de colas con reintento y correlación entre conexiones para el análisis de canales de señalización dedicados y la obtención de un indicador analítico óptimo para el balance de tráfico entre celdas, ambas en tecnología GSM. También se ha propuesto un modelo de auto-ajuste de parámetros basado en lógica difusa para un escenario heterogéneo conjunto GSM-UMTS. Los resultados expuestos han dado lugar a publicaciones en revistas y aportaciones a congresos internacionales.

Por todo ello, consideran que esta Tesis es apta para su presentación al tribunal que ha de juzgarla. Y para que conste a efectos de lo establecido en el artículo 8° del Real Decreto 778/1998 y Real Decreto 56/2005, reguladores de los Estudios de Tercer Ciclo-Doctorado, AUTORIZAN la presentación de esta Tesis en la Universidad de Málaga.

Málaga, a ____ de _____ de 2010

Fdo.: Dr.D. Matías Toril Genovés

Fdo.: Dr.D. Mariano Fernández Navarro

E.T.S.I.Telecomunicación, Campus de Teatinos, 29071-MÁLAGA, Tlf.: 952131440, Fax: 952132927

UNIVERSIDAD DE MÁLAGA
ESCUELA TÉCNICA SUPERIOR DE INGENIERÍA DE TELECOMUNICACIÓN

Reunido el tribunal examinador en el día de la fecha, constituido por:

Presidente: Dr. D. _____

Secretario: Dr. D. _____

Vocales: Dr. D. _____

Dr. D. _____

Dr. D. _____

para juzgar la Tesis Doctoral titulada **Modelling and Optimisation of GSM and UMTS Radio Access Networks** realizada por D. Salvador Luna Ramírez y dirigida por el Dr. D. Matías Toril Genovés y el Dr. D. Mariano Fernández Navarro,

acordó por _____

otorgar la calificación de _____

y para que conste, se extiende firmada por los componentes del tribunal la presente diligencia.

Málaga, a ____ de _____ de 2010

El Presidente: El Secretario:

Fdo.: _____ Fdo.: _____

El Vocal: El Vocal: El Vocal:

Fdo.: _____ Fdo.: _____ Fdo.: _____

Cómo no...
a papá, mamá,
Rocío (y familia) y Pilar

Acknowledgements

Saying that this work would not have been carried out without the help of other people is more than a common statement. I feel it like a complete truth.

First, I wish to thank the help from my supervisors, Mariano and Matías. Their help have been essential at different stages, and with complementary functions, along this thesis. Thanks once again. This work is also yours.

I would also like to thank all my colleagues, and, concretely those in the mobile network optimisation group. Special mention to Fernando, remembering all the questions, doubts, discussions, and time spent in our simulator development. It is also worthwhile to mention all the people from the *Wireless Access Research Center*, and Sean McGrath and Ronan Skehill in a special way. Thanks for your kind welcome and all the help you gave me during my days in Limerick.

The financial support given by the projects mentioned below, together with the research group TIC-102 *Ingeniería de Comunicaciones*, has been equally important. They all made possible to present my work, and learn from other colleagues, in conferences, journals, workshops and research projects. I don't want to omit all people involved in the European GANDALF project, where I learned so much and met many brilliant people.

Finally, regarding less technical but equally important acknowledgements, I strongly thank my family for insisting me to work hard in my thesis. I don't know how much additional time I would have needed if they haven't insisted. Also thanks to my closer friends, always asking and encouraging me to press ahead with this work. And thanks to God, who always walked next to me in this road.

With no doubt, I spent less time in this page than in any other one in my thesis...

This work has been supported by the Excellence Research Program (Andalusian government, project TIC-4052) and by the Spanish Ministry of Science and Innovation (grant TEC2003-07827, TEC2008-06216 and TEC2009-13413).

Contents

Abstract

The size and complexity of mobile communication networks have increased in the last years making network management a very complicated task. GSM/EDGE Radio Access Network (GERAN) systems are in a mature state now. Thus, non-optimal performance does not come from typical network start-up problems, but, more likely, from the mismatching between traffic, network or propagation models used for network planning, and their real counterparts. Such differences cause network congestion problems both in signalling and data channels. With the aim of maximising the financial benefits on their mature networks, operators do not solve anymore congestion problems by adding new radio resources, as they usually did. Alternatively, two main strategies can be adopted, a) a better assignment of radio resources through a re-planning approach, and/or b) the automatic configuration (optimisation, in a wide sense) of network parameters. Both techniques aim to adapt the network to the actual traffic and propagation conditions. Moreover, a new heterogenous scenario, where several services and Radio Access Technologies (RATs) coexist in the same area, is now common, causing new unbalanced traffic scenarios and congestion problems. In this thesis, several optimisation and modelling methods are proposed to solve congestion problems in data and signalling channels for single- and multi-RAT scenarios.

First, a new proposal intends to solve the dimensioning of dedicated signalling channels in GERAN. Current models do not consider retrials nor time correlation between arrivals in signalling traffic. Thus, congestion problems arise even when idle resources can be found. A new signalling traffic and system model is proposed, which can be tuned by means of network performance statistics. This proposal is validated with the comparison between model performance indicators and live network statistics. Such a novel model can be used by operators by means of re-assigning traffic resources more efficiently.

Secondly, an optimal load sharing scheme is described for localised congestion caused by the non-uniform spatial concentration of traffic demand in GERAN. Two network models are constructed and analytical expressions are obtained as a balancing criterion to perform an optimal traffic sharing between cells. Traffic sharing is carried out through the modification of radio resource management algorithm parameters. Optimal traffic sharing criterion is compared with other heuristic load balancing criteria. To assess the optimal traffic sharing criterion, several heuristics methods, used by operators, are compared with the optimal method in several realistic scenarios constructed from a live network.

Finally, a parameter auto-tuning scheme is proposed in an scenario with a strongly unbalanced traffic time distribution and an heterogenous network comprising UMTS and GSM technologies. This scheme contains a Fuzzy Logic Controller (FLC) as the central entity, modifying parameters from the inter-system handover algorithm. Parameter changes are guided by the analysis of joint network performance indicators. To validate such scheme, a joint network-level simulator has been developed for GSM and UMTS. Load sharing capabilities have been tested and different configurations for the FLC were checked out in order to speed up the convergence of the auto-tuning process.

Resumen

Las redes de comunicaciones móviles actuales han crecido significativamente en tamaño y complejidad durante los últimos años. Así, la gestión de la red se ha convertido en una tarea complicada. Las redes de comunicaciones móviles GSM/EDGE (GERAN) se encuentran actualmente en una fase madura en su desarrollo y, por tanto, un posible rendimiento infra-óptimo no vendría originado por los problemas típicos que surgen en la puesta en marcha de la red, sino, posiblemente, por aquellos problemas que surgen debido a las diferencias existentes entre los modelos y predicciones usados en la planificación de red original y las condiciones y características actuales de tráfico. Estas diferencias provocan problemas de congestión en canales de datos y señalización. Los operadores ya no solucionan estos problemas de congestión añadiendo recursos radio adicionales. En vez de esto, y con el objetivo de maximizar los beneficios en sus redes ya existentes, se pueden diferenciar dos tipos de estrategias, a) la reasignación de los recursos radio existentes con métodos de replanificación de la red, y, b) la configuración automática (optimización, en un sentido amplio) de parámetros de red. Ambas técnicas buscan una mejor adaptación de la red a las condiciones actuales de tráfico. Además, en la actualidad también existen escenarios heterogéneos, donde varios servicios y redes de acceso radio conviven bajo un mismo área geográfica. Este nuevo escenario provoca nuevas situaciones de desequilibro de tráfico, y, por tanto, congestión en la red. Esta tesis propone diversos métodos de modelado y optimización para solucionar problemas de congestión en canales de datos y señalización, tanto para escenarios con una sola tecnología radio como con varias.

En primer lugar, una primera propuesta intenta solventar el problema del dimensionamiento de los canales dedicados de señalización en GERAN. Los modelos actuales no consideran las características de reintento ni de correlación en el tráfico de señalización dedicado, y, por ello, aparecen problemas de congestión en los canales de señalización, aun habiendo recursos disponibles. Teniendo en cuenta las características de reintento y correlación, se propone un nuevo modelo que, además, es ajustado con datos recogidos de una red real. Esta nueva propuesta es validada a través de la comparación entre los indicadores de rendimiento del modelo y los estadísticos de red real. El nuevo modelo de tráfico de señalización proporciona al operador una herramienta muy útil para implementar una reasignación más eficiente de los recursos radio en la red.

En segundo lugar, se describe un esquema de reparto de carga para solventar congestiones locales causadas por la concentración espacial no uniforme de tráfico de voz en GERAN. Se han construido varios modelos de red y, a partir de ellos, se han obtenido

expresiones analíticas como criterio de balance para implementar un reparto de carga óptimo entre celdas. El reparto de tráfico es llevado a cabo a través de la modificación de parámetros en algoritmos de gestión de recursos radio. Para la evaluación de las propuestas óptimas de reparto de carga, se realiza una comparación con otros criterios heurísticos de reparto usados por los operadores en diversos escenarios de red realistas.

Finalmente, se propone un esquema de auto-ajuste de parámetros en un escenario de tráfico cambiante a lo largo del tiempo, dentro de una red heterogénea con tecnologías GSM y UMTS. El esquema de auto-ajuste tiene como entidad principal un controlador basado en lógica difusa que modifica parámetros del algoritmo de traspaso inter-sistema con el objetivo de alcanzar el balance de carga entre tecnologías. Los cambios en los parámetros están guiados por el análisis de los estadísticos de rendimiento de la red. La validación de este esquema se ha realizado sobre un simulador conjunto de redes GSM y UMTS. Dicha validación se centra en la capacidad del esquema propuesto para ejecutar el reparto de carga entre tecnologías. De manera adicional, se han probado distintas configuraciones del controlador difuso con el objetivo de acelerar el proceso de convergencia en el auto-ajuste de parámetros.

List of Abbreviations

AC	Admission Control
BH	Busy Hour
BCR	Blocked Call Rate
BLER	BLock Error Rate
BPB	Blocking Probability Balancing
BR	Blocking Ratio
BSC	Base Station Controller
BSS	Base Station Subsystem
BTB	Blocked Traffic Balancing
BTS	Base Transceiver Station
CDMA	Code Division Multiple Access
CHT	Channel Holding Time
CPICH	Common PIlot CHannel
CR	Congestion Ratio
CRS	Cell ReSelection
CRRM	Common Radio Resource Management
CS	Circuit Switched
DCCH	Dedicated Control Channel
EC	Emergency Call
EDGE	Enhanced Data rates for GSM Evolution
FDMA	Frequency Division Multiple Access
FLC	Fuzzy Logic Controller
FER	Frame Error Rate
GERAN	GSM-EDGE Radio Access Network
GH	GHost seizure
GoS	Grade of Service
GSM	Global System for Mobile communications
GSM2U	GSM to UMTS
HO	HandOver
HOC	HandOver Control
ID	IMSI Detach
IBP	Incremental Blocking Probability
IS-CR	Inter-System Cell Reselection

IS-HO	Inter-System HandOver
JAC	Joint Admission Control
JRRM	Joint Radio Resource Management
KPI	Key Performance Indicator
LA	Location Area
LB	Load Balancing
LTE	Long Term Evolution
LU	Location Update
MCD	Mean Call Duration
MHT	Mean Holding Time
MMS	Multimedia Messaging System
MOC	Mobile Originated Call
MTC	Mobile Terminated Call
NMS	Network Management System
NSS	Network and Switching System
OB	Optimal Balancing
OFDMA	Orthogonal Frequency Division Multiple Access
OSS	Operation Support System
PASTA	Poisson Arrivals See Time Averages
PCU	Packet Control Unit
PI	Performance Indicator
POC	Power Control
QoS	Quality of Service
RACH	Random Access CHannel
RAT	Radio Access Technology
RE	call Re-Establishment
RM	Retrial Model
RMCA	Retrial Model with Correlated Arrivals
RNC	Radio Network Controller
RRM	Radio Resource Management
RxLEV	Received signal LEVel
SIR	Signal to Interference Ratio
SMS	Short Message Service
SNR	Signal to Noise Ratio
SON	Self-Organised Networks
SS	Supplementary Service
SDCCH	Stand-alone Dedicated Control CHannel
TCH	Traffic CHannel
TDMA	Time Division Multiple Access
TR	Transition Rate
TSL	Time SLot
TTT	Time To Trigger
UE	User Equipment
UMTS	Universal Mobile Telecommunication System

U2GSM	UMTS to GSM
WAP	Wireless Application Protocol
WCDMA	Wideband Code Division Multiple Access
WLAN	Wireless Local Area Network
2D	Two dimensions
2G	Second Generation
2.5G	Second and a half Generation
3D	Three dimensions
3G	Third Generation

List of Symbols

Chapter 2 and Appendix A

A_c	Carried traffic (measured)
$A_{c_{rm}}$	Carried traffic estimate for RM
$A_{c_{rmca}}$	Carried traffic estimate for RMCA
α	Retrial rate (inverse of the mean time between retrials)
BR	Blocking ratio (measured)
BR_{rm}	Blocking ratio estimate for RM
BR_{rmca}	Blocking ratio estimate for RMCA
CR	Congestion ratio (measured)
CR_{rm}	Congestion ratio estimate for RM
CR_{rmca}	Congestion ratio estimate for RMCA
$\mathbf{D_m}$	Sub-matrix located at the diagonal of $\mathbf{Q}$ matrix
$\overline{e}$	Column vector of ones
E	Erlangs
$\mathbf{L_m}$	Sub-matrix located at the lower diagonal of $\mathbf{Q}$ matrix
$\lambda_{LU}, \lambda_{MOC}...$	Arrival rate for different signalling services
λ_r	Arrival rate for retrial services
λ_{nr}	Arrival rate for non-retrial services
$\lambda_{LU_{on}}, \lambda_{LU_{off}}$	LU arrival rate during the *on* and *off* periods
$\lambda_{r_{on}}, \lambda_{r_{off}}$	Retrial services arrival rate during the *on* and *off* periods
M	Maximum number of users in the orbit
μ	Service rate (inverse of the mean channel holding time)
N	Number of sub-channels)
$N_{LU}, N_{MOC}...$	Number of signalling attempts in one hour for different signalling services
N_s	Number of system states
N_{sam}	Number of measured samples in NMS
$NSAE_{brgs,m}$	Normalised sum of BR absolute errors of revenue-generating services
$O(f(n))$	Set of functions that grow no faster than $f(n)$
$\Pi(i,j)$	Probability of having i busy sub-channels and j users in the orbit
$\Pi(i,j,k)$	Probability of having i busy sub-channels, j users in the orbit and k state
$\mathbf{Q}$	Infinitesimal generator matrix
$\rho_{on\text{-}off}, \rho_{off\text{-}on}$	Switching rates between LU *on* and *off* states
r	Ratio between the duration of *on* and *off* periods

$\overline{SSE}_m$	Average sum of squared errors for performance indicators with method m
θ	Retrial probability
τ_{on}, τ_{off}	mean duration of the *on* and *off* states in RMCA
T_c	Duration of an *on-off* cycle
$TR_{u,v}$	Transition rate from state u to state v
$\mathbf{U_m}$	Sub-matrix located at the upper diagonal of $\mathbf{Q}$ matrix

Chapter 3 and Appendix B

A_{bT}	Total blocked traffic
A_{bi}	Blocked traffic in cell i
a_i	Coverage area of cell i
A_i	Traffic offered to cell i
A_{lbi}, A_{ubi}	Lower and upper bounds on cell i traffic due to spatial considerations
A_T	Total offered traffic in the network
$\beta(A_i, A_{fi}, c_i)$	Balance indicator for optimal traffic sharing in refined model
c_i	Number of channels of cell i
C_m	Relative network capacity
$C_{m,const}$	Relative network capacity with constraints
CHT_i	Channel Holding Time in cell i
Δ	Deviation parameter for traffic bounds
$E(A_i, c_i)$	Erlang-B blocking probability for cell i
L_i	Average traffic load in cell i
λ_f	Arrival rate of fresh calls
λ_{ho}	Arrival rate of handover requests
λ_i	Arrival rate of calls to cell i
λ_T	Overall user call rate
MCD	Mean call duration
MHT	Mean holding time
μ	Service rate
μ_{ci}	Average connection service rate in cell i
N	Number of base stations
r_{cvg}	Coverage radius
s_i	Service area of cell i
θ_{cvg}	Maximum angle off the antenna bearing

Chapter 4

BCR_{GSM}	Blocked call rate for GSM Radio Access Technology

BCR_{UMTS}	Blocked call rate for UMTS Radio Access Technology
BCR_{total}	Blocked call rate in all Radio Access Technologies
CR_{y_z}	Congestion ratio in link y for radio technology z
$\Delta OFFcell_{U2GSM}$	Increment value for $OFFcell_{U2GSM}$ parameter
ΔT_{3A_U2GSM}	Increment value for T_{3A_U2GSM} parameter
$DEVoff_{U2GSM}$	Deviation of the current offset value from the default one
E_c/N_o	Signal to interference and noise energy level ratio
H_{3A_UMTS}	Hysteresis parameter for U2GSM IS-HO
H_{3A_GSM}	Hysteresis parameter for U2GSM IS-HO
i	Origin cell in IS-HO algorithm
j	Destination cell in IS-HO algorithm
$\mu_x(CR_{y\_z})$	Membership function of linguistic term x for CR_{y_z}
$OFFcell_{U2GSM}$	Offset parameter for U2GSM IS-HO
$OFFcell_{GSM2U}$	Offset parameter for GSM2U IS-HO
S_c	Configuration c for fuzzy controller
T_{3A_U2GSM}	UMTS quality level lower bound for U2GSM IS-HO
T_{3A_GSM}	GSM signal level lower bound for U2GSM IS-HO
T_{3A_GSM2U}	GSM signal level lower bound for GSM2U IS-HO
TTT_{3A_U2GSM}	Time To Trigger for event 3A in U2GSM IS-HO

Chapter 1

Modelling and Optimisation in Mobile Networks

With the aim of offering an easy reading of this thesis, this first chapter introduces the main topics covered in this work. The main research objectives are enumerated later in this chapter, and the structure of the document is described to conclude.

A mobile communication network can be considered as an extremely complicated engineering work. A lot of entities, protocols, terminals, algorithms, etc, are grouped into one of the most challenging and successful communication systems. The design of a mobile network needs a lot of procedures that, taking into account the initial target criteria (grade of service, coverage area,...), result in how the network must be structured, the radio resources distributed or protocols designed. Such a design process uses lots of models for those calculations. For instance, the designer uses models for the user behaviour, traffic/service rates, radio resource management or mobile channel attenuation. These models work after certain simplifications about the object to be modelled. Good models (i.e., those close to the real behaviour) allow not only a good network functioning, but also significant savings in time and effort, since good models avoid the excessive repetition of the *design-implementation-test* process.

In parallel to mobile network evolution, models also get complicated, creating new sources of innacuracies. The models used for the network design are often not close to the real network behaviour. Likewise, with time evolution, the initial design conditions are no longer valid, so the network configuration is not suitable for the present conditions. As a consequence, the initial network design does not perform properly, which is usually reflected in congestion or blocking problems. Different approaches can be applied to solve these problems by trying to make the network performs better. These enhancement approaches are called, in a wide sense, *network optimisation*, since they optimise (i.e., improve) the network performance through changes in some network characteristic.

Whatever the optimisation strategy is, network models are necessary to design how the network must be optimised (e.g., how radio resources must be re-distributed, or how a network parameter must be re-configured).

1.1 Network Modelling

Building a model consists of developing an entity or idea which reflects a portion of reality. A model imitates the reality. Such a model tries to behave as similar as possible to the real object, although, with generalisation purposes, it usually assumes some simplifications about the real behaviour of the system to be modelled. In mobile communications, a network model is constructed, basically, to make the analysis and development of a mobile network easier. A network model is widely used as the main platform to predict the network behaviour and performance when an specific configuration, algorithm or network structure is tested.

During the design of a mobile network, a network model is essential. A network model tries to make a good prediction (i.e., as close as possible to the real behaviour) of the performance of the network to be developed later on. In this scenario, a network model where experiments can be performed over is very useful. These experiments during the network design process can save a lot of problems later when the real mobile network is developed.

When a mobile network has already been developed and services are being offered to the user, the use of a network model becomes different. In this scenario where a mobile network is already functioning, traffic and technology evolution demand changes in mobile network configuration. The application of changes to an old mobile network configuration could be also tested over the real mobile network. This approach would result in absolutely exact (and real) measures, and the conclusions would be so irrefutable when results are obtained from live tests. However, this strategy is not possible, since it implies different and important disadvantages:

a) Live tests have a high operational cost. A lot of technical staff is necessary to carry out this realistic assessment, so live tests are most often used for definitive tests about new policies or network structures (after a previous and long stage using network models).

b) There is a high risk of service outage during a real test over a live mobile network. This is logically critical for operators, so they seldom (and very reluctantly) allow live tests.

c) Unlike network models, the environment and scenario where the test is being developed (user movements and directions, incoming traffic,...) is not under the operator control. Consequently, the result can be obtained under circumstances which are not the most suitable for the test to be developed (e.g., a new policy for a high traffic scenario is tested one day with a very low traffic).

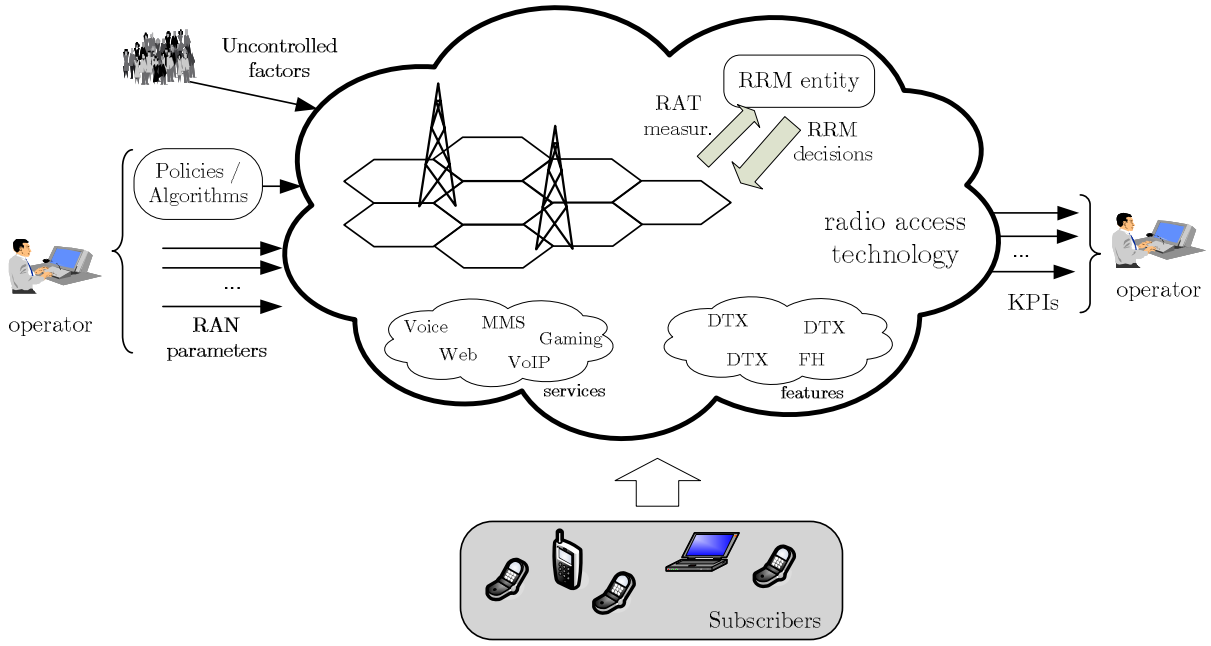


Figure 1.1: Description of a mobile communication system.

Testing changes over a network model would avoid such disadvantages. Instead of using the real mobile network, the use of a network model could asses how a new network configuration or algorithm could manage and how the network would perform, saving time and effort (and, consequently, money).

Mobile networks are extremely complex systems. A lot of entities, algorithms, protocols, parameters configure a mobile communication system. Figure 1.1 illustrates this concept, showing some of the entities involved in mobile networks. Operator parameter configurations, external (and uncontrolled) factors, services and features offered to subscriber or RRM procedures are some of the main entities in a mobile network.

Figure 1.2 plots the reference architecture assumed in this work for a GERAN/UTRAN mobile network where main entities and protocols are included. The figure shows three main systems in a cellular network: the Base Station System (BSS), providing the path between Mobile Stations (MS) and the fixed infrastructure, Network and Switching System (NSS) and Operation Support System (OSS). BSS contains the specific elements in a radio cellular network. In GERAN, BSS contains the Base Transceiver Station (BTS) and Base Station Controller (BSC), comprising radio transmission/reception equipment and control elements of a group of BTSs, respectively. Respective elements in a UTRAN network are the User Equipment (UE), Node B and Radio Network Controller (RNC).

Network models usually do not comprise all the functionalities and characteristics live networks really have, but only those having a large impact over the network performance indicators the designer is interested in. Actually, to decide which functions/characteristics must be considered in the model is a very important point when a network model is being constructed. The more functionalities are considered, the lower risk of getting unrepresentative results, but, at the same time, model construction and assessment is becoming more complicated. In any case, network models are simplifications of the real

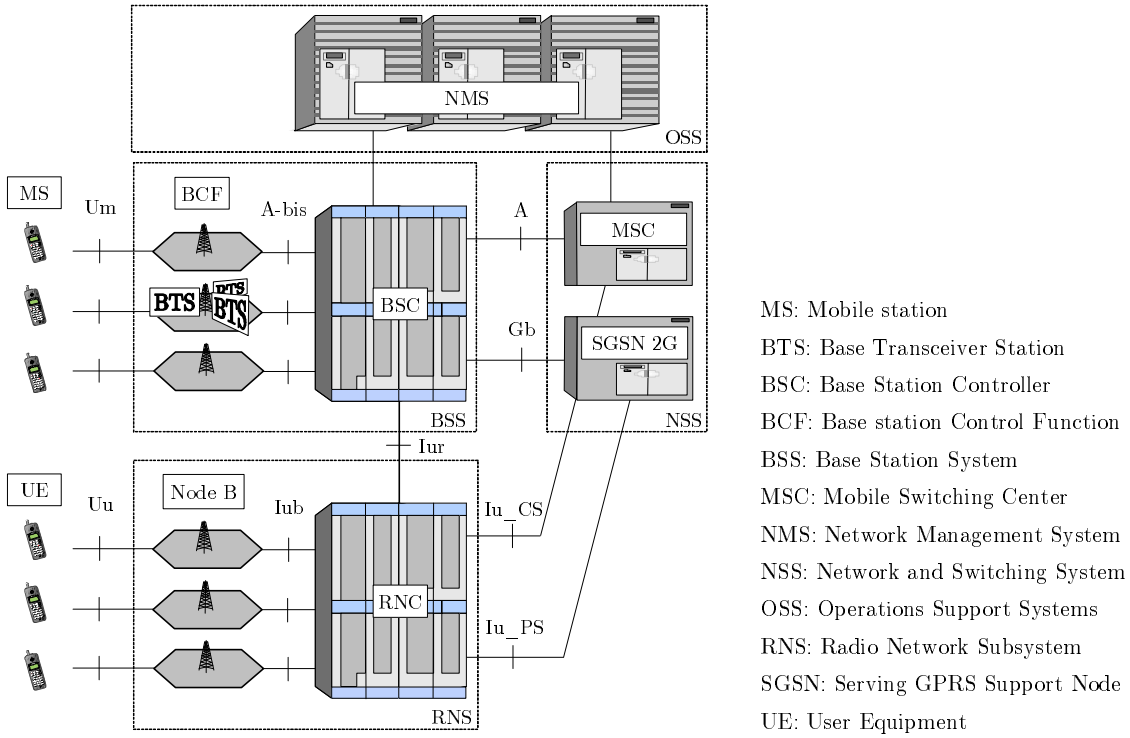


Figure 1.2: A GERAN/UTRAN reference architecture.

system.

Figure 1.3 depicts a broad classification of mobile network modelling. It is divided into two main and different categories depending on a) the application area of the mobile network model, and b) the method of network model design. In the first category, different models are constructed depending on their main use. Four different application areas, with increasing complexity have been identified here, namely,

a) *Performance data analysis.* This is the first step in network modelling. A model is constructed and, next, main performance results are obtained and analysed. This methodology is widely used for the design of mobile networks, previously to their deployment. If possible, model performance results are usually compared to some available live measurements in order to validate the network model itself.

b) *Model parameter tuning.* Once the model is validated, an additional step consists of getting an adequate configuration for model parameters. Since the model is expected to be realistic, testing different model configurations is an useful strategy in order to get a better live network performance. Different sensitivity analysis can be made up over different model parameter settings. At the end of the process, adequate values are exported to the live network.

c) *Impact estimation.* As previously stated, mobile networks are constantly evolving to adapt to the continuously changing environment. Changing some live network characteristic or functionality is a high risk activity, and network performance after the change must be previously assessed over a network model, since no mistakes or network fallouts should occur when a live network is being modified. In this

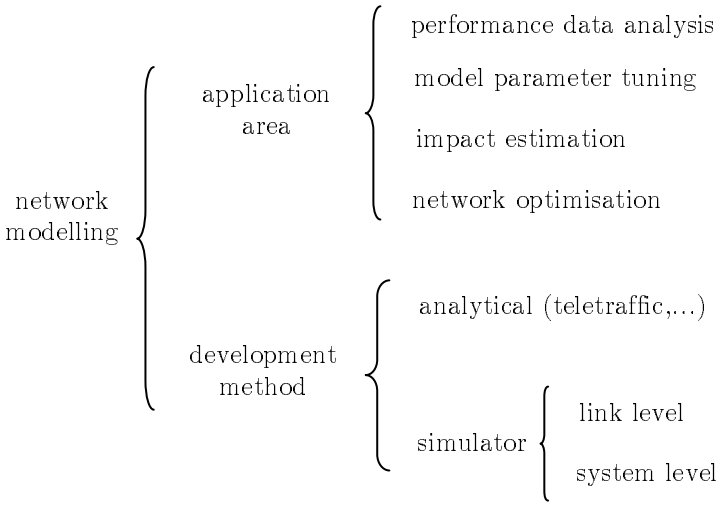


Figure 1.3: A taxonomy of mobile network modelling approaches.

scenario, a network model is a very useful tool to predict the effect of changing some network functionality (e.g., a radio resource management algorithm or the hierarchy structure). The design and development of these models are strongly coupled to the network characteristic to be changed, disregarding in other mobile network functionalities.

d) *Network optimisation.* Depending on their design method, some mobile network models are constructed in such a way that mathematical equations can be extracted for network performance indicators. In this case, classical (i.e., mathematical) optimisation techniques can be applied over the network model, and, then, optimal network performance can be obtained and exported later to a live network. Network models in this category are constructed under teletraffic engineering principles.

A second main classification for network modelling in Figure 1.3 is broken down by how the network model has been developed. Two possibilities can be here found, a) models based on analytical expressions, and, b) based on simulators.

Although several techniques can be cathegorised as analytical modelling (e.g., event-based or deterministic models), teletraffic engineering is the main approach for mobile network modelling. The purpose of teletraffic engineering is intended the "*application of probability theory to the solution of problems concerning planning, performance evaluation, operation and maintenance of telecommunication systems*", [1]. The core of teletraffic is, then, the application of probability theory to model the main parts of a telecommunication system (i.e., users, resources, policies,...). A teletraffic model contains three main parts: the *structure* of the system to be modelled (or part of it), the policies to be applied over the structure (referred to as *strategy*), and statistical properties of the *traffic* incoming the model. In present mobile network models, these three parts are clearly identified by the mobile network structure, the different algorithms and management algorithms (e.g. resource management techniques, frequency planning, etc.), and mobile user behaviour, respectively.

As an advantage, teletraffic models make the derivation of analytical expressions for the main network performance indicators possible, so that classical optimisation techniques can be applied. Nevertheless, models based on teletraffic assume a higher number of simplifications for constructing the model. These simplifications are mainly of a statistical nature (e.g. how the users move or incoming user ratios). Then, as a disadvantage, teletraffic model development requires many assumptions (especially about the statistical behaviour of the traffic and network structure) that might not always be valid.

A second method for model development is based on simulators. A mobile network simulator is a system programmed to perform similarly to a real mobile network. Most of the real mobile network features are included and a mobile simulator is usually quite close to the real system behaviour, including those random elements which are unpredictable, namely source traffic and propagation channel characteristics. Simulators use models for those elements which must be constructed assuming simplifications or random behaviour. Although deterministic elements are programmed in a simulator exactly as they are included in real networks (e.g., AC algorithm, power control calculations), elements in the simulator with a random behaviour are programmed according to different statistic models, trying to reflect their real behaviour as exactly as possible.

Mobile network simulators are classified depending on the part of the system they are focused on. There exist link level and system level simulators. Link level simulators emulate the performance of a mobile radio link between one user and its base station. These simulators are quite focused on the behaviour of an specific link and their design (radio modulation, power control approach, channel behaviour,...), so a big effort is spent on the programming and definition of the radio link, propagation channel and all parameters associated. The main output offered by link simulators is a set of curves of Bit or Block Error Rates (BER or BLER) versus Signal-To-Noise or Signal-to-Interference Ratios (SNR or SIR, respectively), depending on the radio technology to simulate. These results are used as a core element in system level simulators.

System level simulators, or network simulators, can be considered as a higher layer tool in mobile network simulations, and they model and predict the global network performance under some concrete conditions or scenario, comprising very wide simulation areas. In contrast to its link level counterpart, network simulators do not focus on the radio link behaviour (actually BER to SNR curves, the main output in link level simulators, are used as an input to resume and simplify the radio channel behaviour), but on the global network performance monitoring the UE experience along a service requested to the network. The main design effort is put into the programming of the different algorithms (e.g., RRM algorithms) and models (traffic source models) characterising a mobile network system. Global performance indicators (e.g., global blocking rates or handover ratios) are the main output in these simulators. Due to the random behaviour of some network elements, specially traffic sources, reliable results (i.e., representative of the network behaviour) can be obtained only after long simulation instances.

Figure 1.4 plots a typical system level simulator structure. The figure shows the

main processes along a simulation, and it also includes the main models and algorithms applied (dashed boxes). Simulation flow starts with the network initialisation, i.e., the construction of the simulation scenario (e.g., BTS location, antenna diagram, or scenario size) according with the configuration parameters previously supplied. Once the scenario is built, it is important to notice that the network is initially empty, without users. Final network performance indicators can be altered due to this initial simulation stage. A warm-up module creates an initial traffic intensity (i.e., there are on-going services at $t = 0$), accordingly to the traffic rates to apply for the remaining simulation time. This warm-up technique allows the collection of representative performance statistics from $t = 0$, saving some simulation time.

An iterative process take place now, where successive iterations represent simulated time evolution. A first module in a new iteration calculates new user positions. Mobility models are necessary for this movement calculation. Then, user movements cause changes in received signal levels from BTSs, so they must be calculated again according to the propagation model used by the simulator. Both received signal levels and interference levels for each user are obtained in this propagation module. Later, signal levels together with BER/SIR mapping curves supplied by link level simulators allow the calculation of quality link indicators (BER or BLER). Quality indicators are the main parameter for many RRM algorithms (handover, drop calls, resource reassignment,...). Finally in this iterative process, new users can arise, or finish, so new/finish call operation must be activated. Service generation needs of a detailed traffic source model and a radio resource management entity. At the end of this stage, radio resources are suitably assigned to new users, and ending services release the resources they were occupying.

All previous modules in Figure 1.4 update multiple counters, trying to witness all the events taking place during the simulation time. After the iterative process, there are still two additional modules in charge of results management. First, saved counters are processed in order to extract global performance indicators. Second, results are saved for later processing or analysis.

Generally speaking, simulators are the best option when a very complicated system must be modelled. Nevertheless, simulators are hard to code and manage, and the computation of reliable performance indicators requires more time and effort. As advantages, they are more realistic systems and suitable when only a few assumptions about the mobile network can be made. Simulators are quite flexible in their configuration and they are a very powerful tool when different scenarios have to be tested and analysed.

1.2 Network Optimisation

Mobile operators aim to reduce operational and capital expenditures as much as possible nowadays, especially in mature technologies. A continuously increasing level of competition causes cellular operators to provide many services at a minimum cost. In this scenario, operators try to maximise the performance of their existing networks, while

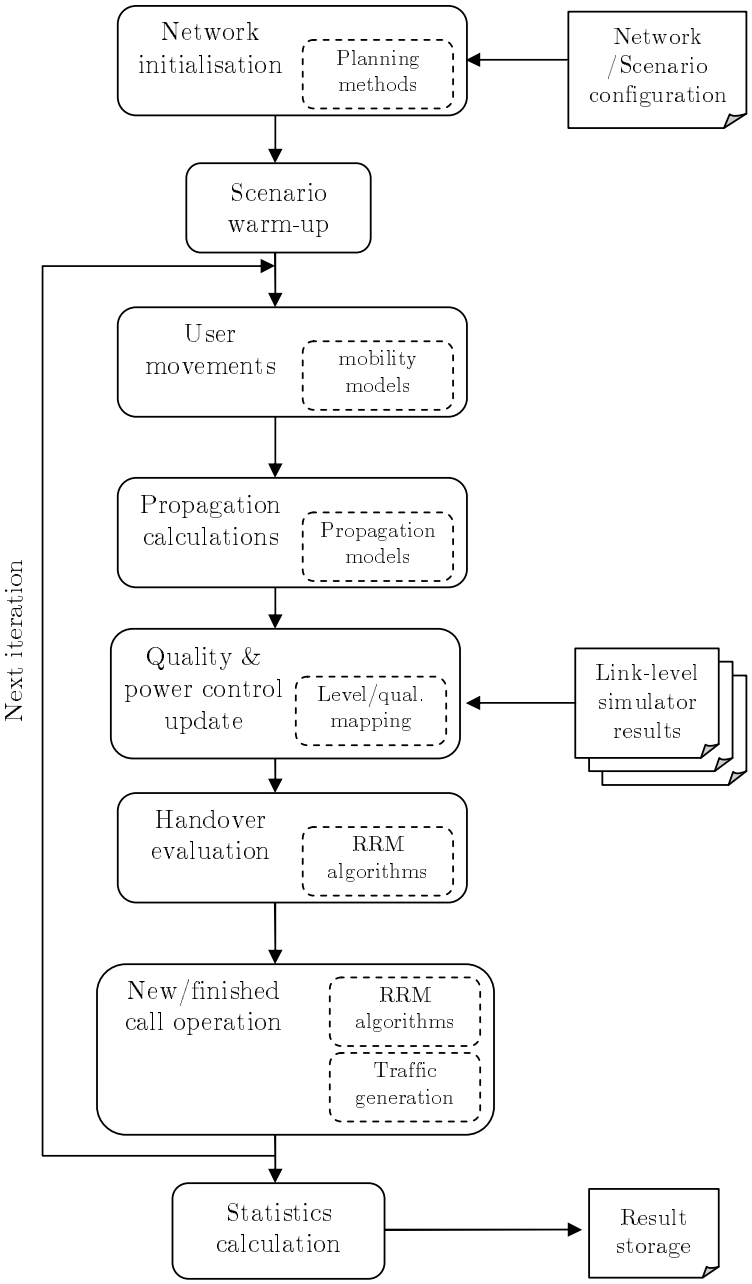


Figure 1.4: A typical system level simulator structure.

keeping an acceptable level of Quality of Service (QoS).

Figure 1.5 shows different network levels having a significant impact into the mobile network performance, and, therefore, they are the main target of optimisation. Its pyramidal structure aims to reflect the same order that operator follows when a live network is being optimised. Over a well-designed radio platform (the lowest level in the figure), the optimisation process starts by ensuring that the network is fault-free. Fault detection area ranges from analysis of alarms to the identification of a bad configuration of network elements. Next levels proceed to the adjustment of physical Base Station (BTS) parameters (e.g. antenna down-tilt or the maximum transmitted power) and the improvement of the adjacency and frequency plans in the network. The assignment of cells to Packet Control Units (PCUs) in a Base Station Controller (BSC) can be optimised in the next level.

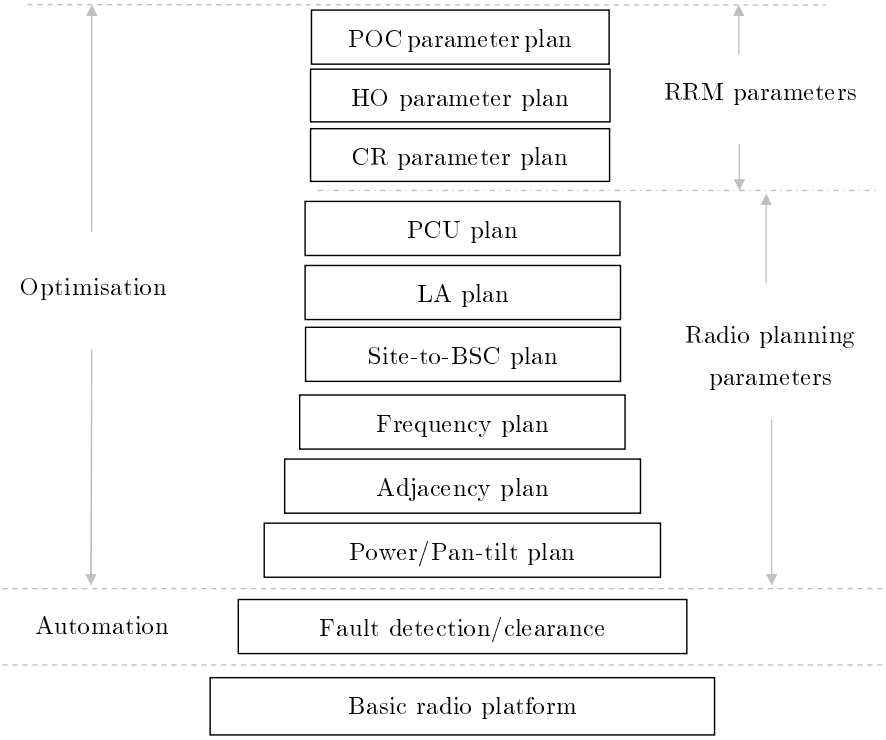


Figure 1.5: Optimisation levels in celullar networks, [2].

Finally, the parameters of RRM algorithms in the BSC, such as Cell ReSelection (CRS), HandOver (HO) and POwer Control (POC), can be tuned to obtain optimal performance.

Depending on the optimisation level, terms *replanning* or *reconfiguration* are used. Replanning refers to those strategies changing network characteristics which are considered stable with time (e.g., network hierarchy, radio resources assigned to cells), and, consequently, those strategies are applied not very frequently. By reconfiguring, it is meant those techniques changing certain network parameters especially in RRM algorithms. Although any optimisation approach can improve network performance for any level in Figure 1.5, operators usually prefer those strategies implying no changes in the existing infrastructure, which is the case of reconfiguration techniques. Due to this non-intrusive characteristic, reconfiguration approaches are executed more often than replanning strategies.

Network models are a basic tool for network optimisation. As observed in Figure 1.3, network models can be constructed to assess the impact of changes in a mobile network, before the real change in the live network. Operators validate optimisation techniques first in network models to minimise unpredictable effects and define as clearly as possible the new network configuration or structure.

Operators are also pushing to automate mobile network optimisation techniques as much as possible. The automation of different processes leads to the design of algorithms changing network configuration in an autonomous way (referred to as *self-organised* or *self-adjusted* networks). In new multi-technology network scenarios, not only with mature technologies, a lot of new scenarios for the application of automatic parameter configu-

ration (or *auto-tuning*) schemes are multiplied. Where several radio technologies coexist in the same geographical area, joint entities must manage traffic flows, indicators and measurements of a very different nature, and the different services provided demand very different features from the mobile network. This results in very changing traffic conditions with time which can only be dealt with by auto-tuning schemes. Automatic adjustment of network parameters is usually implemented by imitating the wide knowledge and experience from the operator technical staff. Thus, the design of auto-tuning schemes is based on previous experience, without any proof of optimality (in a mathematical sense) of these techniques. The use of network models, especially those based on teletraffic, would allow the definition of optimal strategies for network reconfiguration. Thus, an optimal network performance could be obtained.

1.3 Research Objectives

The main goal of this thesis is to design optimisation techniques for mobile networks, by auto-tuning network parameters, based on network models constructed by teletraffic theory and simulators. Such a goal is applied to different problems:

a) A method to redimension signalling resources in cells for GERAN, and constructing a teletraffic model for dedicated signalling channels in the network.

b) A proposal for radio resource management parameter modifications in GERAN to solve localised congestion problems in the network by deriving an optimal criterion for load sharing between cells.

c) A scheme for load balancing in an multi-technology scenario, under strongly unbalanced traffic conditions. The scheme must take into consideration heuristic approaches previously acquired by the operator, and must be adapted to traffic changes with time.

In a) and b), teletraffic theory is used to construct the model, while c) is based on a system-level simulator. Additionally, b) and c) design some criterion or scheme to modify some network parameters, so these two problems can be considered as auto-tuning proposals, while a) can be considered as a replanning strategy.

1.4 Document Structure

This thesis is divided into those problems to be solved, and they will be treated independently. The structure of this document reflects that division in separated chapters, although, for an easier understanding, the different problems have been treated with an unified structure.

This document consists of five main chapters. This introductory chapter gives a general view of network modelling and optimisation. It also introduces general concepts and terms to be used along the rest of the document. Chapters 2, 3 and 4 deal with the research goals defined in section 1.3. These three chapters have a similar structure, beginning with a brief introduction of the problem to be solved, then, describing the proposed scheme and/or model as a solution to the problem in an additional section, and, finally, presenting the main results of the analysis.

Three appendices are also included in this thesis. Appendix A details how Gaver's method is used in the resolution of linear equation systems for the problem described in Chapter 2. Appendix B develops the mathematical procedure to get the optimal traffic sharing conditions presented in Chapter 3, and Appendix C gives a brief summary of this thesis in Spanish.

Chapter 2

Traffic Modelling of Dedicated Signalling Channels in GERAN

This chapter deals with signalling traffic and network modelling in GSM/EDGE Radio Access Network (GERAN) to solve congestion problems in signalling channels. A bad dimensioning strategy in signalling channels leads to congestion problems in mobile networks, even if enough radio resources are devoted to signalling channels. After describing the problem, different network models are proposed to include retrial and time correlation characteristics in signalling data. Models are adjusted with live network performance indicators following a classical optimisation approach. Performance assessment is based on live data. A preliminary analysis shows the need of a re-planning strategy. A comprehensive performance analysis is finally included to show how new model proposals fit much better with network statistics.

2.1 Introduction

Global System for Mobile (GSM) was the first digital (also called *second generation*, 2G) mobile network with an spectacular development since its appearance in 1992. GSM has experienced a worldwide success, with around 3,500 millions subscribers, mainly caused by its pan-European conception, [3]. This success caused the appearance of *GSM/EDGE Radio Access Network* (GERAN), usually classified as 2.5G (second and a half generation) technology. Later, mobile technologies such as Universal Mobile Telecommunication System (UMTS) or Long Term Evolution (LTE) can be considered as the logical evolution in mobile networks. However, operators usually employ 3G networks as the packet switched data bearer, while GERAN is employed for voice traffic, mainly due to its global coverage. 4G networks are still in an early deployment stage.

GERAN systems are in a mature phase. As a consequence, operators and equipment suppliers have a very wide knowledge about network design, development, usual problems, etc. Thus, problems in GERAN are not presently coming from troubles in the roll-out procedure, but arise due to differences between the current situation compared to the conditions assumed when planning the network. Network replanning comprises all those techniques aiming to solve the mismatching between original design and current traffic conditions.

In GERAN, initial network planning allocates frequencies and transceivers along the coverage area based on subscriber estimations and traffic models. Traffic models estimate the behaviour of traffic sources. The better the estimations and models are, the better network performance estimates are obtained. This is equally valid either for signalling traffic channels or voice data channels. In GERAN, signalling capacity largely depends on the capacity of the *Stand alone Dedicated Control CHannel* (SDCCH). SDCCH transmits all signalling information required for mobility management procedures, namely call set-up, mobile station registration and location update, as well as in data services, such as *Short Message Service* (SMS), *Multimedia Messaging Service* (MMS) and *Wireless Application Protocol* (WAP), [4].

Congestion problems are mostly referred to Traffic CHannel (TCH). If a new user demands an empty TCH when all TCHs are busy, the call is blocked. But blocking can also ocurr during the call establishment stage. Call establishment is carried out through signalling channels. A call establishment is also blocked (and, then, the call is not carried) if enough SDCCH resources are not available. Hence, SDCCH congestion must be avoided to minimise revenue loss.

During network design, operators have to estimate the required number of SDCCHs on a cell basis. Traditionally, the Erlang B formula has been used to estimate the minimum number of these channels based on predictions of the signalling traffic, [5], expressed as:

$$E(A,c) = \frac{\dfrac{A^c}{c!}}{\displaystyle\sum_{j=1}^{c} \frac{A^j}{j!}}\,, \tag{2.1}$$

where A and c are the offered traffic and the number of channels, respectively. The application of this formula assumes that:

a) The request arrival process is a Poisson process. A Poisson process mainly includes several assumptions: a) there is no correlation between users and, b) time between consecutive calls is exponentially distributed.

b) A new call demands only one resource block. In other words, the maximum number of users in the system is c (i.e., the number of channels).

c) Blocked attempts are cleared, i.e. the user only tries once (no retrial mechanism is considered).

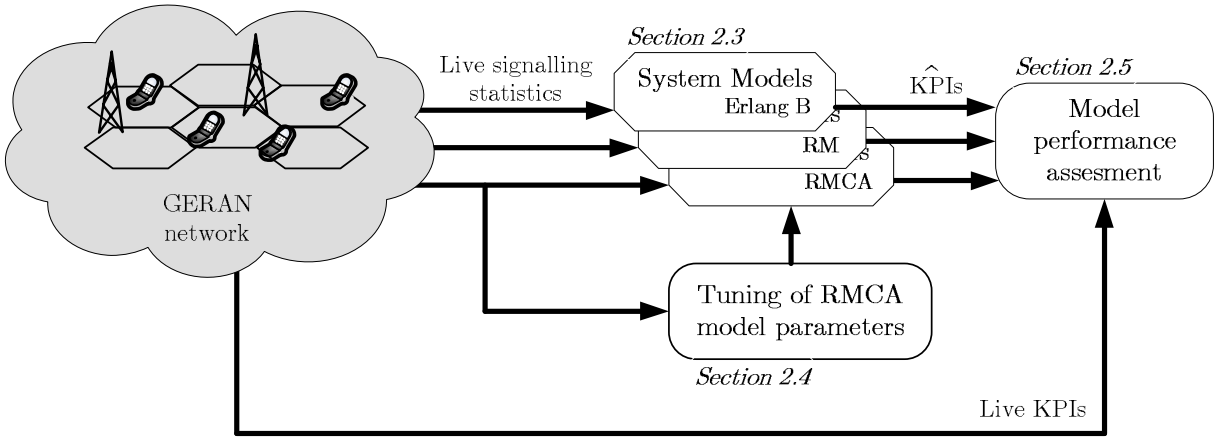


Figure 2.1: General working scheme for Chapter 2.

d) The number of users is large, i.e. offered traffic, A, remains the same no matter how many users are accepted in the system.

Although these assumptions are known to be valid for voice traffic, [6], many of them necessarily do not hold true for the SDCCH traffic. On the one hand, automatic retrial/redial mechanisms incorporated in mobiles cause repeated attempts during congestion periods in SDCCH, [7]. On the other hand, SDCCH requests are correlated in some cells, as will be shown later. In addition, some cells show large signalling traffic from a few terminals. In all these cases, the Erlang B formula fails to give accurate predictions. Due to these limitations, operators used to over-dimension SDCCH resources in the early days of GERAN, reducing blocking problems in signalling channels at the expense of underused resources. However, such an approach is not financially viable anymore as cellular operators have to maximise the usage of each and every time slot in the network to maximise their return on investment.

In this chapter, a comprehensive analysis of the SDCCH is performed using measurements from a live GERAN system. Figure 2.1 shows the structure of this chapter graphically. A first analysis shows that the Erlang B model fails to predict SDCCH congestion and blocking in many cells. Having identified retrials as a source of inaccuracy, a simplified queueing model is presented to evaluate the influence of retrials on SDCCH performance, referred to as *Retrial Model* (RM). Such a model extends that in [8] by considering a mixture of services with and without retrials. Then, the model is improved by including correlated arrivals by a Markov-Modulated Poisson Process, [9]. The resulting model, referred to as *Retrial Model with Correlated Arrivals* (RMCA) includes parameters that can be tuned on a cell basis using statistics in the Network Management System (NMS). Model assessment is carried out by comparing Key Performance Indicators (KPIs) obtained by the model (indicated by the symbol '$\hat{}$') against measurements taken from a live network. Results show that, once the proposed model is tuned on a cell basis, it clearly outperforms models currently used by operators to re-plan SDCCH resources.

The structure of the chapter is organised as follows. Section 2.2 outlines the *Standalone Dedicated Control CHannel* (SDCCH) re-planning problem from the operator's

point of view and introduces traffic and channel modelling, as well as the retrial problem in teletraffic issues. Section 2.3 presents two retrial queueing models for the SDCCH. Section 2.4 describes configuration techniques for the models previously presented in Section 2.3. Section 2.5 compares performance estimates obtained by the models with real network measurements. Finally, Section 2.6 presents the conclusions of this chapter.

2.2 Problem Formulation

In this section, the SDCCH congestion problem is presented. The reasons for congestion in SDCCH channels are first introduced. Then, the state of research in the topic is detailed. The issues presented here will justify the need for the models and tools presented in the next sections.

2.2.1 The SDCCH Dimensioning Problem

Mobile users are not uniformly distributed in space. Moreover, space distribution is not the same with time, showing changes (i.e., global user movements) as time goes by. For example, users concentrate in working or residential areas during the morning or evening periods, respectively. As a consequence, cellular traffic tends to be unevenly distributed both in time, [10], and space, [11].

Temporal traffic fluctuations are a combination of short and long-term trends. Long-term fluctuations comprises population growth, premises openings or seasonal changes. Short-term changes take place in a shorter time scale, e.g., weekly, daily and hourly fluctuations. Fast fluctuations in traffic demand are dealt with by complex Radio Resource Management (RRM) features. As an example, modifying HandOver (HO) margins can adapt cell service areas for a better match between traffic demand to cell radio resources, [12]. In contrast, permanent congestion problems can only be solved by proper dimensioning of traffic resources on a cell basis. A new premise opening with its associated increase of traffic demand is best solved by adding new radio resources to the existing cell permanently. Since new users generate both data and signalling traffic, temporal and spatial traffic fluctuations are reflected in both data and signalling channels. Thus, a higher spatial user concentration in a cell causes not only a high offered data traffic, but also a high signalling traffic, [13].

Figure 2.2 shows the SDCCH busy hour carried traffic distribution on a cell basis in a live network. It is observed that traffic distribution is far away of being uniform. Most cells carry a low amount of signalling traffic, and, at the same time, a few cells carry a very high amount of traffic. This is a clear indication of the non-uniformity of SDCCH traffic. Moreover, Figure 2.3 illustrates the temporal traffic fluctuation along a day for both data and signalling traffic. As seen in the figure, both traffic flows are highly correlated and experience large fluctuations during the day.

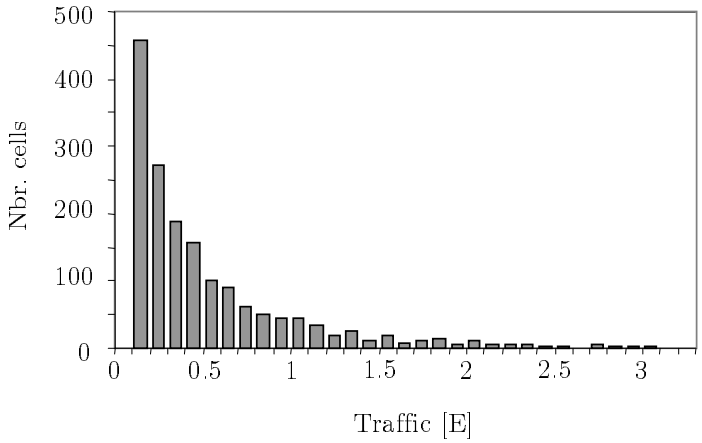


Figure 2.2: SDCCH carried traffic distribution in a cell level.

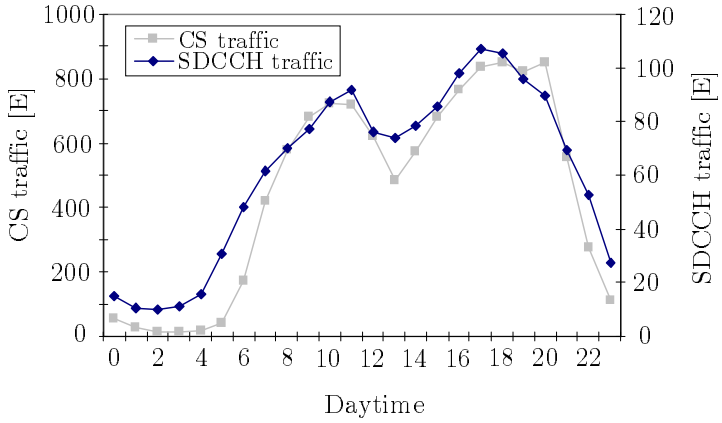


Figure 2.3: SDCCH traffic distribution on an hourly basis.

A major problem in dimensioning signalling resources is the fact that the SDCCH is used for several purposes. SDCCH is used in GSM to provide a reliable connection for signalling messages from and towards the user. Some of the main signalling procedures using this channel are mobile call delivery, either *Mobile Originated Call* (MOC), or *Mobile Terminated Call* (MTC), and *Location Updates* (LU). SDCCH also supports the *Short Message Service* (SMS) and other like *Emergency Call* (EC), *call Re-Establishment* (RE), *IMSI Detach* (ID), *Supplementary Service* (SS), and *GHost seizure* (GH), [14]. The latter reflects SDCCH seizures that time out due to false requests in the Random Access CHannel (RACH). The planning of SDCCH resources aims to minimise persistent congestion problems by a proper selection of SDCCH capacity on a cell basis. The main design parameter is the number of time slots dedicated to SDCCH on a permanent basis. Each time slot can comprise 4 or 8 sub-channels, [15]. Therefore, the number of SDCCH sub-channels in a cell, N, is a multiple of 4. In some networks, one of these sub-channels is used for the Cell Broadcast CHannel (CBCH), in which case N takes values in the set $4i - 1$, $i \in \mathbb{N}_+$.

After the network design stage, SDCCH resources are assigned on a cell basis accordingly to traffic estimations. Due to long-term temporal fluctuations, or an inaccurate design criterion, the initial SDCCH plan does not perform well, and a high blocking is experienced. Re-planning strategies try to solve design failures. For re-planning purposes,

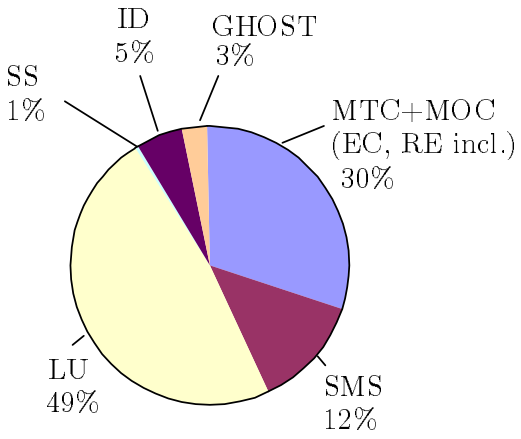


Figure 2.4: SDCCH traffic divided by establishment causes.

SDCCH statistics are collected by the NMS on an hourly basis. KPIs are the SDCCH blocking ratio (i.e., ratio of blocked attempts), BR, and the SDCCH congestion ratio (i.e., ratio of time without free sub-channels), CR. Vendor equipment also provides the average SDCCH carried traffic, A_c, the mean SDCCH holding time, MHT, and the number of offered, blocked and successfully carried SDCCH attempts per hour. It should be pointed out that the latter counters include both fresh and retrial attempts from the same user, as the network cannot differentiate between them. In addition, the number of carried attempts is also broken down by establishment causes (i.e., MOC, MTC, EC,...).

Operators usually employ Erlang B formula, (2.1), to assign radio resources to cells, assuming some statements about traffic and user behaviour (described in Section 2.1). Some of those statements do not hold true for SDCCH traffic. Two main sources of unaccuracy can be found.

First, operator's approach does not usually consider UE retrials. Retrial mechanism can be rejected when congestion ratios are very low, i.e., the user is accepted in its first attempt, so no second or additional attempts occur. A low congestion ratio cannot be assumed in SDCCH, as it will be seen later. Additionally, retrial/redial mechanisms are very easy to implement in current terminals, being automatically made by the UE or just pushing one key. One of the main effects of retrials over network performance indicators is that blocking ratio is different and higher than congestion ratio (unlikely to Erlang B traffic where congestion and blocking ratios are the same), due to additional attempts coming from the same user. Thus, Erlang B dimensioning approach underestimates blocking ratio if offered traffic experiences retrials.

Second, user mobility patterns cause some special phenomena in location management traffic, carried through SDCCH. Figure 2.4 shows the share of SDCCH traffic components in a live network divided by the establishment causes. Figure 2.4 proves that LU requests are a high percentage of the network-wide amount of SDCCH messages. An LU message is originated when the User Equipment (UE) crosses a Location Area (LA) border. Thus, it is expected that signalling traffic will be very related to user mobility trends in the network.

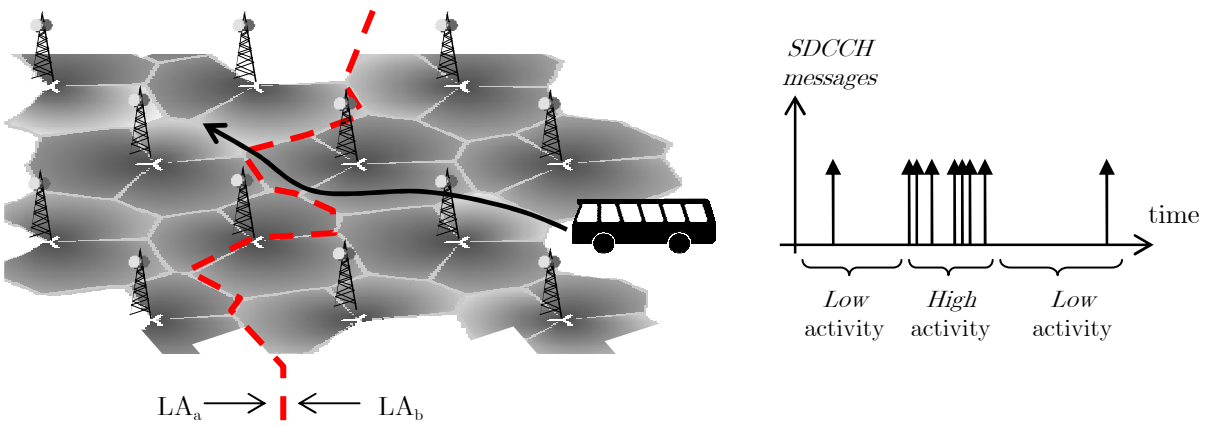


Figure 2.5: Example of SDCCH messages correlation.

Figure 2.5 illustrates an example of time correlation of traffic associated to user movement patterns. LU messages are triggered when an UE changes its LA. The UE communicates the change to the Base Station Controller (BSC) through the SDCCH. If a significant amount of users crosses an LA border at the same time (e.g. public buses, trains or traffic lights), LU messages will be concentrated in a short period of time. Such a time correlation of LU attempts defines an interval with a high traffic rate, referred to as *high* activity period. Likewise, a *low* activity period is defined. Consequently, the conditions enumerated for a SDCCH dimensioning design following (2.1) could not be assumed.

The above mentioned reasons cause a bad SDCCH resource distribution. In the absence of good teletraffic models, operators can re-assign SDCCH resources, based on SDCCH messages measurements. Such a re-planning task is done, at most, on a weekly basis. Due to bad planning, some cells experience unacceptable SDCCH blocking during operation. Network operators consider blocking ratios larger than 10^{-2} unacceptable. Note that if a MOC or MTC attempt is blocked on the SDCCH, the call is lost. Even if some schemes allow using spare traffic channels temporarily for signalling purposes, this cannot be relied on as peaks of signalling and call traffic tend to be correlated, as seen in Figure 2.3, [13]. Hence, operators counteract SDCCH blocking by increasing the number of sub-channels, N, in problematic cells. Subsequent addition of new cells often causes that SDCCH resources on existing cells become unnecessary, which cannot be detected without a precise performance model. Unfortunately, such a model is not currently available due to retrials and correlated arrivals in the SDCCH. As a result, SDCCH resources are over-dimensioned in many cells and under-dimensioned in others, [13]. This problem can be solved by an improved performance model that can be tuned on a cell basis. As main benefit, many time slots unnecessarily assigned to SDCCH could be converted into Traffic CHannels (TCHs), increasing network capacity.

2.2.2 State of Research

Many teletraffic models have been proposed for cellular networks as a tool for an easier design or replanning. Teletraffic theory applies probability theory to telecommunication

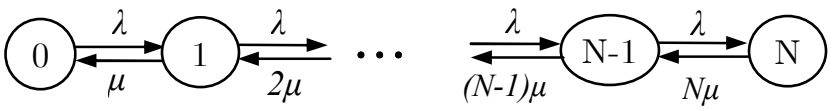


Figure 2.6: Markov chain for a simple network model with successive state transitions.

systems. It provides analytical equations for the planification, performance analysis or maintenance of telecommunication systems. Teletraffic engineering was first employed in fixed telephone network design. Erlang B equation, (2.1), is the milestone of teletraffic engineering. With that equation, several network parameters can be designed (e.g., number of channels) and network performance can be estimated (e.g., blocking ratio). Teletraffic models are a central issue in most of research areas where a flow of requests and sparse resources must be managed. This section will focus on the state of research in mobile network and retrials in traffic models, which are the main tools and concepts to apply in this chapter.

The basic teletraffic model for mobile networks was presented by [16]. The model relates network performance with data traffic, including both fresh and handed over calls. Several priority schemes are considered for handover incoming connections. The main assumptions taken in [16] are the ones enumerated in section 2.1. Thus, signalling traffic specific characteristics are not taken into account. Subsequent studies have tried to spread the model by considering (or eliminating) different assumptions in the original proposal. A first generalisation is presented in [17]. While [16] assumed a negative exponential distribution for the channel holding time (i.e. the user resides in a cell by an exponential distribution), Fang and Chlamtac propose a general distribution for the cell residence time in mobile networks. They also introduce the importance of a correct user mobility model.

Previous network models only considered one service, usually real time voice calls. Thus, all incoming calls maintained statistic characteristics of radio resource consumption and holding time parameters. Several references extend classical models to consider multiple services, with real time characteristics, [18][19][20]. Each different service needs a different amount of radio resources, complicating the model analysis. From different traffic sources a state transition matrix is defined, and, later, performance indicators are obtained from that matrix. When single service networks were modelled, a Markov chain contains all possible transitions between states (actually, from a i-state, only the $i-1$ and $i+1$ states are possible, as shown in Figure 2.6). When multiple services arise, multiple transition between states are now possible and a state transition matrix is usually defined. Following mobile network evolution, hierarchical (also called *multi-layered*) cellular networks have been also modelled in several references, [21][22][23][24]. Network layers appeared in GSM as the main tool for deployment of new cells when traffic increase demands additional resources, not available in already existing cells. As an additional evolution, heterogenous scenarios include several radio access technologies, with different radio resources, traffic models and management policies, [25][26][27].

All the previous models have been conceived for user traffic channels. As mentioned

above, it has been assumed in the literature that those models are still valid for signalling channels. Therefore, the same methodology has been translated from traffic in signalling channels. However, to the author's knowledge, no study has been published checking the validity of these models for dedicated signalling channels based on real network data.

As the main difference with user traffic channels, signalling traffic experience automated retrials. A large number of papers have studied the problem of retrials in both wired and wireless networks. For a very deep and exhaustive survey on retrial queues, the reader is referred to [28] or [29]. Earlier references on the effects of retrials are [30] and [31]. Performance analysis of standard multi-server retrial systems, considering Poisson arrivals, exponential service times and exponential inter-retrial times, is presented in [32] or [33]. These early references concern with the analytic solution, if possible, of multiserver retrial queues including the retrial phenomenon. That analytical solution, however, has not been obtained for more than a few servers, [29], which is not the case of a typical SDCCH scenario (in this work, 79% of cells have 7 or more servers, carrying 83% traffic). Therefore, the aim of current studies is to develop efficient numerical methods to estimate performance expressions. With this goal, one of the typical assumptions is the homogenisation of the state space beyond a given number of users in the retrial orbit (i.e., users retrying). This implies that performance measures do not differs much once users retrying are more than M. Due to their use in this thesis, truncated methods, [34][35], are here emphasized as one of the approaches for such an assumption in the retrial orbit. In [36], performance analysis is extended to retrial systems with correlated arrivals.

In the context of cellular networks, additional characteristics must be included to retrial models. Thus, previous models have been extended, still with user traffic channels, to consider handovers, [8][37][38], automatic equipment retrials and user's redials, [7][39], and more general distributions of inter-arrival, service and inter-retrial times, [9][38].

The problem treated here has similarities with that reported in [7]. In that paper, a simple analytical model was proposed to estimate, for each cell, the average number of retrials and redials per fresh call attempt in user traffic channels by using only NMS measurements. The main differences for the SDCCH are: a) the mixture of services with very different properties, and b) the presence of correlated arrivals. In this chapter, all these well-known principles and techniques of retrial queues in literature are applied for the first time to the analysis of the SDCCH. The main contributions are: a) to show the limitations of the Erlang loss model for dedicated signalling traffic in GERAN, b) to prove that such limitations are due to time correlation between arrivals, c) to propose an accurate retrial queueing model for the SDCCH, which, unlike more refined models, can easily be tuned from network statistics, and d) to compare SDCCH performance estimates obtained by different queueing models against real network measurements.

To the author's knowledge, there is no data or analysis over signalling traffic in live GERAN showing the validity of dimensioning strategies used by the operators for dedicated signalling channels. In this thesis, models for dedicated signalling traffic are formulated, including retrial and correlation characteristics. The models are validated

with live GERAN data. This work uses retrial formulation presented in [8].

2.3 System Models

This section describes two queueing models for the SDCCH. Models in this section consider the peculiarities of signalling traffic, in contrast to existing proposal for payload traffic flows. A first model considers retrials. A second model extends retrials by adding time correlated arrivals.

2.3.1 Retrial Model

A new queueing model for signalling channels considering retrial phenomenon is presented here. A non-retrial model establishes that a user attempt can either be served or blocked. A third possibility arises when retrial mechanisms are introduced: a user that has been initially blocked waits for a new attempt, which will be triggered by the same user in a short period of time. Then, a third state is possible, where the user is waiting for a new attempt after a failed channel access.

The basic retrial model proposed here is based on that presented in [8]. Such a model considers a single cell in which repeated attempts occur. As shown in Figure 2.7, different states can be experienced by the user: *idle*, *active* and *wait-for-reattempt*. After finishing a transaction, a terminal goes back to idle state until its next fresh attempt is generated. In case of rejection when a channel is requested, the terminal enters the wait-for-reattempt state (also referred to as *retrial orbit*) with retrial probability θ or abandons with probability (1-θ). The durations of all states are assumed to be exponentially distributed, and, hence, the system can be modelled by a Markov chain. For simplicity, it is assumed here that the population in a cell is infinite, i.e., offered traffic keeps the same regardless of the number of users being carried in the model.

To make this model more general, different traffic flows are defined, differentiating between services with and without retrials in the SDCCH. The resulting model, referred to as *Retrial Model* (RM), is shown in Figure 2.8(a). As mentioned before, the arrival flow is divided into two components, namely retrial and non-retrial traffic, depending on whether blocked attempts are repeated or not. All services arriving at the system (LU, MOC, MTC,...) are summarised in those two categories. For simplicity, it has been assumed that all services except GHost seizures (GH) are repeated until success (i.e., θ=1). This assumption is reasonable, because UEs usually implement automatic retrials. Even if automatic retrials by the terminal fail, re-dialing is only a matter of pushing a button by the user in current handsets. Ghost seizures are fictitious attempts since they occurred due to channel fadings and boosts, causing a false request in RACH, and, consequently, GH is characterised as a non-retrial traffic flow. Figure 2.8(b) shows the state transition diagram of RM, where the state of the system (i, j) is described by

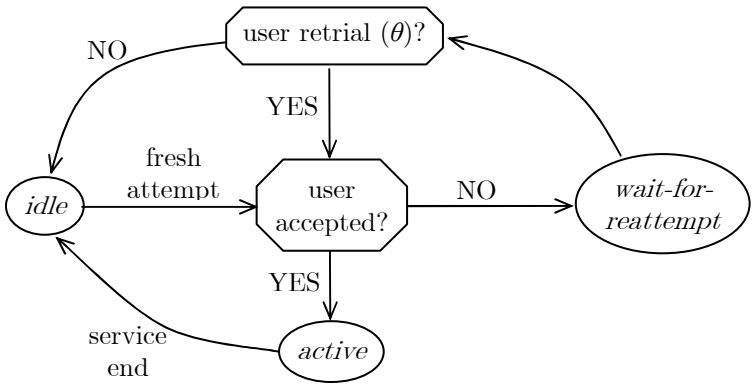


Figure 2.7: User states in a retrial model.

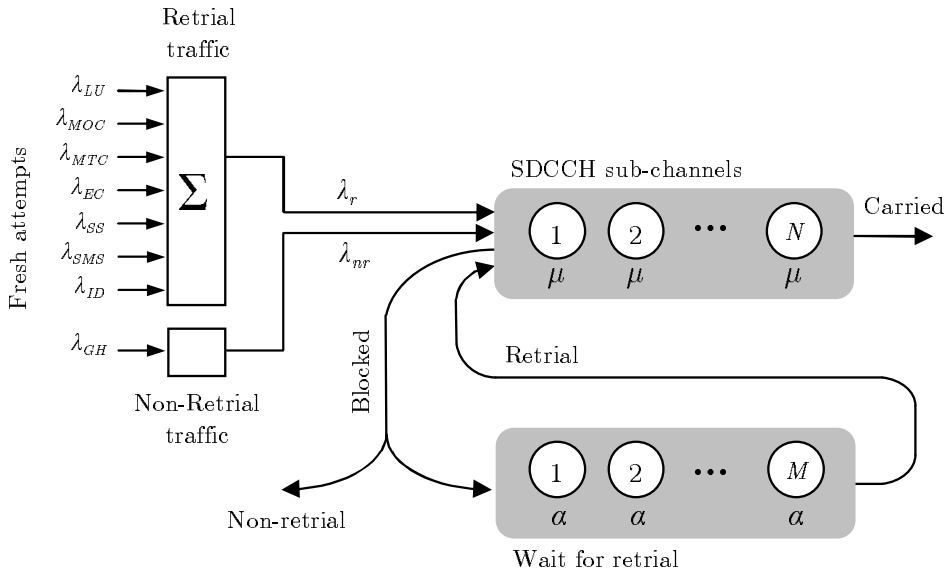
the number of busy SDCCH sub-channels, i, and the number of requests waiting for re-attempt (i.e., the number of users in the orbit), j.

The main parameters in the model are the total arrival rate for services with and without retrials, λ_r and λ_{nr}, the service rate (i.e., the inverse of the mean channel holding time), μ, the retrial rate (i.e., the inverse of the mean time between retrials), α, the number of sub-channels, N, and the size of the orbit, M. The values of all parameters in the model can be obtained from measurements gathered on a cell and hourly basis in the NMS. The total arrival rate for services with retrials, λ_r, is obtained directly from the number of seizures per hour (note that, for these services, offered and carried traffic coincides, since it has been assumed that θ=1). For services without retrials (i.e., GH), the arrival rate, λ_{nr}, is estimated from the congestion ratio, CR, as

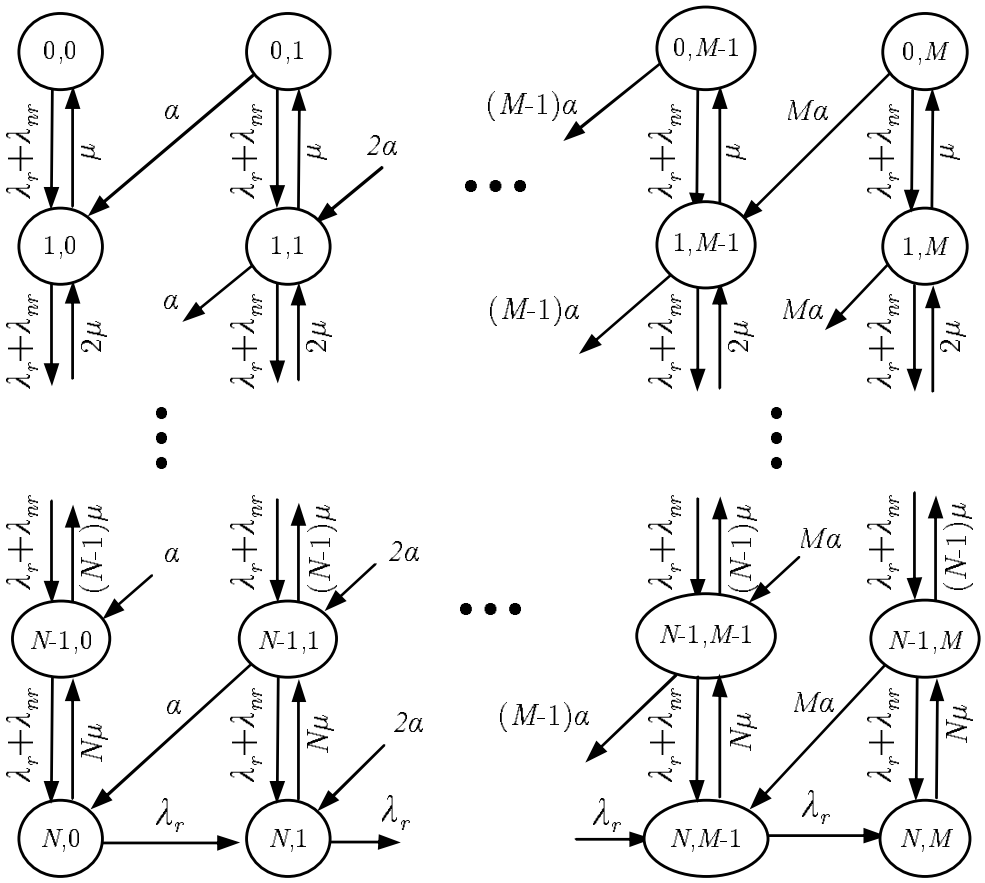
$$\lambda_{nr} = \frac{N_{GH}}{3600(1-CR)} \quad , \tag{2.2}$$

where N_{GH} is the number of ghost seizures in one hour. The service rate, μ, is the inverse of the SDCCH mean holding time. The retrial rate, α, is fixed to the value configured network wide by the operator. In the previous model, it is assumed that the time between consecutive retrials is exponentially distributed to ensure mathematical tractability, even if this parameter takes a deterministic value in a real system. It is expected that this assumption has a negligible impact on key performance indicators, as shown in [39].

The size of the orbit, M, is an important parameter in the model. A strictly exact retrial model should include an infinite orbit, so no retrial request is eliminated from the system. An infinite orbit would lead to an infinite number of states in Figure 2.8(a). However, the size of the orbit must be finite for computational efficiency and mathematical tractability. Additionally, the number of states for a RM system, N_s, is linearly dependent on the M value, $N_s = (N+1)(M+1)$, as seen in Figure 2.8(b). Thus, high values of M lead to an excessively large number of N_s. For computational efficiency, the number of users in the orbit is artificially limited to M to keep the number of system states finite and not very high, [40]. When this limit is introduced, those system states with users $j > M$ can not be considered in the performance analysis. Different authors describe some



(a) System model



(b) State transition diagram

Figure 2.8: The basic retrial model.

techniques to analyse retrial models with a finite orbit without errors or maintaining the error below certain limits, [34][35]. Basically, the size of the orbit, M, must be chosen so that the probability that the orbit is full is negligible, which depends on offered traffic conditions.

The reader is referred to [29] for a performance analysis of this classical retrial queue shown in Figure 2.8. It should be pointed out that an analytical expression for steady-state probabilities in RM is only available for $N = 1$ and 2. For $N \geq 3$, the problem does not preserve the birth-and-death structure and, consequently, no closed-form expression can be found, [29]. Hence, teletraffic performance indicators can only be obtained by computing the stationary distribution of the Markov chain describing system dynamics numerically. Thus, all steady-state probabilities (i.e., the probabilities of being in the state (i, j)) are obtained, and teletraffic performance indicators can then be calculated. Such a computation process starts with the solution of a system of linear equations, expressed as

$$\overline{\Pi}\,\mathbf{Q} = 0\,,\ \ \overline{\Pi}\,\overline{e} = 1\,,\ \ \overline{\Pi} \geq 0\,, \tag{2.3}$$

where $\overline{\Pi}$ is the steady-state probability vector, which contains all the probabilities of being in (i, j) state, $\mathbf{Q}$ is the infinitesimal generator matrix[1] and e is a column vector of ones, [41]. The reader is referred to [34] for a detailed description of the values of $\mathbf{Q}$ for the retrial queue in Figure 2.8. For the proposed RM, $\mathbf{Q}$ has a block tri-diagonal structure if states are enumerated column-wise, suggesting the use of block gaussian elimination for solving (2.3). This is the approach followed by Gaver, Jacobs and Latouche, [42]. Appendix A details the formulation and efficient resolution of (2.3) for RM with Gaver's method.

Once the stationary distribution, $\overline{\Pi}$, is obtained, teletraffic performance indicators can be calculated. The main indicators are the carried traffic, $\hat{A}_{c_{rm}}$, in Erlangs, the congestion ratio (i.e., the probability of all resources are occupied), $\hat{CR}_{rm}$, and the blocking ratio, $\hat{BR}_{rm}$ (i.e., the probability of a service attempt to be blocked). Note that a blocked user join the orbit in the case of retrial services, so a specific retrial user could cause several blocked attempts. Performance indicators are expressed as a function of the steady-state probabilities as

$$\hat{A}_{c_{rm}} = \sum_{i=0}^{N} \sum_{j=0}^{M} i\,\Pi(i,j)\,, \tag{2.4}$$

$$\hat{CR}_{rm} = \sum_{j=0}^{M} \Pi(N,j) \tag{2.5}$$

[1] For any off-diagonal element, values in $\mathbf{Q}$ matrix, $q_{a,b}$, are calculated as the transition rate from state $a \in \{1,\ldots,N_s\}$ (row) to state $b \in \{1,\ldots,N_s\}$ (column). Off-diagonal elements are all positive and $\mathbf{Q}$ matrix has a dimension of N_sxN_s. Row sum must equal to zero, so diagonal elements are all non-positive and they are calculated as the complementary sum of their respective rows (i.e, $q_{a,a} = -\sum_{k=1}^{N_s} q_{a,k}$ $\forall k \neq a$)

and

$$\hat{BR}_{rm} = \frac{\sum_{j=0}^{M} \left[(\lambda_r + \lambda_{nr} + j\alpha)\, \Pi(N, j) \right]}{\sum_{i=0}^{N} \sum_{j=0}^{M} \left[(\lambda_r + \lambda_{nr} + j\alpha)\, \Pi(i, j) \right]} \ , \tag{2.6}$$

where $\Pi(i, j)$ is the probability of having i busy sub-channels and j users in the orbit, previously obtained by solving (2.3). Roughly, $\hat{A}_{c_{rm}}$ is the sum of all steady-states probabilities multiplied by the users in the system for each state. $\hat{CR}_{rm}$ accumulates all the probabilities having RM resources fully occupied ($i = N$). Finally, $\hat{BR}_{rm}$ is the quotient of the sum of all blocked attempts to all traffic attempts in the system.

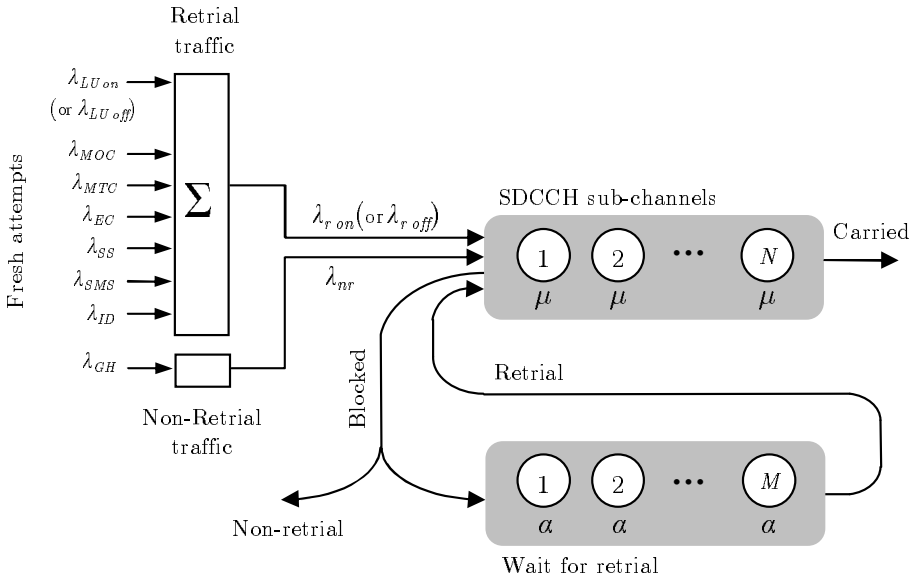
2.3.2 Retrial Model with Correlated Arrivals

In mobile networks, user location is constantly updated by LU requests sent through the SDCCH. A LU request is triggered when a subscriber crosses the border of location areas into which the network is divided. For subscribers moving in groups (e.g., public transport), the boundary-crossing event is synchronised, [43]. As a result, LU requests tend to concentrate in short periods of time in cells on the border of location areas, as shown in Figure 2.5. Signalling traffic models do not usually take this effect into account and, consequently, consecutive LU attempts are not correlated.

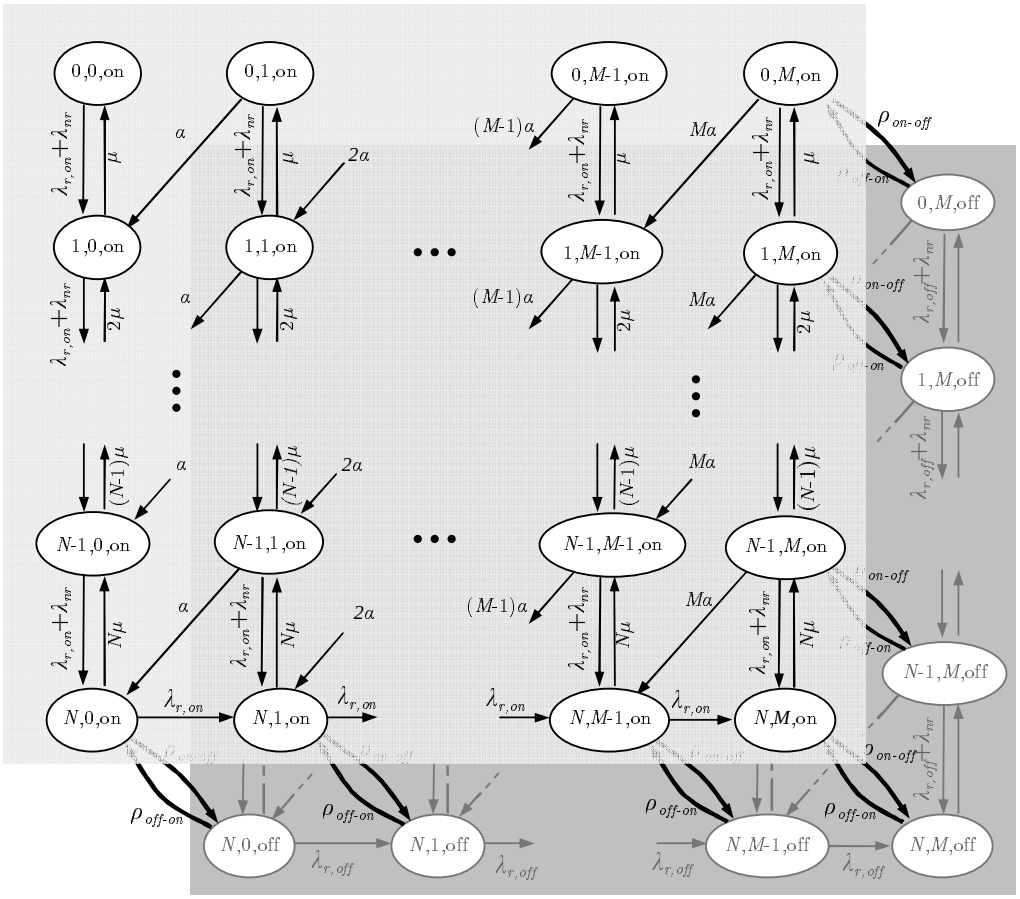
A new model with correlated arrivals is presented here. To account for this effect, the proposed model considers time correlation of LU attempts. For simplicity, time correlation between fresh LU attempts is modelled by a switched Poisson process, [44], where the system changes alternately between two different states. In this case, the LU arrival rate alternately switches between two values (denoted as *on* and *off* values) with a certain frequency, showing two different states in dedicated signalling traffic. Other services (i.e., SMS, MTC, MOC...) do not present correlation, as LU traffic does. Thus, those non-correlated services do not show differences between *on* and *off* states in RMCA (i.e., the arrival rate during the *on* period is the same as in the *off* period). It is assumed that the duration of the *off* and *on* periods is exponentially distributed.

The resulting model, referred to as *Retrial Model with Correlated Arrivals* (RMCA), is shown in Figure 2.9. Figure 2.9(a) details the system model, quite similar to RM except the two possible LU values, and Figure 2.9(b) presents the state-transition diagram. Figure 2.9(b) plots a tri-dimensional state transition diagram, which extends that in Figure 2.8(b) by considering two LU traffic intensity states, denoted as *on* and *off*. The third dimension in the diagram represents the *on* and *off* system states (front and background planes, respectively).

Compared to RM, the new parameters in the model are: a) the LU arrival rate during the *on* and *off* periods, $\lambda_{LU_{on}}$ and $\lambda_{LU_{off}}$, and b) the switching rates between LU activity



(a) System model



(b) State transition diagram

Figure 2.9: The proposed retrial model with correlated arrivals.

states, $\rho_{on\text{-}off}$ and $\rho_{off\text{-}on}$ (or, equivalently, the mean duration of the *on* and *off* states, τ_{on} and τ_{off}). An important RMCA indicator is the measured average LU arrival rate, λ_{LU}, which is the weighted average of the attempt rates during the *on* and *off* periods

$$\lambda_{LU} = \frac{\lambda_{LU_{on}}\tau_{on} + \lambda_{LU_{off}}\tau_{off}}{\tau_{on} + \tau_{off}} \, . \tag{2.7}$$

The state of the system is now described by three values, (i, j, k), where i is the number of busy SDCCH sub-channels, j is the number of requests waiting for re-attempt, and k is a new parameter, the LU-activity state.

Computation of Queueing Performance Indicators

Note that there is no closed-form expression that relates most important performance indicators and model parameters in RMCA. Thus, queueing performance can only be estimated by computing first the stationary distribution numerically from (2.3). For brevity, the reader is referred to [36] for the value of $\mathbf{Q}$ for the retrial queue in Figure 2.9, as a special case of the MAP/M/C retrial queue. As in RM, $\mathbf{Q}$ has again a block tri-diagonal structure if states are enumerated in lexicographic order, so that Gaver's method can be used again. Appendix A details Gaver's methodology and its formulation for the RMCA problem. As it will be explained later in this chapter, the RMCA case has to solve (2.3) many times in order to tune the new parameters introduced by the *on-off* transitions. Consequently, finding an efficient way of solving (2.3) is key for RMCA. Considerations about the complexity of different methods solving RM and RMCA systems are also detailed in Appendix A.

Once (2.3) is solved, the resulting stationary distribution, $\overline{\Pi}$, is used to compute queueing performance indicators, i.e., the carried traffic, $\hat{A}_{c_{rmca}}$, the congestion ratio, $\hat{CR}_{rmca}$, and the blocking ratio, $\hat{BR}_{rmca}$, as

$$\hat{A}_{c_{rmca}} = \sum_{i=0}^{N}\sum_{j=0}^{M}\sum_{k\in\{on,off\}} i\,\Pi(i,j,k) \, , \tag{2.8}$$

$$\hat{CR}_{rmca} = \sum_{j=0}^{M}\sum_{k\in\{on,off\}} \Pi(N,j,k) \tag{2.9}$$

and

$$\hat{BR}_{rmca} = \frac{\sum_{j=0}^{M}\left[(\lambda_{r_{on}} + \lambda_{nr} + j\alpha)\,\Pi(N,j,on) \; + \; (\lambda_{r_{off}} + \lambda_{nr} + j\alpha)\,\Pi(N,j,off)\right]}{\sum_{i=0}^{N}\sum_{j=0}^{M}\left[(\lambda_{r_{on}} + \lambda_{nr} + j\alpha)\,\Pi(i,j,on) \; + \; (\lambda_{r_{off}} + \lambda_{nr} + j\alpha)\,\Pi(i,j,off)\right]} \, , \tag{2.10}$$

where $\Pi(i, j, k)$ is the probability of having i busy sub-channels and j users in the orbit in the LU-activity state k, where $k \in \{on, off\}$. The calculation of these indicators is similar to RM approach, but adding the *on-off* state probabilities in each case.

Parameter Sensitivity Analysis

RMCA has introduced new parameters that add flexibility to the model. RMCA can model additional scenarios not previously considered in RM. In the absence of an analytical expression that relates system performance and model parameters, it is interesting to describe the effect of these new parameters on queueing performance for RMCA. The analysis is focused on parameters that reflect time correlation between new SDCCH requests (i.e., $\lambda_{LU_{on}}$, $\lambda_{LU_{off}}$, τ_{on} and τ_{off}). Note that RMCA reduces to RM when $\lambda_{LU_{on}} = \lambda_{LU_{off}}$. For each combination of parameter values in the sensitivity analysis, a new matrix $\mathbf{Q}$ is generated and a new set of steady-state probabilities are obtained by the Gaver's algorithm. For simplicity, it is assumed in this sensitivity analysis that all SDCCH traffic is due to LUs (i.e., $\lambda_r = \lambda_{LU}$, $\lambda_{nr} = 0$). The retrial rate, α, is set to $1/6\ s^{-1}$, as fixed by cellular operators. Likewise, the service rate, μ, is set to $1/6\ s^{-1}$, according to real measurements of the SDCCH mean holding time.

The first experiment, described in Figure 2.10(a), aims to quantify the impact of concentrating traffic demand in short periods of time. For this purpose, the duration of an *on-off* cycle, T_c $(= \tau_{on} + \tau_{off})$, is fixed to one measurement period in the NMS (i.e., one hour). Then, the number of attempts in each state, $\lambda_{LU_{on}} \cdot \tau_{on}$ and $\lambda_{LU_{off}} \cdot \tau_{off}$, is fixed to be the same and half of the global number of LU attempts measured along both *on* and *off* states, i.e.,

$$\lambda_{LU_{on}} \tau_{on} = \lambda_{LU_{off}} \tau_{off} = \frac{\lambda_{LU}(\tau_{on} + \tau_{off})}{2} . \tag{2.11}$$

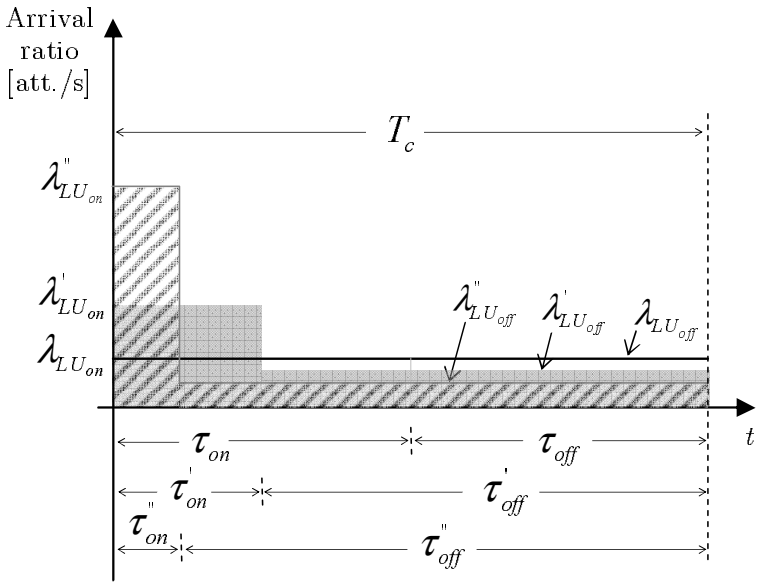
Finally, the length of the *on* period is progressively reduced. As observed in Figure 2.10(a), as τ_{on} is reduced (and, consequently, τ_{off} is increased to maintain T_c), the value of $\lambda_{LU_{on}}$ increases and the value of $\lambda_{LU_{off}}$ decreases to satisfy (2.11). Thus, the degree of time correlation between arrivals is controlled by the ratio $r = \tau_{on}/\tau_{off}$. The lower r, the shorter τ_{on} and longer τ_{off}. From (2.7) and (2.11), it can be derived that

$$\lambda_{LU_{on}} = \frac{\lambda_{LU}(\tau_{on} + \tau_{off})}{2\tau_{on}} = \frac{\lambda_{LU}}{2}(1 + \frac{1}{r}) \tag{2.12}$$

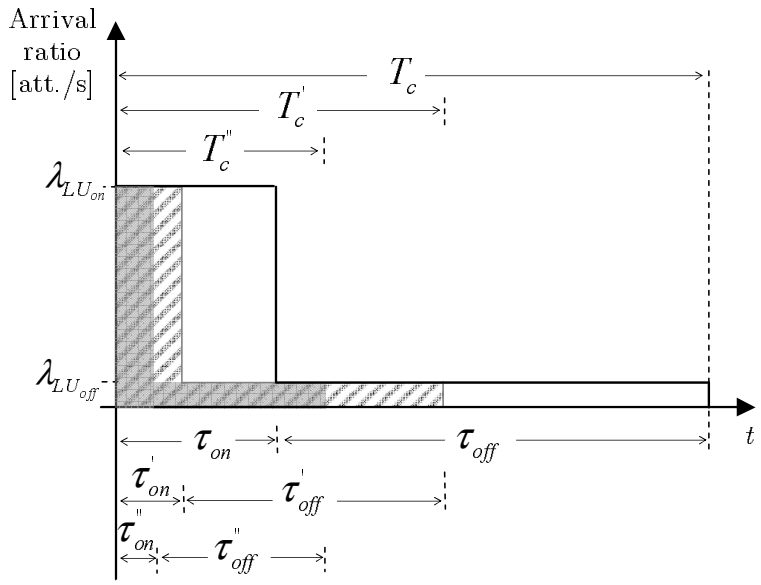
and

$$\lambda_{LU_{off}} = \frac{\lambda_{LU}(\tau_{on} + \tau_{off})}{2\tau_{on}} = \frac{\lambda_{LU}}{2}(1 + r) . \tag{2.13}$$

From (2.12) and (2.13), it can be deduced that, if $r = 1$ (i.e., $\tau_{on} = \tau_{off}$), $\lambda_{LU_{on}} = \lambda_{LU_{off}} =$



(a) Modification of r (T_c fixed)



(b) Modification of T_c (r fixed)

Figure 2.10: Sensitivity analysis for RMCA model.

λ_{LU} (i.e., new arrivals are uncorrelated) and RMCA is reduced to RM. Recall that, unlike the Erlang loss model, RM still considers retrials. In contrast, if $r \to 0$ (i.e., $\tau_{on} \to 0$), $\lambda_{LU_{on}} \to \infty$, so that new arrivals are highly correlated. From (2.12) and (2.13), it can also be deduced that r must not be greater than 1 to ensure that $\lambda_{LU_{on}} \geq \lambda_{LU_{off}}$.

Figure 2.11 shows congestion and blocking ratios (solid and dotted lines, respectively) with increasing traffic demand for different values of r for the particular case $N = 3$. For comparison purposes, the Erlang Loss model is also included in the figure (denoted as Erl-B). In the latter, $\hat{BR}_{ErlB}$=$\hat{CR}_{ErlB}$. As expected, the Erlang Loss Model, considering neither retrials nor correlated arrivals, has the lowest BR and CR. Note that an Erlang Loss Model user only tries once and attempts from different users are not concentrated in time, causing less network congestion and a lower number of blocked attempts.

In RMCA, the lower the value of r (i.e., the higher the concentration of traffic demand during the *on* period), the larger BR. When r is low, many attempts are concentrated in a short period of time, τ_{on} is low, so many attempts are blocked even when the channel is mostly idle for the global T_c (actually, for the *off* period) and CR is low. The same trend is observed for CR for small offered traffic (i.e., CR increases when r decreases for $A_c < 1.5$ in Figure 2.11). However, the opposite trend is observed in CR for large offered traffic. This result can be explained by the fact that time correlation between new arrivals makes congestion possible, even for very small average offered traffic (traffic can be highly concentrated in a very short *on* period). But, for very large average offered traffic, time correlation between arrivals ensure long periods of low activity, which leads to a low CR. More formally, equation (2.12) shows that, if $r \to 0$, $\lambda_{LU_{on}} \to \infty$, and, obviously, $\tau_{off} \to T_c$, since $\tau_{on} \to 0$.

The second experiment, described in Figure 2.10(b), evaluates the influence of the switching rate between the *on* and *off* periods. For this purpose, the ratio $r = \tau_{on}/\tau_{off}$ is fixed, while still satisfying (2.11), and the duration of an *on-off* cycle, T_c, is progressively reduced.

Figure 2.12 shows congestion and blocking ratios with increasing traffic demand for different values of T_c (in seconds) for $r = 0.1$ (i.e. τ_{on} is ten-times shorter than τ_{off}) and $N = 3$. In the figure, it is observed that the larger T_c (i.e., the lower the switching rate, and the larger τ_{on} and τ_{off}), the higher BR. This is due to the fact that, for large switching rates, the effect of the transient regime between the *on* and *off* states becomes more evident. Thus, for low values of T_c, the short *on* period might not be long enough to cause congestion after the queue becomes empty during the long *off* period. It can also be observed that, as T_c decreases, RMCA tends to perform as RM ($r = 1$ in Figure 2.11), despite the fact that $r = 0.1 << 1$ (i.e., $\lambda_{LU_{on}} >> \lambda_{LU_{off}}$). In this case, RMCA tends to behave as if one only average traffic flow, λ_{LU}, is approaching the system.

2.4 Tuning of RMCA model parameters

RM can be configured directly from live network data, namely attempts counters per cause, mean holding time and number of channels per cell. This is not the case of RMCA, where not all parameters can be obtained directly from measurements in the NMS. Unlike RM, time correlation parameters in RMCA are not known by the network. Note that the values of $\lambda_{LU_{on}}$, $\lambda_{LU_{off}}$, τ_{on} and τ_{off} are not available, since only the average LU arrival rate, λ_{LU}, is measured. Neither is possible to define global default values for these parameters as experience shows that they greatly vary from cell to cell (e.g., cells located at the border versus those in the middle of a LA). Hence, time correlation parameters must be estimated from SDCCH performance measurements on a cell basis. Such a problem is referred to as *inverse problem* in queueing theory.

This section manages the problem of adjusting RMCA parameters to get performance estimates as close as possible to live network data. This problem is formulated as an

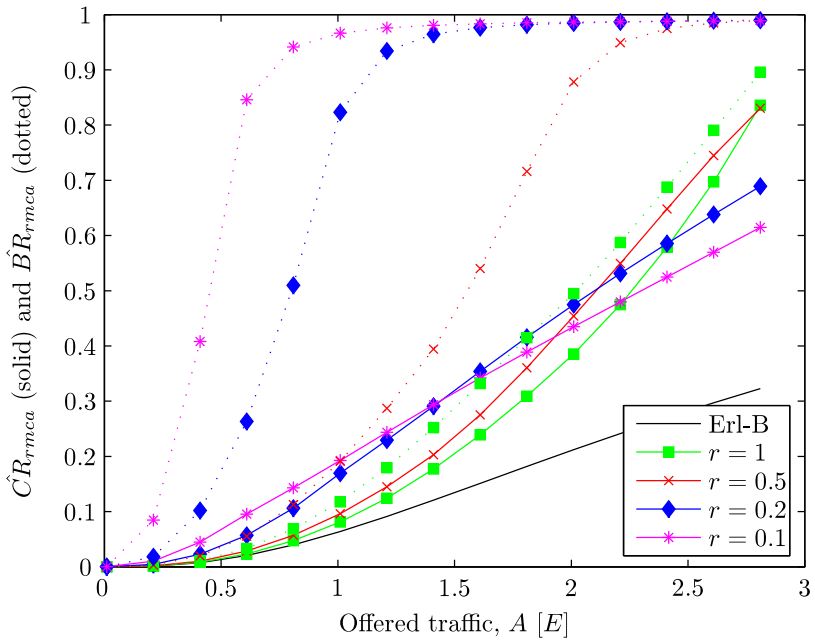


Figure 2.11: Influence of time correlation between new arrivals on SDCCH performance (case $N = 3$).

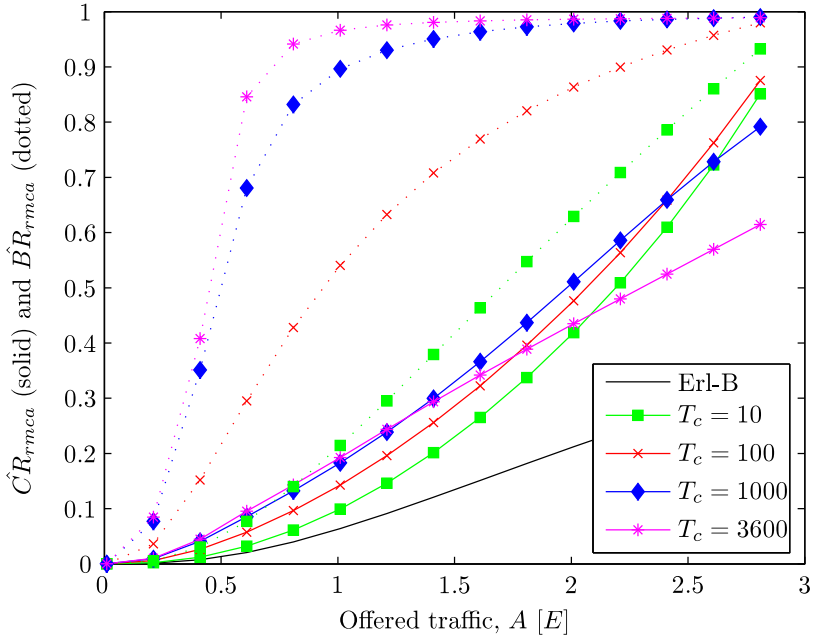


Figure 2.12: Influence of switching rate between *on* and *off* periods on SDCCH performance (case $N = 3$).

optimisation problem first. The second part of this section analyzes the feasibility of such an optimisation problem.

2.4.1 RMCA tuning as an optimisation problem

It can be observed in (2.7) that several combinations of $\lambda_{LU_{on}}$, $\lambda_{LU_{off}}$, τ_{on} and τ_{off} might give the same λ_{LU}, but completely different CR and BR, as shown in Section 2.3.2. Hence, the right set of λ and τ values must be found if a real RMCA performance (i.e., close to those measured in the network) is desired.

From the analysis in the previous section, some initial considerations about RMCA parameters can be made. For instance, if $\lambda_{LU_{on}} >> \lambda_{LU_{off}}$ and $\tau_{on} << \tau_{off}$ (i.e., most traffic demand is concentrated in a short period of time), CR is moderately low and BR is high. Therefore, and conversely, a measured value of BR much larger than the value of CR is an indication of time correlation between new arrivals. Based on this observation, the values of $\lambda_{LU_{on}}$, $\lambda_{LU_{off}}$, τ_{on} and τ_{off} in RMCA can be tuned on a cell and hourly basis so that key performance indicators given by RMCA, namely average traffic load, blocking ratio and congestion ratio, resembles as much as possible those measured in the real network.

Such a tuning process can be performed by considering the values of $\lambda_{LU_{on}}$, $\lambda_{LU_{off}}$, τ_{on} and τ_{off} as the components of a 4-vector (i.e., a point in $\Re^4$) and solving the nonlinear programming problem, [45],

$$\begin{aligned}
\text{Minimise} \quad & \left(\frac{A_c - \hat{A}_{c_{rmca}}(\lambda_{LU_{on}}, \lambda_{LU_{off}}, \tau_{on}, \tau_{off})}{N} \right)^2 \\
& + \left(CR - \hat{CR}_{rmca}(\lambda_{LU_{on}}, \lambda_{LU_{off}}, \tau_{on}, \tau_{off}) \right)^2 \\
& + \left(BR - \hat{BR}_{rmca}(\lambda_{LU_{on}}, \lambda_{LU_{off}}, \tau_{on}, \tau_{off}) \right)^2 \qquad (2.14) \\
\text{subject to} \quad & \frac{\lambda_{LU_{on}}\tau_{on} + \lambda_{LU_{off}}\tau_{off}}{\tau_{on} + \tau_{off}} = \lambda_{LU} \,, \qquad (2.15) \\
& \lambda_{LU_{on}} \geq \lambda_{LU_{off}} \,, \qquad (2.16) \\
& \tau_{on} + \tau_{off} \leq 3600 \,, \qquad (2.17) \\
& \lambda_{LU_{on}}, \lambda_{LU_{off}} \geq 0, \;\; \tau_{on}, \tau_{off} \geq 1 \,, \qquad (2.18)
\end{aligned}$$

where $\lambda_{LU_{on}}$, $\lambda_{LU_{off}}$, τ_{on} and τ_{off} are the decision variables, A_c, CR and BR are measurements in the Network Management System (i.e., constants), and $\hat{A}_{c_{rmca}}$, $\hat{CR}_{rmca}$ and $\hat{BR}_{rmca}$ are performance estimates given by RMCA (i.e., functions of the decision variables).

The objective function (2.14) reflects the goal of minimising the sum of squared errors between measurements and RMCA estimations for the average load, congestion ratio and

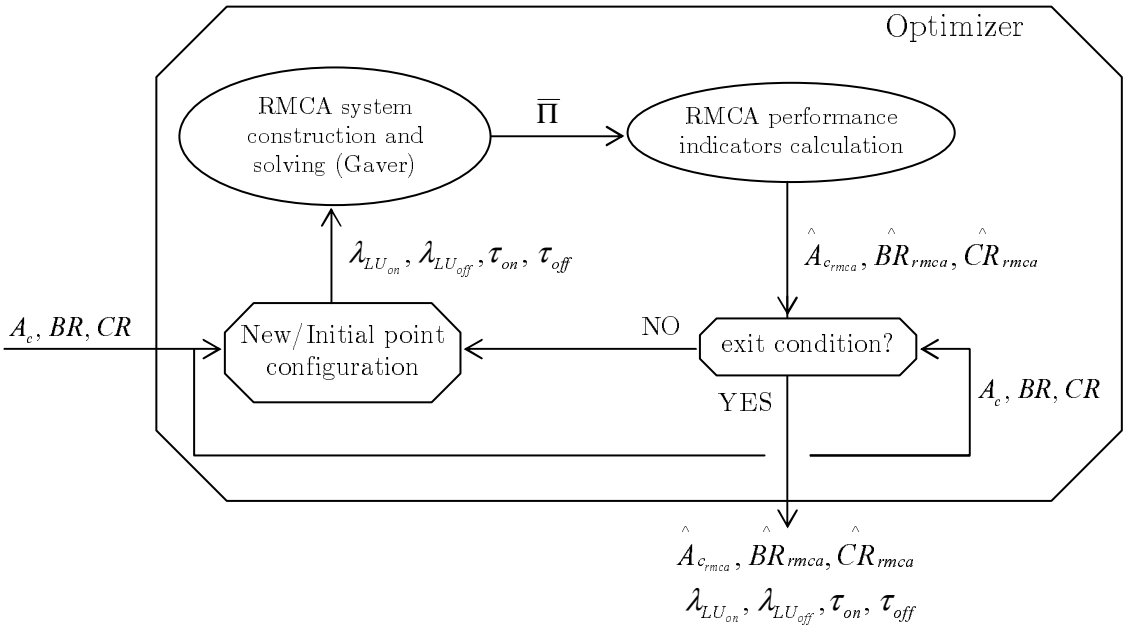


Figure 2.13: One cell RMCA optimisation flow chart.

blocking ratio. Note that estimates do not only depend on the LU traffic characteristics (defined by the decision variables), but also depend on the attempt rates due to other causes (e.g., $\hat{A}_{c_{rmca}}(\lambda_{LU_{on}}, \lambda_{LU_{off}}, \tau_{on}, \tau_{off}, \lambda_{MOC}, \lambda_{MTC}, ...)$). Since the latter are measurements (i.e., constants), they have been omitted in (2.14) for clarity. The nonlinear equality constraint (2.15) enforces the relationship between decision variables so that the average LU arrival rate coincides with the measured value. The linear inequality constraint (2.16) eliminates the symmetry in the model by forcing that the *on* period is that with the largest LU traffic. The remaining inequality constraints (2.17) and (2.18) ensure that the values of the decision variables correspond to a realistic case, where correlated LU arrivals are due to group boundary-crossing events. In practice, a typical crossing event lasts a few seconds, whereas the period between crossing events may be of up to several minutes. Constraint (2.17) discards excessively low switching rates, whose period would be larger than the measurement window in the NMS (i.e., one hour), while (2.18) are the lower-bound constraints ensuring that attempt rates are non-negative and the *on* and *off* periods last more than 1 second.

Note that the optimisation problem needs to test many points before reaching the final solution. A new RMCA system must be solved any time a new solution point (i.e. $\lambda_{LU_{on}}$, $\lambda_{LU_{off}}$, τ_{on} and τ_{off} new values) is tested. Such a process can be repeated hundreds or even thousands of times for each cell in the network. Figure 2.13 depicts the optimisation process. Any time a new point is desired to be tested, the RMCA system must be constructed and solved, RMCA performance indicators are calculated, and exit conditions are checked. The final solution is reached if at least one of the exit conditions is fulfilled. These conditions are minimum error between measured and estimated RMCA indicators, minimal distance between consecutive solution points and minimal derivative.

2.4.2 Feasibility Study for the Optimisation Problem

From a mathematical point of view, it is interesting to prove that there is always a feasible solution to the problem (2.14)–(2.18). A single point satisfying all the constraints would show that there is always a feasible solution, at least.

Constraint (2.15) is the only nonlinear equation, while (2.16)–(2.18) are linear inequalities. $\lambda_{LU_{on}}$, $\lambda_{LU_{off}}$, τ_{on} and τ_{off} are the unknowns in the optimisation process, while λ_{LU} is the only input parameter in the constraints. Note that λ_{LU} is the measured LU arrival rate and, therefore, λ_{LU} is a non-negative real value fixed in advance. Constraints (2.17) and (2.18-right) are satisfied at points where $\tau_{on} = 1$ and $\tau_{off} = 3599$. If $\lambda_{LU_{on}} = \lambda_{LU_{off}} = \lambda_{LU}$ is also set (i.e., uncorrelated new arrivals), then all the remaining restrictions are satisfied and hence there exists at least one feasible solution, regardless of the value of λ_{LU} (which always values ≥ 0). Actually, this solution is used as the starting point for the iterative optimisation algorithm.

2.5 Model Performance Assessment

Once different models have been presented and analysed, the aim of this section is threefold:

a) to show the limitations of the Erlang B formula for dedicated signalling channels in GERAN, justifying the need for new models considering effects not taken into account in the Erlang B model;

b) to prove that such limitations are due to time correlation between attempts, and

c) to prove that both retrials and correlated arrivals must be considered to estimate performance on these channels accurately, showing that RMCA is a suitable model for SDCCH performance analysis.

For this purpose, SDCCH measurements were collected in a large geographical area of a live GERAN system. Such measurements comprise both SDCCH traffic demand and queueing performance on a cell basis and hourly intervals. From measured traffic demand values, different models are constructed and solved, and key performance indicators are estimated on a cell and hourly basis by the theoretical models with and without retrials and correlated arrivals. Finally, model assessment is carried out by comparing performance estimates from the models and real measurements throughout the network. For clarity, the analysis set-up is first introduced and results are then presented.

BTS name	BH	N	A_c	CR	BR	Gl.attemps	Gl.Blocked	LU	IMSI	SS	···
···	···		···			···		···		···	
MÁLAGA5	9:00	3	2,41	0,58	0,65	2094	1363	653	1	0	···
CENTER3	18:00	3	1,58	0,26	0,42	2963	1258	1650	1	0	···
MARITI2	00:00	3	0,61	0,02	0,05	434	20	176	33	4	···
···	···		···			···		···		···	

Table 2.1: Example of SDCCH data collected.

2.5.1 Analysis Set-up

SDCCH data were collected over 8 days from 1730 cells in a live GERAN system. Such data are stored in a NMS covering about half of the operator's network. Each sample corresponds to the SDCCH Busy Hour (SDCCH-BH) of a day in a cell, so the most intense traffic conditions are considered. Thus, the original dataset consists of 8·1730=13840 samples of SDCCH performance data (8 samples per cell). The analysis is focused on cells with $N = 3$, 7 and 15, as these are the most common SDCCH configurations in the network. Each data sample contains different pieces of information, namely global number of attempts, number of attempts per cause, carried traffic, congestion ratio and blocking ratio, as shown in Table 2.1. To obtain reliable estimations, different samples are discarded when a) SDCCH traffic is less than 0.1 Erlang and b) ratio of ghost attempts is larger than 50%. Finally, data collection contains 10241 valid samples. This dataset is representative of the whole network area as it comprises 75% of cells and these samples comprise 90% of the total SDCCH traffic. Likewise, robust estimations are expected, since the dataset covers a large geographical area (i.e., 120000 km^2) with very different traffic and user mobility characteristics. Additionally, such a wide range of data ensures that a model performing these statistics is flexible enough to represent most of SDCCH traffic scenarios.

A preliminary analysis of the data shows that 19% of the 1730 cells experience unacceptable averages of SDCCH-BH blocking (i.e., $BR > 1\%$). At the same time, 10% of the cells have SDCCH TSLs that remain unused. Note that only data during the SDCCH BH were collected. These two observations (i.e., significant blocking figures and unused resources) are a clear indication that the current dimensioning approach used by operators is not working properly.

Performance results will compare three different queueing models: the Erlang Loss Model (denoted as Erl-B), the basic retrial model (RM) and the retrial model with correlated arrivals (RMCA). In both retrial models, the retrial rate, α, is set to $1/6\ s^{-1}$, as configured by the operator. A heuristic algorithm is used to fix M for both retrial models on a per-sample basis (i.e., cell and hour) so as to reduce the number of states as much as possible while ensuring that the probability that the orbit is full is negligible (<0.01). Thus, the value of M ranges from 1 to 100, depending on incoming traffic measurements.

It is worth noting that, although all models take the number of SDCCH channels, N, and traffic demand, λ and μ, from NMS data, RMCA is tuned for each sample (i.e., cell

and hour) with real data, whereas the same Erl-B and RM are applied to all samples in the network. Therefore, $\lambda_{LU_{on}}$, $\lambda_{LU_{off}}$, τ_{on} and τ_{off} in RMCA are calculated for each sample by solving the optimisation problem (2.14)–(2.18) with real network statistics, namely A_c, BR, CR and λ_{LU}.

Performance estimates for Erl-B are computed by the Erlang B formula, (2.1), where $\hat{CR}_{ErlB}$=$\hat{BR}_{ErlB}$=$E(A,c)$, c=N (i.e., the number of signalling subchannels), and A is the total offered SDCCH traffic, with no distinction between retrial, non-retrial and *on-off* rates. Carried traffic for Erl-B is easily calculated as $\hat{A}_{c_{ErlB}} = A(1-\hat{CR}_{ErlB}))$. In the absence of an equivalent expression for RM and RMCA, performance is estimated by numerical methods. For each sample and retrial model, a new matrix $\mathbf{Q}$ is generated and the stationary distribution is computed by solving (2.3) by the Gaver's algorithm. Then, key performance indicators are calculated as in (2.4)–(2.6) or (2.8)–(2.10). In the case of RMCA, the model is tuned by solving (2.14)–(2.18) with the *fmincon* function in MATLAB Optimisation Toolbox, [46], initialised to $\lambda_{LU_{on}} = \lambda_{LU_{off}} = \lambda_{LU}$, $\tau_{on} = 1$ and $\tau_{off} = 3599$, following the flow chart depicted in Figure 2.13. During the tuning process, (2.3) must be solved several times for each cell and hour, as $\mathbf{Q}$ in RMCA changes with different parameter settings.

Model assessment is based on two figures of merits. From the operator side, the most important criterion is the error in determining the number of fresh blocked attempts for revenue-generating services (i.e., MTC, MOC, EC, SS and SMS). Such an error has a direct translation to economical revenues. For instance, a model underestimating BR will cause more fresh blocked attempts than expected, with an associated revenue loss in the real network. On the other side, overestimating BR leads to waste signalling resources, which were assigned expecting a higher traffic than it is presently being offered. Therefore, an adequate goodness-of-fit measure is the normalised sum of absolute errors for the blocked arrival rate of revenue-generating services, $NSAE_{brgs,m}$,

$$NSAE_{brgs,m} = \frac{\sum_{i=1}^{N_{sam}} \lambda_{rgs}(i) \left| CR(i) - \hat{CR}_m(i) \right|}{\sum_{i=1}^{N_{sam}} \lambda_{rgs}(i)\, CR(i)}, \tag{2.19}$$

where N_{sam} is the number of samples (i.e., cells and hours), $\lambda_{rgs}(i)$ is the total fresh arrival rate of revenue-generating services in sample i, $CR(i)$ is the measured congestion ratio in sample i, and $\hat{CR}_m(i)$ is the congestion ratio for sample i suggested by model m, where $m \in \{\text{Erl-B}, \text{RM}, \text{RMCA}\}$. Note that fresh attempts of these services are Poisson arrivals and, therefore, CR equals the probability of finding all sub-channels busy. This figure of merit is dominated by cells and hours with a larger revenue-generating traffic (λ_{rgs} high), i.e, if a small error in CR estimate occurs in a cell with a high $\lambda_{rgs}(i)$, that cell has a large contribution in the global $NSAE_{brgs,m}$ value. Thus, a few samples from the global dataset (those with a large λ_{rgs}) have the major contribution to $NSAE_{brgs,m}$.

All cells, services and performance indicators are equally important from the academic

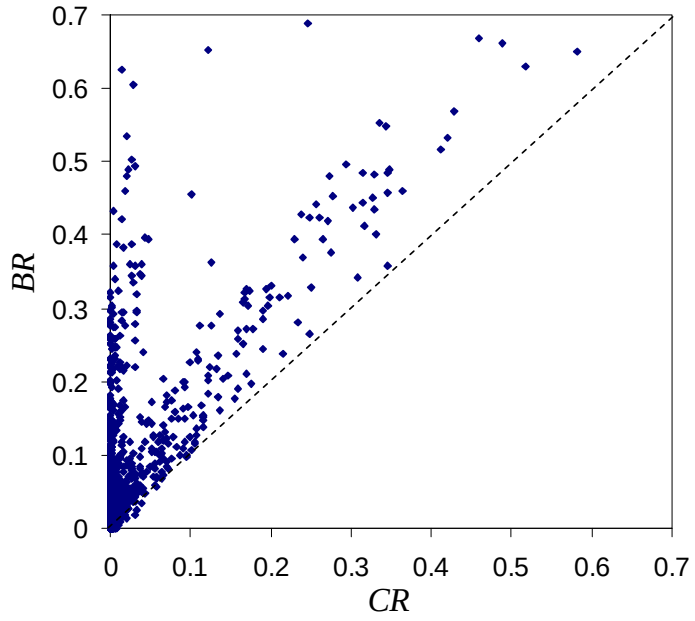


Figure 2.14: SDCCH congestion ratio versus blocking ratio in a live network.

side. On this premise, a more adequate goodness-of-fit measure is the average sum of squared errors for the average load, congestion ratio and blocking ratio, $\overline{SSE}_m$,

$$\overline{SSE}_m = \frac{\sum_{i=1}^{N_s} \left(\left(\frac{A_c(i) - \hat{A}_{c_m}(i)}{N(i)} \right)^2 + \left(CR(i) - \hat{CR}_m(i) \right)^2 + \left(BR(i) - \hat{BR}_m(i) \right)^2 \right)}{N_s}, \quad (2.20)$$

where $A_c(i)$, $CR(i)$ and $BR(i)$ are measurements, $\hat{A}_{c_m}(i)$, $\hat{CR}_m(i)$ and $\hat{BR}_m(i)$ are estimates from model m, and $N(i)$ is the number of SDCCH sub-channels in the cell of sample i. In this second figure, estimate errors are not weighted by the traffic, so similar errors in low or high traffic cells are equally important.

2.5.2 Results

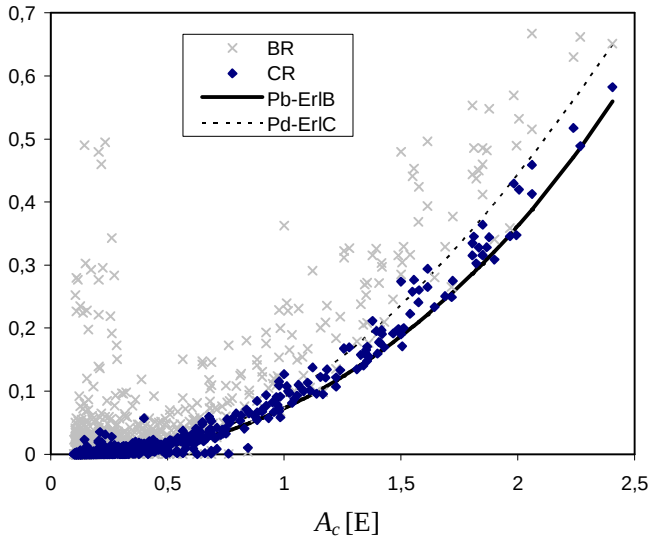
The first experiment checks if the data arrival process is Poisson distributed by checking the *Poisson Arrivals See Time Averages* (PASTA) property, [47], over real SDCCH measurements. Figure 2.14 shows a scatter plot of CR versus BR, together with a dashed line representing $BR = CR$ (representing Poisson arrival processes). It is observed that, in many cases, BR is significantly larger than CR. This is a clear indication that the Poisson assumption does not hold for the SDCCH traffic.

Figure 2.15 confirms this statement by representing CR and BR measurements in terms of the average carried traffic for cells with 3, 7 and 15 sub-channels. For comparison purposes, congestion values given by the Erlang B and C formulas (i.e., blocking and delay probability in a loss and delay system, respectively) are superimposed. Note that an Erlang C system is close to a retrial system with retrial rate $\alpha \to \infty$. The analysis is

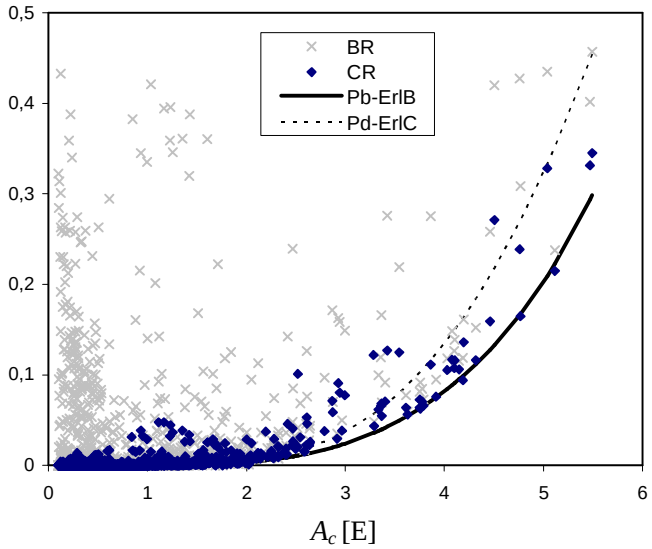
first focused on cells with N=3, represented in Figure 2.15(a). It is observed that, in most cases, both CR and BR are above the Erlang B curve. Thus, for large traffic values, BR can be up to twice the value predicted by the Erlang B formula. Similar results are observed in cells with N=7 and 15, presented in Figure 2.15(b)-(c). A more detailed analysis (not shown here) reveals that the value of the blocking probability given by the Erlang B formula falls outside the 95% confidence interval of BR in 15% of the samples, so Erlang B cannot model SDCCH traffic behaviour for those samples at all. More important, the problematic samples are those with a larger blocking ratio, comprising 26% of the total SDCCH traffic in the network. Therefore, these samples are the main focus of network re-planning procedures.

These estimation errors are partly due to retrials. On the one hand, retrials make the network behave, in some sense, like a delay system. Thus, retrials tend to enlarge congestion periods (and, hence, increase CR), as channels are occupied by users in the orbit as soon as they become free. As a result, CR tends to be larger than Erlang B blocking probability for the same value of A_c. This is confirmed by the fact that most CR samples in Figure 2.15(a) lie between Erlang B and C curves, which is a well-known property of retrial queues, [32]. On the other hand, due to retrials, attempts are not statistically independent, but are concentrated around congestion periods. This justifies that $BR >> CR$, which is also a well-known effect of retrials, [29]. Similar trends are observed in cells with N=7 and 15, represented in Figure 2.15(b)-(c). However, while $BR > CR$ in both figures, CR is well above the Erlang C curve in many samples in Figure 2.15 (c). From this observation, it is envisaged that RM will fail to explain congestion in cells with a large number of sub-channels.

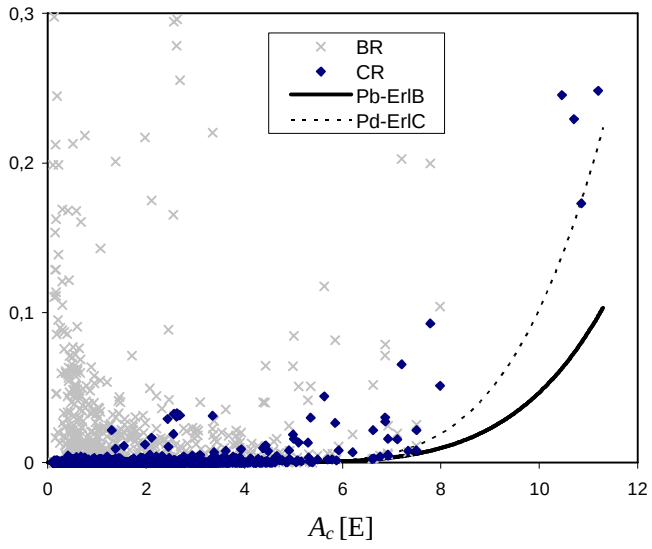
The previous hypotheses are confirmed by the overall estimation results. Table 2.2 presents the values of the two goodness-of-fit measures, $NSAE_{brgs}$ and $\overline{SSE}$, for the three queueing models. Results have been broken down by number of sub-channels and last column shows the values for all the samples. In the table, it is observed that Erl-B (1^{st} and 4^{th} rows) is the worst model, as it gives the largest value of $NSAE_{brgs}$ and $\overline{SSE}$ for any number of channels. From the former indicator, it can be inferred that the error in estimating the number of blocked fresh attempts of revenue generating services by the Erlang B formula is 23%, 42% and 82% for N = 3, 7 and 15, respectively, which is a very large value. As already pointed out, the error is much larger in cells with large N, where the Erlang B formula fails to explain congestion without retrial and correlation mechanisms. In contrast, $\overline{SSE}$ for Erl-B decreases with N due to the fact that there are fewer cells with congestion problems when N is large, and $\overline{SSE}$ gives equal importance to all performance indicators and cells. Similarly, RM shows large values of $NSAE_{brgs}$ and $\overline{SSE}$, close to Erl-B estimates. Thus, it can be concluded that retrials can only explain a small fraction of blocking. In contrast, RMCA gives the lowest values of $NSAE_{brgs}$ and $\overline{SSE}$ for any number of sub-channels, and, consequently, for the whole network. From the last column, it can be deduced that, with RMCA, the overall $NSAE_{brgs}$ and $\overline{SSE}$ are reduced by 63% and 77%, respectively, when compared to Erl-B. This is a clear indication of the superior accuracy of RMCA. For $NSAE_{brgs}$, the benefit is more evident for large N, since this figure is dominated by a few cells where only RMCA can explain blocking.



(a) $N = 3$



(b) $N = 7$



(c) $N = 15$

Figure 2.15: SDCCH congestion and blocking performance.

Measure	Model	N=3	N=7	N=15	N={3,7,15}
$NSAE_{brgs}$	Erl-B	0.2272	0.4198	0.8242	0.4904
	RM	0.2176	0.4065	0.6500	0.4247
	RMCA	0.1519	0.1742	0.2244	0.1835
$\overline{SSE}$ $[\cdot 10^{-2}]$	Erl-B	0.4745	0.1812	0.0733	0.2430
	RM	0.3700	0.1742	0.0696	0.2046
	RMCA	0.0383	0.0863	0.0452	0.0566

Table 2.2: Performance of queueing models for different number of sub-channels.

In contrast, the benefit in $\overline{SSE}$ is more evident for small N, as this indicator is an average of many cells and there are more cells with congestion problems in relative terms for small N.

The previous result was expected, since parameters in RMCA are adjusted on a cell/hour basis to fit performance data, and, consequently, performance estimate errors should be very low after the optimisation process. More interesting is the analysis of the final RMCA settings, which can reveal the amount of cells with a high time correlation in SDCCH traffic. Such an analysis shows that 8% of the samples display values of $(\lambda_{r_{on}} + \lambda_{nr})/(\lambda_{r_{off}} + \lambda_{nr}) > 2$, i.e., the signalling traffic rate during the *on* periods doubles the traffic rate during the *off* period. In order to give a qualitative estimation of that value, note that this is more than half of the samples where the Erlang B formula failed (i.e., 15%). More important, these samples comprise 59% of the total blocked attempts for all the samples. These results clearly indicate the need for considering time correlation between new arrivals when estimating SDCCH performance in a real network.

2.5.3 Implications for SDCCH Re-dimensioning by operators

SDCCH models can be used in the network design stage. Once the network is in its operational stage, and having observed how poorly Erl-B/C predict SDCCH performance, it is clear that a precise performance model is needed to re-dimension SDCCH resources on a cell basis based on network statistics, as RM and, specially, RMCA do. The main challenge is to identify cells where the number of SDCCH sub-channels is unnecessarily high, as the opposite situation (i.e., excessively low) can simply be detected from BR statistics. Results have shown that the offered SDCCH traffic cannot be estimated by the sum of carried and blocked attempts, since part of the latter are retrials from the same original attempts. Likewise, the Erlang B formula currently in use is not adequate for many cells. To re-dimension SDCCH resources, the operator should first check if $CR \simeq BR$ and then check if both indicators coincide with the blocking probability obtained by the Erlang B formula, $P_{b_{ErlB}}$. In cells where $CR < P_{b_{ErlB}} < BR$ or $P_{b_{ErlB}} < CR < BR$, RMCA can be adjusted from network statistics to derive the actual offered traffic and its temporal distribution. Once tuned, RMCA can be used to predict blocking performance with a different number of sub-channels. In the rare cases where $CR < BR < P_{b_{ErlB}}$, none of the previous models is valid. The latter situation is typical of cells with limited

population and large SMS traffic due to WAP-over-SMS traffic, [48]. For these cells, a finite source queueing model (e.g. Engset distribution, [1]) is more adequate.

The main drawback of the proposed methodology when compared to the current approach is the increased computational load. Specifically, the execution time for the 10241 samples (i.e., all the data in a NMS) in a 2.4GHz 2GB-RAM Windows-based computer is 2 hours, most of which is spent in samples with a large value of N and a high level of congestion, where M has to be set to 100 to get accurate results. Note again that tuning RMCA requires running the Gaver's algorithm hundreds of times per sample. The execution time might be reduced by substituting the current finite truncation approach in the orbit by generalised truncated methods, [35]. Nonetheless, the current execution time is low enough for network re-planning purposes.

2.6 Conclusions

Due to financial pressure, operators are increasingly forced to maximise the financial return on their investment in GERAN. To achieve this aim, operators rarely add resources to solve congestion problems, but try to ensure that every time slot is assigned to the most suitable usage, signalling or traffic, i.e., SDCCH or TCH. To assist operators in this undertaking, a comprehensive performance analysis of dedicated signalling channels for operator re-planning purposes has been performed in a live GERAN system.

Preliminary analysis has shown that the Erlang B formula, currently used by operators, fails to give adequate estimates of the SDCCH blocking ratio in 15% of measurements, proving that Erlang B cannot model SDCCH traffic behaviour for those samples at all. More important, the problematic samples correspond to cells with the largest SDCCH blocking, receiving most of the attention from the operator. A first analysis has been carried out introducing retrials. Such a retrial model showed important limitations when the number of signalling channels is large, performing very similar to Erlang B models. Therefore, estimation errors are not reduced significantly when only considering retrials.

To overcome these limitations, a retrial queueing model with correlated arrivals has been proposed. Time correlation between new arrivals is modelled by a switched Poisson process with *on* and *off* states. Location Update requests are the only traffic component considered to show time correlation between attempts. The resulting model is simple and flexible enough to be adjusted on a per-cell and per-hour basis using statistics in the Network Management System. Such a tuning problem is formulated as an optimisation problem. For computational efficiency, a special method has been used for solving the system state-transition diagram in this correlation model. With the proposed model, the sum of squared residuals for the main performance indicators is reduced by 77% when compared to the current, i.e., based on the Erlang B formula, approach.

It is clear that more complex models considering differences between retrials and redials, [7][9][39], or more general distributions of inter-arrival, service and inter-retrial times,

[9][36], could obtain more accurate predictions. However, such models are more difficult to tune, as they require knowledge of the traffic attributes on a cell basis. Such knowledge can only be obtained by a time-consuming analysis of traffic traces, which are seldom available.

Chapter 3

Optimal Traffic Sharing in GERAN

In a mobile communications network, the uneven spatial distribution of traffic demand and the spatial re-distribution with time can be dealt with by sharing traffic between adjacent cells. In this chapter, two analytical teletraffic models for the traffic sharing problem in GSM-EDGE Radio Access Network (GERAN) are presented. From these models, a closed-form expression is derived for the optimal traffic sharing criterion between cells through service area redimensioning. Then, performance analysis is detailed through the comparison of several traffic balance criteria in different scenarios. These scenarios are built according to live data from GERAN networks, and they include restrictions in the traffic sharing process, as it is expected to find in real networks. Finally, results compare different balance criteria through network traffic performance indicators.

3.1 Introduction

In the last few years, the success of mobile communication services has caused an exponential increase of traffic in cellular networks. More than 500 million GSM mobile users can be found nowadays only in western Europe, [3]. The initial design of a mobile network takes into account traffic forecast. But, in addition to those forecasts, lots of models about the behaviour of the different network elements are also needed: propagation channel, user daily movements, traffic flow characterisation, etc. Good traffic models allow better dimensioning of network resources during the design stage. Work in Chapter 2 is a good example of a model improvement for a better resource dimensioning. The better the model and forecasts are, the better the mobile network performances.

However, due to network evolution and traffic increase and re-distribution, the matching between the initial planned distribution of resources and actual traffic demand becomes looser. As stated in Chapter 2, present network and traffic conditions can be quite

different from their initial values. Current research activities over GERAN are focused on network replanning and optimisation, specially trying to bridge the gap between original design conditions and actual traffic scenario. Such a mismatching is usually reflected in congestion problems. Hence, the availability of tools to manage the dynamic nature of traffic demand becomes crucial during the network operational stage and network adaptation becomes a key feature in mobile networks. The need for adaptation starts up a re-planning process that usually concludes with the addition or re-allocation of network resources. But these resource changes are not used very frequently since it supposes too high operational and capital expenditures. In the meantime, traffic management becomes the main tool for solving dynamic network congestion problems and for maintaining the Quality of Service (QoS). *Traffic management* refers to the set of algorithms and policies that allow the network to provide adequate QoS, with existing resources and infrastructure, [12], which is precisely its main benefit.

Managing Circuit-Switched (CS) traffic in a cellular network mainly consists of selecting the base station to which every mobile station is attached. Traffic management algorithms are implemented by Radio Resource Management (RRM) algorithms. Three are the main RRM algorithms dealing with CS traffic management in GERAN: admission control, congestion control and load balancing. Basically,

- *Admission Control* (AC) evaluates the initial serving BTS by checking the availability of radio resources and interference levels; if minimum signal levels are not reached or no radio resources are available, the call is blocked;
- *Congestion control* (CC) is in charge of detecting and managing excessively high traffic demand situations in the network;
- Finally, *load balancing* (LB) is in charge of traffic re-distribution between cells to avoid congestion.

Fast traffic fluctuations in GERAN are dealt with by advanced RRM features, such as dynamic load sharing, [49], and dynamic half-rate coding, [50]. However, advanced RRM procedures are unable to solve localised congestion problems caused by spatial concentration of traffic demand, as shown in [12]. In the long term, these problems are solved by re-planning strategies, such as the extension of transceivers or cell splitting. Nevertheless, in the short term, the adaptation of cell service areas is the only solution for cells that cannot be upgraded quickly. Cell resizing is performed by modifying base station transmit power, [51], antenna uptilting/downtilting, [52], or adjusting RRM parameter settings, [53][54][55]. Traffic load can be shared with surrounding cell by tuning parameters in the HandOver (HO) algorithm, [56]. HO is usually seen as the main network mechanism to keep a global coverage for the user. In addition, HO defines the cell service area as the area wherein all users are connected to a specific BTS. Service area boundaries are the locations where a user leaves a BTS to be handed over towards the neighbour cell. If some BTS service area can be reduced (and the service area of neighbouring cells is consequently enlarged), users located at the border are sent to surrounded cells, and BTS load

in the origin cell is reduced. As a consequence, the load of surrounded cells is increased. Changes of service area can be implemented by modifying HO margins, [57][58][59].

Nonetheless, there is still the need of finding the best strategy to modify HO margins. Most cell resizing approaches aim at minimising the total blocked traffic in the network (i.e., maximising the total carried traffic). But this general goal must be translated into an easy-to-implement rule for network equipment, and, then, such a global goal is usually broken down into local, and heuristic, criteria, e.g., to equalise the blocking probability or blocked traffic of any cell in the network. Operators generally implement the load balancing heuristic criterion of equalising some performance indicator on a cell basis, with the hope that the total blocked traffic is thus minimised. As an example, operators often equalise call blocking ratios across the network, i.e.,

$$BR_1 = BR_2 = \cdots = BR_i = \cdots = BR_N \,, \tag{3.1}$$

where N is the number of cells in the network and BR stands for Blocking Ratio. Thus, parameter changes are implemented so that blocking ratios are equalised in the network. Some other performance indicators could also be used to be equalised (e.g., cell traffic load or blocked traffic on a cell basis). Assuming that parameter changes are correctly made, (3.1) is usually fulfilled. However, there is no proof that, by enforcing (3.1), the global optimal solution (i.e., the total number of blocked calls in the network) is achieved.

This chapter presents a novel method for determining the best (i.e., the optimal) call traffic sharing criterion between cells in a GERAN system. The proposed method computes an optimal indicator on a cell basis, given a certain distribution of network resources and a spatial distribution of traffic. The main decision variables are the traffic demand originated by fresh calls and handovers in each cell. Call traffic is moved between adjacent cells to equalise the optimal indicator. Two queueing models are presented, corresponding to the cases when traffic is reallocated by tuning parameters in the call admission control or the handover algorithm. For both models, a closed-form expression of the optimal traffic sharing criterion is derived based on the properties of the Erlang-B formula. The analysis shows that the common rule of balancing call blocking rates between adjacent cells is not the optimal strategy (i.e. total blocked traffic is not minimal). Performance assessment is carried out in a set of realistic test cases. During the tests, the proposed exact method is compared with other heuristic balancing criteria currently used by network operators.

Figure 3.1 illustrates this chapter's structure in a graphical way. Two new system models are proposed and each optimal balance criteria are extracted from them. Model performance assessment is carried out through the comparison of optimal criteria with some other heuristic balance criteria. These optimal and heuristic approaches are evaluated in several scenarios, constructed from real GERAN network statistics and configuration. The main contribution in this chapter is a novel criterion for balancing circuit-switched traffic between adjacent cells, which can easily be integrated in automatic network optimisation

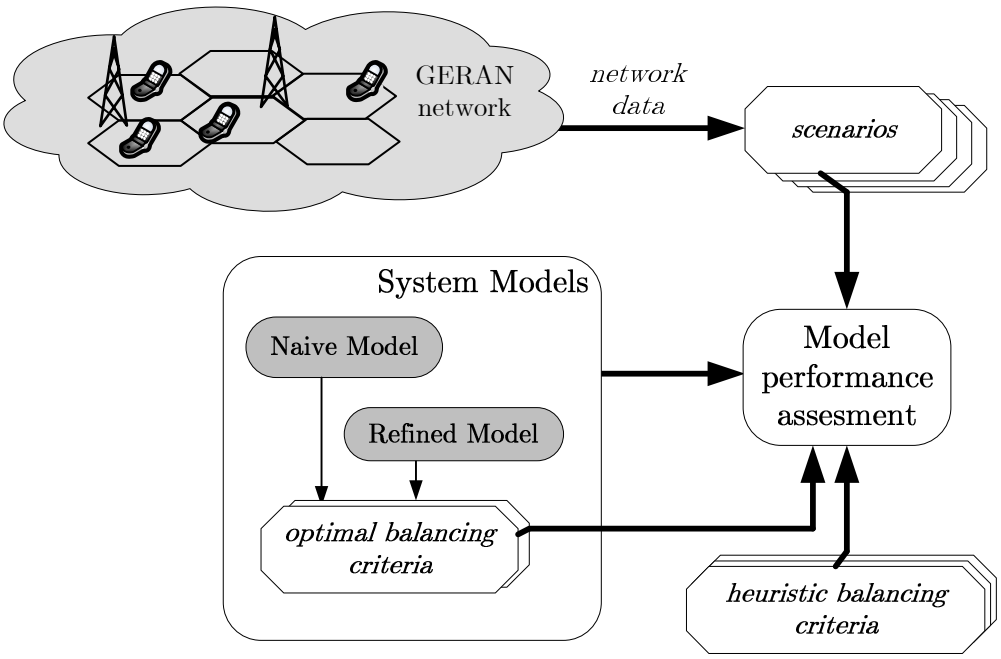


Figure 3.1: General working scheme for Chapter 3.

tools.

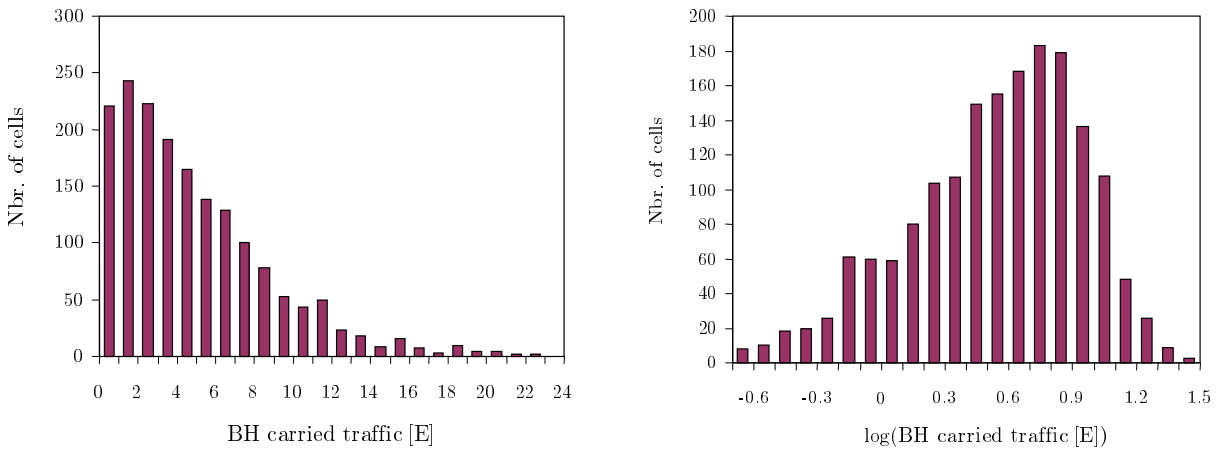
The rest of the chapter is organised as follows. Section 3.2 details the traffic sharing problem for GERAN networks. Section 3.3 presents two network models for traffic sharing, and each optimal balance criteria are extracted and analysed. Section 3.4 implements the performance analysis and results of both optimal and heuristic traffic sharing criteria for different scenarios and, finally, Section 3.5 presents the main conclusions of the study.

3.2 Problem outline

In this section, the problem of traffic sharing in GERAN is presented. First, the origin of congestion in mobile networks is analysed. Then, teletraffic models are described as an important tool for mobile network design and optimisation. Finally, the state of the research and technology related to the topic is detailed. The issues presented here will justify the need for the models and tools presented in next sections.

3.2.1 The Traffic Sharing Problem in GERAN

With continuous increase of user demand, mobile operators are forced to increase network capacity (e.g., by adding new transceivers or new cells grouped in different network hierarchy levels). In the past, the decision for allocating new resources was usually taken based on weekly averages of the daily peak traffic, [60]. As a consequence, the network was most often over-dimensioned. Nowadays, such an approach is no longer valid due to



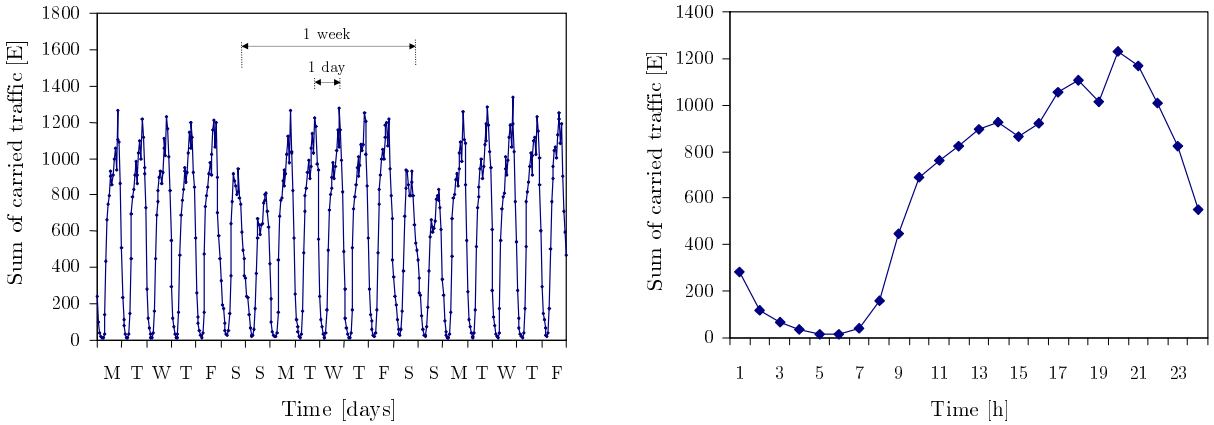
(a) Linear (b) Logaritmic

Figure 3.2: Traffic distribution in a cell level in a live network, [2].

explosive traffic increase, scarcity of radio resources and pressure on operators to reduce capital expenditures, especially in mature technologies such as GERAN. Thus, the addition of new resources must be kept to a minimum. As a consequence, traffic growth leads to network congestion and, consequently, user call blocking and revenue losses. Even if the total amount of resources in the network is enough, the uneven distribution of traffic demand, both in time and space, and the mismatching between traffic demand and initial resource distribution in the network are also major contributors to network congestion.

Figure 3.2 reflects the non-homogeneity of the spatial traffic distribution, showing the histogram of BH traffic in a live network. As pointed out in [10], the spatial traffic distribution on a cell basis can be modeled by a log-normal distribution, as shown in Figure 3.2 (b). This is the result of concentration of users in urban environments, where traffic is located in highly loaded areas, referred to as *hot-spots*, located around business activities, [11]. Moreover, traffic distribution also reflects a strong time correlation. Temporal traffic profiles can be classified in short-term and long-term trends, [11]. Long-term trends refer to overall traffic growth, seasonal changes or population movements that remain stable (e.g., premise openings). In a short-term scale periodic events can be found, such as daily or hourly traffic fluctuations along a day. Figure 3.3 plots an example of the temporal fluctuation on an hourly basis in a Base Station Controller (BSC) for 3 weeks. Figure 3.3(a) shows that working days are the ones with the largest traffic. Figure 3.3(b) shows that, within a day, several traffic peaks are observed, corresponding to the morning, afternoon and night periods.

Although traffic patterns are repeated periodically, traffic peaks are not the same in each cell, both their magnitude and the hour of the day vary. Differences are in a short-time scale due to the randomness of traffic demand. Even for a larger time scale, where traffic randomness is averaged out, important traffic patterns differences can be found, [2]. As an example, residential and business areas experience similar patterns but shifted in time.



(a) 3 weeks (b) 1 day

Figure 3.3: Traffic distribution on an hourly basis in a live BSC, [2].

One of the most interesting conclusions from the previous traffic analysis is that congestion is not only caused by the global lack of radio resources in the network. On the contrary, congestion can also be seen as a local effect where a highly loaded cell is surrounded by underused cells. In this situation, load sharing becomes an efficient way of solving congestion in loaded cells. Excessive traffic can be redirected to surrounded, and underused, cells.

Load sharing can be performed by several means, amongst which is admission control. AC re-directs incoming connection requests during call setup to balance traffic load in the network. Thus, a new incoming user can be assigned not to the best (from the radio perspective) BTS, but to any other one which is less loaded. Unfortunately, once the user is accepted in the cell selected by AC, the network has no control on user movements. Thus, the user can be handed back to the original cell with a high load and, therefore, the initial AC decision would have little effect. Thus, load sharing decisions taken by AC cannot be maintained throughout the call duration because of user movement. Figure 3.4 shows this scenario. Symbols '+' represent BTS sites, solid lines are the borders of the cell service areas, and the grey area represents the coverage area of the omnidirectional cell 1. In this scenario, a new call from the user could be attached to cell 1 by AC for traffic sharing purposes. That assignment to cell 1 is possible since the user is under the coverage area of cell 1. Nevertheless, once the call is in progress, a HO process is triggered and the user is handed over to cell 5, which is the best serving cell in this scenario.

Such a shortcoming in AC decisions is the main reason for using HO parameters for load sharing. As a call progresses, the user might leave the service area of the initial assigned cell and enter that of a surrounding cell. This is more likely as service areas become smaller, as it is usually the case in urban areas. The HandOver Control (HOC) process ensures that a user is always connected to the best serving cell and one user can be connected sequentially to several cells in the same call. HOC decisions might cause that balancing actions taken by AC become ineffective, as HOC prevails over other mechanisms. In [2], a wide description of HO algorithm and parameters and how load sharing can be implemented through HO parameter changes are found.

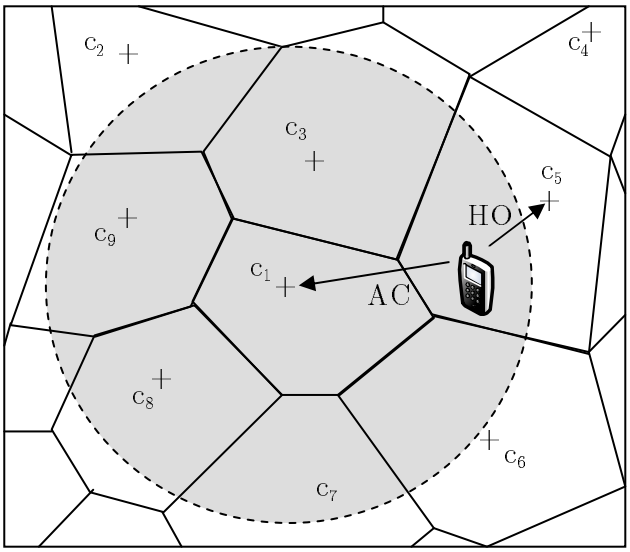


Figure 3.4: AC and HO different cell assigments.

Teletraffic models arise here as a powerful tool for analysing mobile network systems and, more important, checking the impact of different traffic sharing criteria on network performance. A teletraffic model consists of: a) assumptions about the general system or traffic behaviour (e.g., there are no user retrials or there is an infinite number of users), b) some probabilistic models of user behaviour (e.g., Poissonian call arrival rate or exponential service time), c) model parameters (e.g., incoming average traffic demand or amount of available resources), and d) definitions of Performance Indicators (PI), computed from the value of system state probabilities.

In some cases, the model state probabilities can be expressed in a closed form and, consequently, an analytical formula for model performance indicators can be obtained without the need of solving the state-transition diagram. In such cases, performance analysis is simple. In contrast, RM and RMCA in Chapter 2 are examples for a non-analytical solution, which is expressed by a set of state-transition probabilities.

Once network behaviour is modelled by equations, a goal function can be defined that relates network parameter and performance indicators. Thus, a classical optimisation process can be used to tune network parameters under the control of the operator. As in any other optimisation process, constraints can also be introduced. Figure 3.5 shows all the process and how network models and optimisation are used. The behaviour of a network is influenced by multiple parameters. Some parameters are controlled by the operator (e.g., RRM parameters or BTS configuration), and some others are not (user movements or propagation channel behaviour). Network optimisation modifies parameters under operator's control to achieve a goal function. Network models are used as a platform to test different parameter settings before getting the definitive values to be downloaded to the real mobile network.

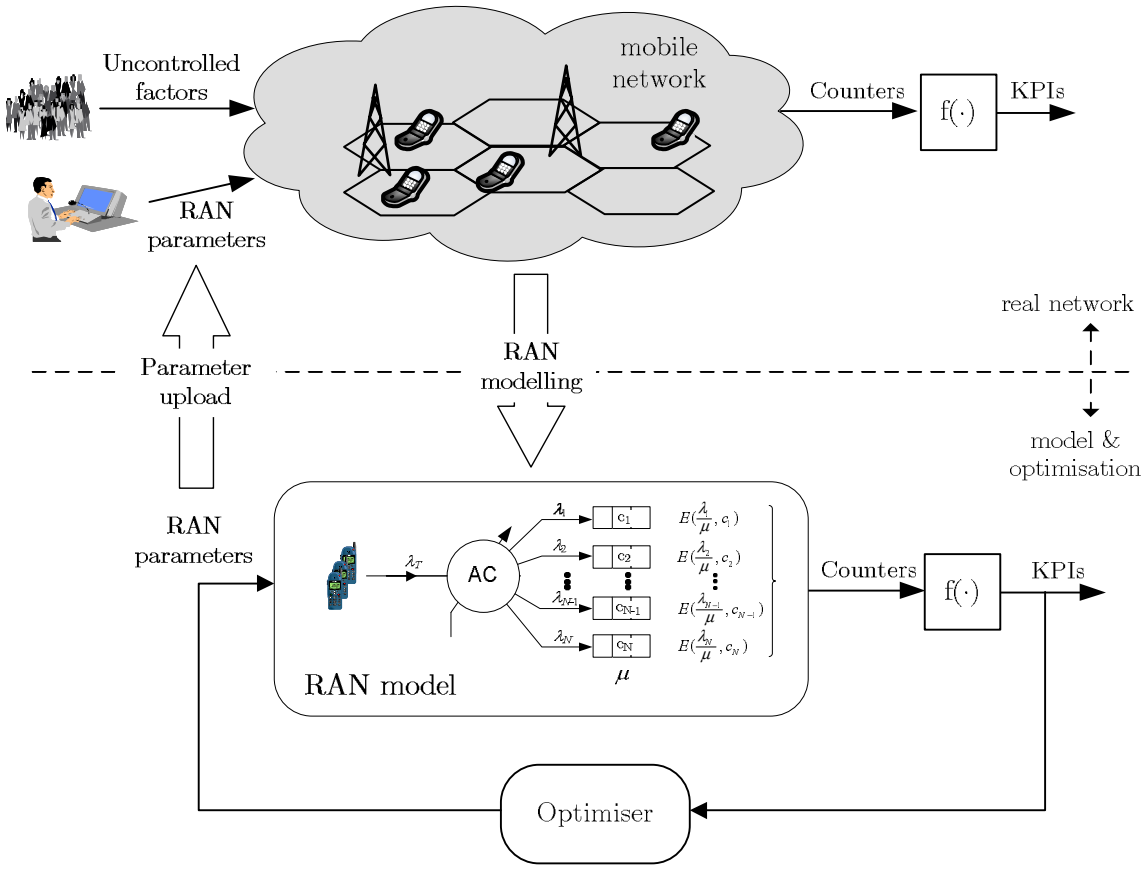


Figure 3.5: Network optimisation using teletraffic models.

3.2.2 State of Research

The traffic sharing problem in mobile networks has two main focus areas in the literature: development of teletraffic models and the optimisation (in a wide sense) of the performance of different teletraffic models.

A first group of references present analytical teletraffic models for Time-Frequency Division Multiple Access (TDMA/FDMA) cellular networks. Hong and Rappaport, [16], proposed a traffic model and analysis for cellular mobile radio telephone systems with handover. Several resource management schemes are analysed, with and without priorisation for handover attempts. They formulate a classical Markov chain model with Poisson fresh call and handover arrivals, and exponential service times to evaluate the performance of prioritising and queueing handover requests. Hong and Rappaport's model can be considered as a milestone for cellular teletraffic engineering. Subsequent references extend the initial proposal. Guerin, [61], extended the model to consider queueing of both fresh call and handover requests in the system with a similar approach to that used in Chapter 2 (i.e., a 2-D state transition diagram). Other references, [62][63], extend the traffic model by a multi-flow scheme (i.e., users with different rates and dwell times). [64] improved the model by considering a general distribution of channel holding time. [64] also considers tuning of model parameters for a better adjustment to live network data.

Additional extensions have been progressively introduced, such as time correlation of incoming calls, [43][65], or user retrials, [8][9]. Another step forward in network system modelling is the introduction of multiple services, [20][66]. Different user connections do not only differentiate in temporal characteristics (e.g., service time), but they can also demand different amount of network resources. Such a new scenario requires new indicators, since the network does not behave the same depending on the service, e.g., a new streaming service can be rejected by one cell, but a simple audio call can be accepted right after. These new multi-service network models include both real and non-real time services. Non-real time traffic demand the addition of new characteristics to the system model, such as different levels of prioritisation, or specific buffering queues and resource reservation, [66].

To cope with the explosive increase of mobile traffic, multilayered (or hierarchical) networks have been proposed as a solution for congested areas. These networks increased the capacity with several cells in the same geographical area, but different frequency allocation. Teletraffic models for hierarchical networks are presented in [21][24].

As a logical consequence, all the previous models models, conceived for TDMA/FDMA systems, have been extended to other radio access technologies, namely Code Division Multiple Access (CDMA), [67][68], and Orthogonal Frequency Division Multiple Access (OFDMA), [69].

Most of the previous work uses queueing models to evaluate the performance of novel RRM algorithms. However, not so many references use these models to find an optimal configuration of the network model parameters. In [70], the design of a multi-layered network is solved as an optimisation problem, whose goal is to minimise total system cost. The decision variables are cell sizes and the number of channels per layer. In [71], the minimum number of channels per traffic class in a channel reservation scheme is obtained based on an analytical model of a multi-service scenario. More related to this work, [57] and [58] formulate the traffic sharing between adjacent cells as a classical optimisation problem. In their approach, the goal is to minimise call blocking in the network and the decision variables are the handover margins, defined on a per-adjacency basis. For this purpose, the spatial traffic distribution is estimated from measurement reports and mobile positioning data, respectively. In [72], an analytical model of Wideband CDMA (WCDMA) cell capacity is used to minimise the total downlink interference in a real scenario by tuning sector azimuths and antenna tilts. However, none of these references propose a closed-form expression of the optimal solution for the traffic sharing problem. Thus, in these studies, the best solution is found by heuristic search methods, without any proof of optimality.

The main contribution in this chapter is the definition of an optimal criterion for traffic sharing. Previous works define different criteria for load sharing, following a heuristic approach. In this work, a network model has been defined and global performance indicators extracted. Through a classical optimisation process, a new performance indicator to be balanced has been defined.

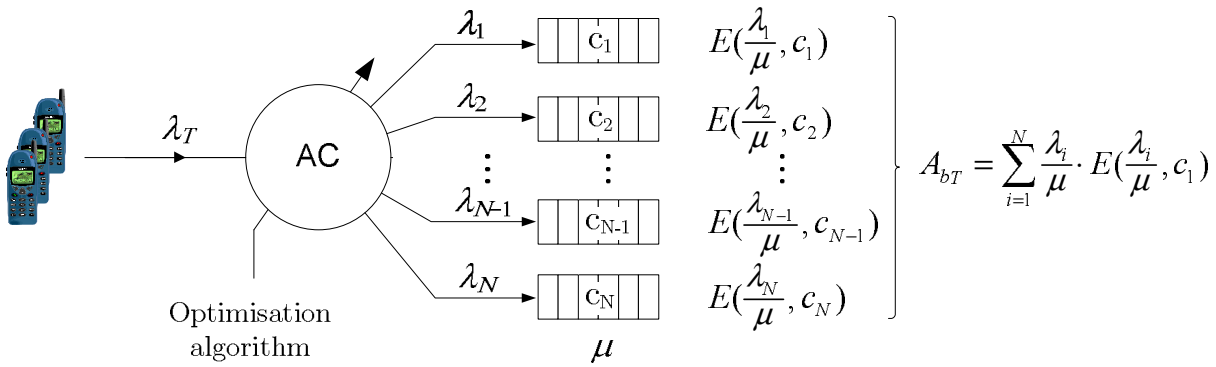


Figure 3.6: A naive model of the traffic sharing problem.

3.3 System Models

This section outlines the problem of determining the optimal spatial traffic distribution when re-planning an existing TDMA/FDMA cellular network. For this purpose, two different network models are presented. A first naive model considers the case of re-distributing traffic demand by the modification of parameters in the call admission control algorithm. Then, a more refined model considers the more realistic case of re-distributing traffic by the modification of cell service areas through handover parameters under constraints on the traffic re-allocation process. For both models, a closed form of the optimality conditions of the traffic sharing problem is derived.

3.3.1 Naive Model

A first network model is presented here, in which a few simplistic considerations are taken into account, and hence the name of *naive* model. The structure of this first model is shown in Figure 3.6. The network consists of a set of N base stations (or cells) serving connection requests from users. These cells are heterogeneous in terms of capacity (i.e., each cell i has a different number of channels, c_i). User demand is modelled by a unique flow of calls that can be freely distributed among cells with absolute freedom, assuming that there is full overlapping between all cells, i.e., the scenario is completely covered by the coverage area of any cell i. The assignment of a user to a cell is performed during connection set-up by the Admission Control (AC) algorithm. The global user flow is then divided in individual cell flows by the admission control. The call arrival process is a time-invariant Poisson process with overall rate λ_T, and so are the cell call rates, λ_i. The service time is an exponentially distributed random variable with parameter $\mu = 1/MCD$, where MCD is the mean call duration. Thus, the total offered traffic in the network is $A_T = \lambda_T/\mu = \lambda_T \cdot MCD$.

The naive model does not consider the handover of a call between cells during the call duration. As a consequence, a) the assignment of a user to a cell by the AC is maintained throughout the call, and b) the channel holding time in a cell coincides with

the call duration and the service rate per channel is identical in all network cells, i.e., $\mu_i = \mu = 1/MCD$. Finally, it is assumed that a call attempt is lost if all channels in the cell are busy, and no retrials are considered (i.e., the network is a loss system).

Most of the previous assumptions are widely used in the literature. The considered call arrival and holding time distributions are standard for voice traffic in telecommunication systems, whether fixed, [73], or mobile, [74]. The loss system assumption is applicable to systems with access control and without queueing or retrials (e.g., [16][68][69]). The permanent association of the mobile to the cell where the call is initiated, equivalent to not modelling user mobility, has also been used in many studies (e.g., [74][75]). Such an assumption is reasonable if cell size is large compared to the distance travelled by the user during the call. More debatable is the condition of full overlapping between cells, as will be discussed later in next section.

Under these assumptions, the call blocking probability in a cell is given by the Erlang-B formula

$$E(A_i, c_i) = \frac{\dfrac{{A_i}^{c_i}}{c_i!}}{\displaystyle\sum_{j=1}^{c_i} \frac{{A_i}^{j}}{j!}} , \tag{3.2}$$

where A_i is the offered traffic, and c_i is the number of channels, both for cell i. $E(A_i, c_i)$ indicates both the congestion and call blocking probabilities. A_i can be expressed as

$$A_i = \frac{\lambda_i}{\mu_i} = \frac{\lambda_i}{\mu} , \tag{3.3}$$

because of the assumption of identical service rate for all cells. The total blocked traffic in the network is the sum of blocked traffic in each cell, computed as

$$A_{bT} = \sum_{i=1}^{N} A_{bi} = \sum_{i=1}^{N} A_i \, E(A_i, c_i) , \tag{3.4}$$

i.e., the sum of the products of offered traffic and call blocking probability in each cell.

Most of AC algorithms take the decision of assigning a fresh connection to a cell according to the current state of the system, e.g., a call is assigned to the cell with more free resources at the moment of the decision. In the naive case, the model has been constructed not to analyse RRM performance, but with the aim of finding the best partitioning of traffic demand among cells (i.e., a traffic sharing strategy, so that the total blocked traffic is minimized). Hence, admission control from the network optimisation perspective is used to defining the size of each cell service area. The underlying optimisation problem can be formulated as

$$\text{Minimise} \qquad \sum_{i=1}^{N} A_i \, E(A_i, c_i) \tag{3.5}$$

$$\text{subject to} \qquad \sum_{i=1}^{N} A_i = A_T \, , \tag{3.6}$$

$$A_i \geq 0 \qquad \forall \, i = 1 : N \, . \tag{3.7}$$

(3.5) shows the goal of minimizing the total blocked traffic, (3.6) ensures that the total offered traffic in the network is A_T, and (3.7) ensures that all offered traffic values are non-negative.

In Appendix B, it is shown that the solution to (3.5)-(3.7) is the one satisfying that

$$E(A_i, c_i) + A_i \, \frac{\partial E(A_i, c_i)}{\partial A_i} = E(A_j, c_j) + A_j \, \frac{\partial E(A_j, c_j)}{\partial A_j} \qquad \forall \, i, j = 1 : N \, . \tag{3.8}$$

The previous equation shows that the total blocked traffic is minimised when the indicator $E(A_i, c_i) + A_i \cdot \frac{\partial E(A_i, c_i)}{\partial A_i}$ is the same for all cells. Such an indicator, hereafter referred to as *incremental blocking probability* (IBP), adds a term to the blocking probability, $E(A_i, c_i)$. This conclusion seems contrary to the common practice of equalising network blocking throughout the network. In a homogeneous network, all cells have the same number of channels (i.e., $c_i = c_j$) and, for symmetry reasons, (3.8) has a trivial solution $A_i = A_j$. In these conditions, equalising any traffic indicator leads to the optimal solution. However, when cells have different capacity (i.e., $c_i \neq c_j$), it is proved in Appendix B that balancing blocking probabilities, i.e., first term on both sides of (3.8), does not lead to the optimal solution.

An intuitive interpretation of the IBP can be given by using that, [76],

$$\frac{\partial E(A_i, c_i)}{\partial A_i} = E(A_i, c_i) \left[\frac{c_i}{A_i} - 1 + E(A_i, c_i) \right] . \tag{3.9}$$

Thus,

$$\begin{aligned} IBP(A_i, c_i) &= E(A_i, c_i) + A_i E(A_i, c_i) \left[\frac{c_i}{A_i} - 1 + E(A_i, c_i) \right] \\ &= E(A_i, c_i) \left[1 + c_i - A_i (1 - E(A_i, c_i)) \right] . \end{aligned} \tag{3.10}$$

By noting that $\{c_i - A_i(1 - E(A_i, c_i))\}$ is the average number of free channels in a cell with offered traffic A_i and c_i channels, $N_{fc}(A_i, c_i)$, (3.10) is re-written as

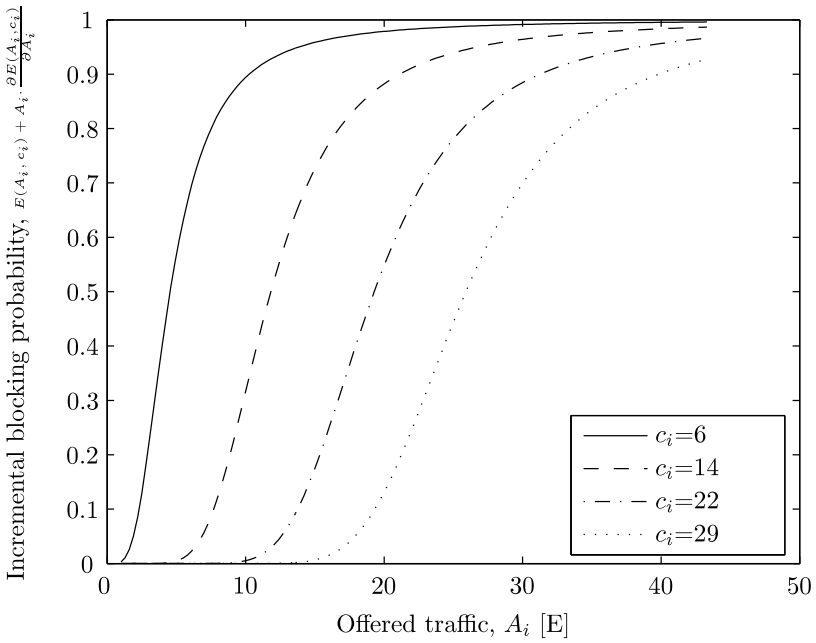


Figure 3.7: Incremental blocking probability with different number of channels.

$$IBP_i = IBP(A_i, c_i) = E(A_i, c_i)\left[1 + N_{fc}(A_i, c_i)\right]. \tag{3.11}$$

Thus, an optimal AC derives fresh connections to different cells with the aim of equalising IBP_i indicators for all cells in the network. Such an equalisation task requires the knowledge of the IBP function behaviour. Figure 3.7 shows the incremental blocking probability in a cell with increasing offered traffic for different number of channels. It is observed that the IBP is a non-decreasing function of the offered traffic, i.e., if two cells experience different IBP values, the one with a lower IBP value has also an A_i lower than the optimal value, A_i^*, and, consequently, AC diverts more traffic demand to that cell. From Figure 3.7, it is also clear that, in cells with a different number of channels, the same value of incremental blocking indicator is reached with different values of offered traffic. As an example, $IBP = 3$ for $A_i \approx 3$ and $A_i \approx 10$ in cells with $c_i = 6$ and $c_i = 14$, respectively. It should be pointed out that the values of c_i in the figure reflect the number of traffic channels (i.e., time slots) in a cell with 1, 2, 3 and 4 transceivers in a live GERAN network.

3.3.2 Refined Model

The naive model made important and simplistic assumption, namely that:

a) users can be freely assigned to any cell in the network, regardless of user and cell locations,

b) the assignment of users to cells is performed during call set-up and not modified later (HO is not considered), and

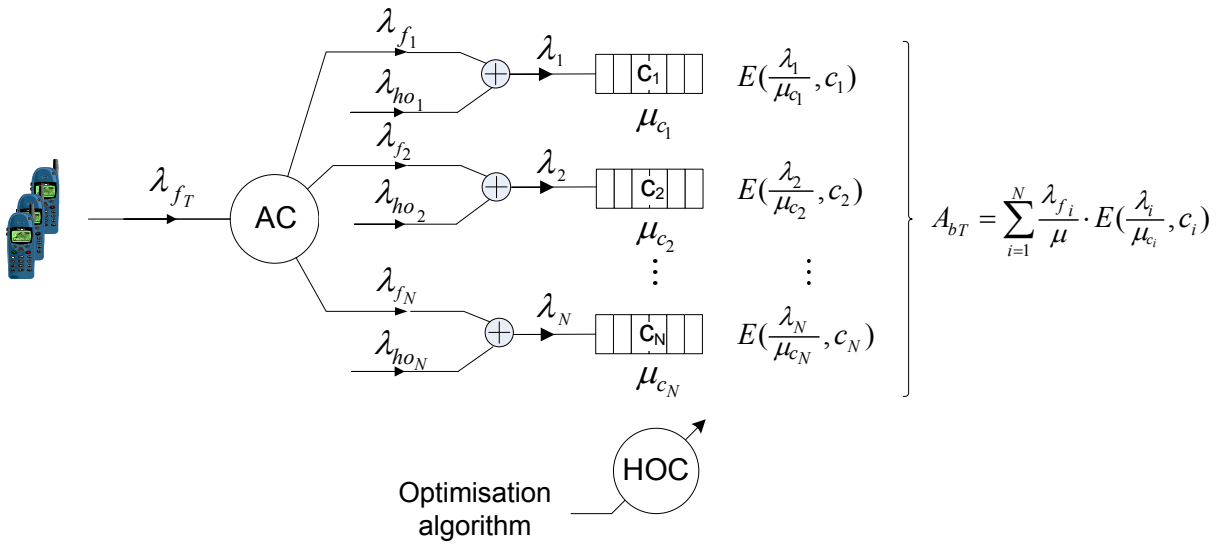


Figure 3.8: A refined model of the traffic sharing problem.

c) all cells can provide adequate coverage during the entire call.

Under these assumptions, the desired balancing effect only relies on the AC procedure. Such a model, albeit intuitive and manageable for a first analysis, is unable to capture two key issues in the cellular environment: user mobility and limited cell coverage. As said in Section 3.2.2, a load balancing strategy should not rely on only AC but also on HO. Hence, the service area of a cell must be controlled by tuning HOC (instead of AC) parameters.

Figure 3.8 presents the refined model, which is based on more realistic assumptions. First, the refined model considers the existence of handovers by modelling a call as a series of connections to several cells, [16]. The total flow of connection requests in a cell is assumed to be the addition of fresh calls being connected to that cell and incoming handover connections from neighbour cells. This total flow is modelled as a Poisson process of rate $\lambda_i = \lambda_{f_i} + \lambda_{ho_i}$, where λ_{f_i} and λ_{ho_i} are the arrival rates of fresh calls and handover requests, respectively, in cell i. The Channel Holding Time (CHT) in a cell, whether for a new call or a handover, is a random variable that can be modelled by a negative exponential distribution with parameter $\mu_c = 1/CHT$. Unlike the naive model, a call can now be handed over between cells, and thus CHT is generally smaller than the mean call duration, MCD. All these are common assumptions in classical models that explicitly consider user mobility (e.g., [16][77][78][79][80]). Although more accurate distributions have been proposed for cell dwell times, channel holding times and handoff inter-arrival times, the exponential assumption is a good approximation, [81][82].

Figure 3.9 illustrates the concepts of call duration and channel holding times. A user starts at cell 1 and travels along four cells before the call is over. Four different CHT_i are obtained, corresponding to the time the user has been attached to cell i. It is worth noting that the channel holding time does not only depend on user mobility (i.e., which cells the user visits and how long stays), but is also influenced by the cell service area defined by HOC settings. Decreasing cell service area by forcing handovers to adjacent

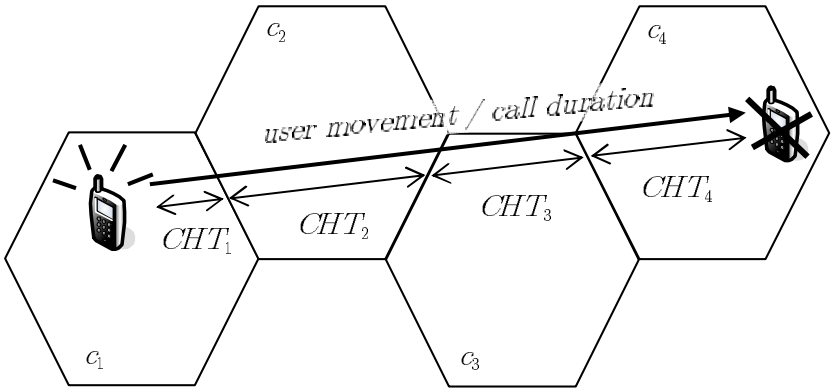


Figure 3.9: Call duration and channel holding time during a user call.

cells leads to a reduction in the channel holding time in the source cell. A smaller cell 1 leads to a wider service area for cell 2 in Figure 3.9, and, consequently, CHT_1 decreases and CHT_2 increases. In the refined model, the statistical distribution of CHT in each cell i is reflected through the service rate parameter, μ_{c_i}. Note that a reduction in service area in cell i not only causes a lower CHT_i, but also leads to an increase of outgoing handover calls, and, consequently, an increase of λ_{ho_i} in neighbouring cells.

Traffic sharing can also be performed by tuning HOC parameters modifying service areas. It is assumed that a call is not dropped when a handover request is blocked, which is reasonable for urban scenarios, where low user mobility, high cell overlapping and few coverage problems exist. Therefore, the aim of tuning is to minimise blocking of fresh calls, which is assumed to be the only source of lost traffic. Such a goal is achieved by adjusting λ_{ho} and μ_c on a per-cell basis to ensure the values of $A_i = \lambda_i/\mu_{ci} = (\lambda_{f_i} + \lambda_{hoi})/\mu_{ci}$ that minimise

$$A_{bT} = \sum_{i=1}^{N} A_{bi} = \sum_{i=1}^{N} \frac{\lambda_{fi}}{\mu} E(A_i, c_i) = \sum_{i=1}^{N} \frac{\lambda_{fi}}{\mu} E(\frac{\lambda_i}{\mu_{c_i}}, c_i) \tag{3.12}$$

(i.e., the total blocked traffic due to the rejection of fresh calls of average duration $1/\mu$).

Hitherto, it has been assumed that users can be freely assigned to network cells. In a live network, a user can only be assigned to cells providing adequate coverage where the call is originated. This fact limits the minimum and maximum offered traffic that can be assigned to a cell, i.e., A_i is generally lower than A_T and higher than 0. A lower bound in cell i is associated to connections in the area where the cell i is the only one providing adequate coverage, so that traffic can only be carried through that cell. An upper bound in cell i corresponds to connections that fall within the coverage area of the cell, and no more traffic could be carried by that cell i. These bounds limit the capability of sharing traffic, causing that the optimal solution to the unconstrained problem cannot be reached. Therefore, the new traffic sharing problem can be formulated as the constrained optimisation problem

$$\text{Minimise} \qquad \sum_{i=1}^{N} \frac{\lambda_{f_i}}{\mu} E(A_i, c_i) \quad \text{or} \quad \sum_{i=1}^{N} \lambda_{f_i} E(A_i, c_i) \tag{3.13}$$

$$\text{subject to} \qquad \sum_{i=1}^{N} A_{fi} \left(1 - E(A_i, c_i)\right) = \sum_{i=1}^{N} A_i \left(1 - E(A_i, c_i)\right), \tag{3.14}$$

$$A_{lbi} \leq A_i \leq A_{ubi} \qquad \forall\, i = 1 : N, \tag{3.15}$$

where A_{lbi} and A_{ubi} are lower and upper bounds on cell traffic due to spatial concentration of traffic demand. Briefly, (3.13) reflects the goal of minimising the total blocked traffic or the total blocking rate. Both goals lead to the same optimal solution since μ is a constant. As previously stated, it is assumed that a call is not dropped when a handover request is blocked. (3.14) formulates such an assumption, ensuring that the total sum of traffic accepted by the system equals to the sum of carried traffic in cells (i.e., non-carried traffic in the system is only coming from the blocking of fresh calls). Finally, (3.15) describes the limits due to spatial concentration of traffic demand. Note that fresh call arrival rates, λ_{f_i}, are not affected by tuning HOC, as they only depend on the AC. Hence, the decision variables are the offered traffic per cell, A_i. The latter are controlled by changes in the average connection service rates, μ_{ci}, and incoming handover rates, λ_{hoi}, caused by tuning HOC parameters. Also note that, unlike in the naive model, in the refined model, the traffic entering the system through a particular cell does not necessarily coincide with the carried traffic in the cell due to the traffic balancing mechanism.

In Appendix B, it is shown that the solution to (3.13)-(3.15) satisfies that the value of β, defined as

$$\beta_i = \beta(A_i, A_{fi}, c_i) = \frac{A_{f_i} \frac{\partial E(A_i, c_i)}{\partial A_i}}{1 - E(A_i, c_i) + (A_{f_i} - A_i) \frac{\partial E(A_i, c_i)}{\partial A_i}}, \tag{3.16}$$

is the same $\forall\, i$. More precisely, the optimal solution is the one satisfying that

$$\beta_i = \beta(A_i, A_{fi}, c_i) = \beta(A_j, A_{fj}, c_j) = \beta_j, \tag{3.17}$$

for all cells i, j where constraint (3.15) is inactive[1], and

$$\beta(A_u, A_{fu}, c_u)|_{A_u = A_{ub}} \leq \beta_i \qquad \text{or} \tag{3.18}$$

$$\beta_i \leq \beta(A_l, A_{fl}, c_l)|_{A_l = A_{lb}} \tag{3.19}$$

for all cells l and u where constraint (3.15) is active due to the lower or upper bound, respectively. Basically, (3.17) shows that, in the absence of traffic constraints, the best

[1] An inequality constraint is *inactive* (or *not binding*) when the equality does not hold.

performance in the refined model is achieved by equalising the indicator β across the network. Likewise, (3.18)-(3.19) suggest that, in those cells where one of the traffic bounds is reached, traffic has to be fixed to the limit value and the traffic excess (or defect) must be re-distributed among the remaining cells. This fact justifies that sharing the traffic between adjacent cells leads to the optimal solution even in the presence of constraints on the offered traffic per cell.

3.4 Model Performance Assesment

In the previous section, the optimal traffic sharing criteria for two teletraffic models of a cellular network have been presented. The following experiments quantify the benefit of the exact approach when compared to common heuristic approaches. For clarity, the analysis set-up is first introduced and results are then presented.

3.4.1 Analysis Set-up

Assessment is carried out over four test scenarios of increasing complexity. Each new scenario adds a new feature, so the impact of such an addition can be observed. The first three scenarios consist of 3 GERAN cells of uneven capacity. In the example, the number of channels per cell is $c_i = 29$, $c_2 = 6$ and $c_3 = 6$, corresponding to 4, 1 and 1 transceivers, respectively. More specifically,

- *Scenario 1* considers the naive model, i.e., static users, full cell overlapping and, consequently, no constraints on traffic sharing.
- *Scenario 2* considers the refined model with no constraints on traffic sharing, where user mobility is taken into account, but full cell overlapping is still assumed (i.e., A_{lb_i}=0, A_{ub_i}=∞). Thus, the impact of the introduction of user mobility can be quantified in this scenario.
- *Scenario 3* considers the refined model with limits to the cell offered traffic to evaluate the impact of limited cell coverage and spatial concentration of traffic demand. This scenario still consists of simple 3-cell scenario still remains.
- *Scenario 4* extends the analysis to a real case built from data taken from a live network. This scenario corresponds to the cells served by a real base station controller. Unlike previous scenarios, much more than three cells are analysed, and cell traffic bounds, A_{lb_i} and A_{ub_i}, are computed on a cell-by-cell basis from geometric considerations, as will be explained later.

Four traffic sharing strategies are tested for each scenario. All methods aim to equalise some performance indicator across the network, differing in the particular indicator balanced. The first three are heuristic methods that equalise the average traffic load, L_i, the

Sharing strategy	Acronym	Indicator	Formula	Type
Load Balancing	LB	L_i	$= \frac{A_i(1-E(A_i,c_i))}{c_i}$	Heuristic
Blocking Prob. Balancing	BPB	$E(A_i, c_i)$	Eq. (3.2)	Heuristic
Blocked Traffic Balancing	BTB	A_{b_i}	$= A_{fi} \cdot E(A_i, c_i)$	Heuristic
Optimal Balance	OB	IBP_i	Eq. (3.11)	Optimal (naive)
		β_i	Eq. (3.16)	Optimal (refined)

Table 3.1: Definition of traffic sharing strategies.

blocking probability, $E(A_i, c_i)$, or the blocked traffic, A_{b_i}, respectively. These methods are hereafter referred to as *Load Balancing* (LB), *Blocking Probability Balancing* (BPB) and *Blocked Traffic Balancing* (BTB), respectively.

Table 3.1 details each strategy through five columns. The names and acronyms of the different strategies are detailed in first and second column, respectively. Third column depicts the performance indicator to balance, and fourth column includes the mathematical definition (i.e., the formula) of such indicators. Finally, fifth column indicates if each method is labelled as heuristic or optimal. While LB is used by most traffic balancing algorithms for real-time purposes, [59][83], BPB is often used by network operators when optimising their networks, [12][55]. The fourth method, referred to as *Optimal Balancing* (OB), considers the optimal sharing criterion in each model (i.e., (3.8) for the naive model and (3.17)-(3.19) for the refined model).

To evaluate network performance, the traffic share among cells is computed for each strategy. For optimal sharing (i.e., OB), this is performed by solving (3.5)-(3.7) and (3.13)-(3.15) analytically. For heuristic strategies (i.e., LB, BPB and BTB), the balancing problem is formulated as the non-linear least squares problem

$$\text{Minimise} \quad \sum_{i=1}^{N-1} \left(\mathcal{I}\left(A_i, c_i\right) - \mathcal{I}\left(A_{i+1}, c_{i+1}\right)\right)^2 \tag{3.20}$$

$$\text{subject to} \quad \sum_{i=1}^{N} A_i = A_T, \tag{3.21}$$

for the naive model and

$$\text{Minimise} \quad \sum_{i=1}^{N-1} \left(\mathcal{I}\left(A_i, c_i\right) - \mathcal{I}\left(A_{i+1}, c_{i+1}\right)\right)^2 \tag{3.22}$$

$$\text{subject to} \quad \sum_{i=1}^{N} A_{fi}(1 - E(A_i, c_i)) = \sum_{i=1}^{N} A_i(1 - E(A_i, c_i)) \ , \tag{3.23}$$

$$A_{lbi} \leq A_i \leq A_{ubi} \quad \forall\, i = 1 : N, \tag{3.24}$$

for the refined model, where $\mathcal{I}(A_i, c_i)$ is the value of the balanced indicator in cell i (i.e., average traffic load, blocking probability or blocked traffic, third column in Table 3.1), expressed as a function of A_i and c_i. Optimisation models are solved by the *fsolve* and *fmincon* functions in MATLAB Optimisation Toolbox, [46]. When possible, the Jacobian matrix is provided to the scripts to speed up computations.

Traffic sharing strategies have the final goal of minimising the global blocked traffic in operator's network. Thus, the total blocked traffic, A_{bT}, and the overall blocking rate, $\frac{A_{bT}}{A_T}$, are the main figures of merit for assessment. Two additional figures are also used for a better comparison between load sharing strategies. First, a measure of network capacity for each strategy is computed as the total offered traffic in the scenario for an overall Grade of Service (GoS) of 2% (i.e., a global network blocking probability of 2%). To quantify the loss of network capacity of not implementing optimal sharing, maximum traffic values are normalised by that of OB to give a relative capacity figure

$$C_m = \frac{A_T|_{GoS=2\%,\, m}}{A_T|_{GoS=2\%,\, OB}} \ , \tag{3.25}$$

where $A_T|_{GoS=2\%,\, m}$ is the network capacity obtained by method m for GoS of 2%. Note that $m \in \{LB, BPB, BTB, OB\}$ and $C_{OB} = 1$, i.e., OB method does not experience loss of network capacity comparing with itself.

Second, to quantify the impact of constraints on traffic sharing strategies, capacity values in the constrained case are normalised by that of OB in the unconstrained case, as

$$C_{m,const} = \frac{A_T|_{GoS=2\%,\, m,\, const}}{A_T|_{GoS=2\%,\, OB,\, unconst}} \ , \tag{3.26}$$

where $A_T|_{GoS=2\%,m,const}$ is the network capacity obtained by method m with spatial constraints on the offered traffic per cell.

3.4.2 Results

Scenario 1

The first scenario considers the naive model with 3 cells of uneven capacity. A first experiment evaluates the performance of traffic sharing strategies for a fixed value of total offered traffic. Specifically, A_T=30 E(rlang), which leads to an overall offered traffic load

$$\rho = \frac{A_T}{c_1 + c_2 + c_3} = \frac{30}{29 + 6 + 6} = 0.73 \quad . \tag{3.27}$$

Table 3.2 presents the results of the different strategies in separate columns. Each row in the table presents the value of a teletraffic indicator. From top to bottom, the rows show total offered traffic (A_T), offered traffic ($\mathbf{A}$), average traffic load ($\mathbf{L}$), blocking probability ($\mathbf{E}$), blocked traffic ($\mathbf{A_b} = \mathbf{A} \cdot \mathbf{E}$), incremental blocking probability ($\mathbf{IBP} = \mathbf{E} + \mathbf{A} \cdot \nabla\mathbf{E}$) and total blocked traffic (A_{bT}). Indicators in bold are represented by vectors with the values in the three cells. Obviously, the second and third components in every vector have the same value, as those cells have the same capacity in this scenario (i.e., $c_2 = c_3 = 6$). Likewise, all three cells show the same value of the indicator to equalise in each strategy (i.e., $\mathbf{L}$ in LB, $\mathbf{E}$ in BPB, $\mathbf{A_b}$ in BTB and $\mathbf{IBP}$ in OB). For clarity, the equalised indicator in each strategy is highlighted in grey.

From the table, it is clear that the minimum total blocked traffic (i.e., A_{bT}, last row) is obtained by equalising the incremental blocking probability (i.e., OB method, last column), resulting in $A_{bT} = 1.486$. As expected, OB performs the best for this value of traffic demand. Nonetheless, it is observed that large imbalances of the latter IBP indicator still give adequate blocking performance. For instance, the BPB method (3^{rd} column) causes a reduction in the incremental blocking probability (7^{th} row) in cells 2 and 3 of 43% respecting to cell 1, i.e., [0.38 0.22 0.22]. Despite this imbalance, A_{bT} only increases by 5.5% compared to the optimal strategy, i.e., 1.567 versus 1.486. In contrast, a 50% increase of blocked traffic is obtained by equalising the average traffic load against OB, i.e., 2.224 versus 1.486. It is worth mentioning that equalising the blocked traffic in cells is worse than equalising the blocking probability in terms of total blocked traffic.

These results have been extracted for a fixed total offered traffic, A_T=30 E. However, the previous conclusions are still valid, regardless of the total offered traffic, A_T. A second experiment, illustrated by Figure 3.10, shows the evolution of the overall blocking rate with total offered traffic for all strategies. As expected, OB obtains the minimum blocked traffic (and, consequently, the maximum carried traffic) for all values of traffic demand. BPB and BTB achieve nearly the same blocking as OB, whereas LB performs much worse.

From Figure 3.10, the total offered traffic for an overall blocking rate of 2% in each strategy can be easily be found. This value is used as a measure of network capacity. Analysis shows that, in this scenario, the network capacity of BTB, BPB and LB relative

Method	LB	BPB	BTB	OB
[Balancing Criterion]	$[L_i]$	$[E(A_i, c_i)]$	$[A_i \cdot E(A_i, c_i)]$	$[E(A_i, c_i)+A_i \frac{\partial E(A_i, c_i)}{\partial A_i}]$
A_T [E]		30		
A [E]	[19.88 5.06 5.06]	[24 3 3]	[21.64 4.18 4.18]	[23 3.50 3.50]
L	[0.67 0.67 0.67]	[0.78 0.47 0.47]	[0.73 0.61 0.61]	[0.76 0.54 0.54]
E [%]	[1.21 19.61 19.61]	[5.22 5.22 5.22]	[2.52 13.03 13.03]	[3.95 8.24 8.24]
$\mathbf{A_b}$ [E]	[0.24 0.99 0,99]	[1.25 0.16 0,16]	[0.54 0.54 0.54]	[0.91 0.29 0.29]
IBP	[0.13 0.58 0.58]	[0.38 0.22 0.22]	[0.22 0.44 0.44]	[0.31 0.31 0.31]
A_{bT} [E]	2.224	1.567	1.634	1.486

Table 3.2: Results of traffic sharing strategies in Scenario 1.

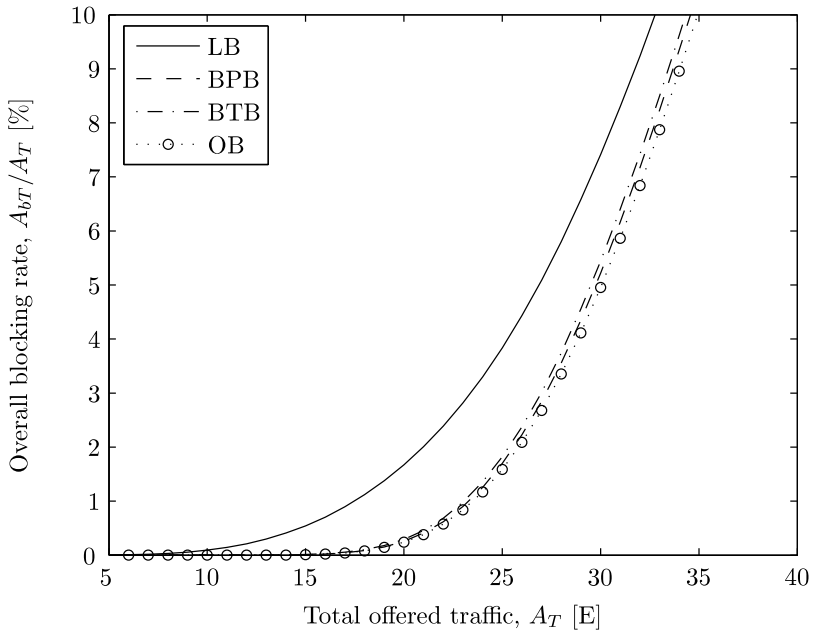


Figure 3.10: Overall blocking rate for different traffic sharing strategies in Scenario 1.

to OB is $C_{BTB} = 0.99$, $C_{BPB} = 0.98$ and $C_{LB} = 0.81$, respectively. These results confirm that, in the naive model, BTB and BPB traffic sharing strategies give near-optimal performance (1% less capacity than OB method), while LB performs much worse. A closer analysis shows that the gradient of the objective function in (3.5) in the direction imposed by constraint (3.6) is small near the optimum. Thus, significant changes in the traffic distribution cause limited performance differences.

Scenario 2

In contrast to Scenario 1, this second scenario considers the refined model, where user mobility is introduced. Cell resizing is performed by changing HO parameters in the HOC. Thus, the HOC can freely define the offered traffic demand to each cell, A_i, given that the fresh call arrival rate in each cell, λ_{f_i}, is fixed. For simplicity, a uniform spatial user distribution is assumed, i.e., $\lambda_{f_i} = \lambda_T/N$. Full cell overlapping is still assumed (i.e., A_{lb_i}=0, A_{ub_i}=∞). Thus, Scenario 2 evaluates the impact of the introduction of user mobility.

Figure 3.11 shows the overall blocking rate of the different strategies with increasing offered traffic in the new scenario. In the figure, it is observed that OB performs the best while LB still performs the worst. BPB is still the best heuristic method, and, more importantly, larger performance differences are now observed between methods. OB method is now significantly better than heuristic strategies. Specifically, the relative network capacity for BPB, BTB and LB is now 0.967, 0.967 and 0.65, respectively. In other words, the improvement of network capacity achieved by OB is 3.3% compared to the best heuristic method and 35% compared to LB. It can be concluded that, in the refined model (i.e., when user mobility is introduced), the benefit of the optimal

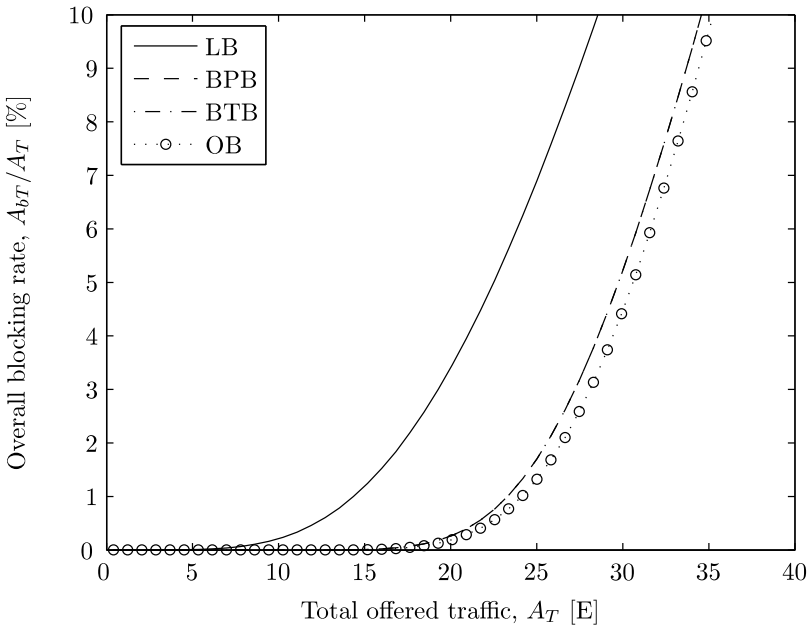


Figure 3.11: Overall blocking rate for different traffic sharing strategies in Scenario 2.

approach becomes more evident. It is also important to note that, under the assumption of uniform spatial distribution (i.e., $A_{f_i} = A_{f_j}$), BPB and BTB lead to the same solution (both curves coincide in Figure 3.11). It happens due to their balance conditions (i.e., $E(A_i, c_i) = E(A_j, c_j)$ and $A_{fi}E(A_i, c_i) = A_{fj}E(A_j, c_j)$, respectively) becomes the same and, consequently, their offered traffic solutions are identical.

Scenario 3

In previous scenarios, it has been assumed that the optimal traffic share can always be reached by tuning HOC parameters, so any offered traffic in any cell, A_i, is possible. However, this is not true in actual networks, where not all users can be handed over to any cell (e.g., the HOC might try to send a user to another cell for load sharing, but the target cell does not reach the mobile). To account for this limitation, in the third scenario, lower and upper bounds, A_{lbi} and A_{ubi}, are included on the offered traffic in cells. Such constraints reduce the feasible solution space, causing that the optimal solution to the unconstrained problem might not be reached.

To evaluate the impact of constraints on methods, bounds are introduced in the optimisation problem, (3.15) and (3.24) for optimal and heuristic strategies, respectively. For clarity, these bounds are gradually relaxed in the scenario. For simplicity, bounds for all cells are controlled by a single parameter, Δ, referred to as *deviation parameter*. This parameter is introduced in the definition of the lower and upper offered traffic bounds as

$$A_{lb_i} = (1 - \Delta)A_i^{(0)} , \tag{3.28}$$

$$A_{ub_i} = (1 + \Delta)A_i^{(0)} , \tag{3.29}$$

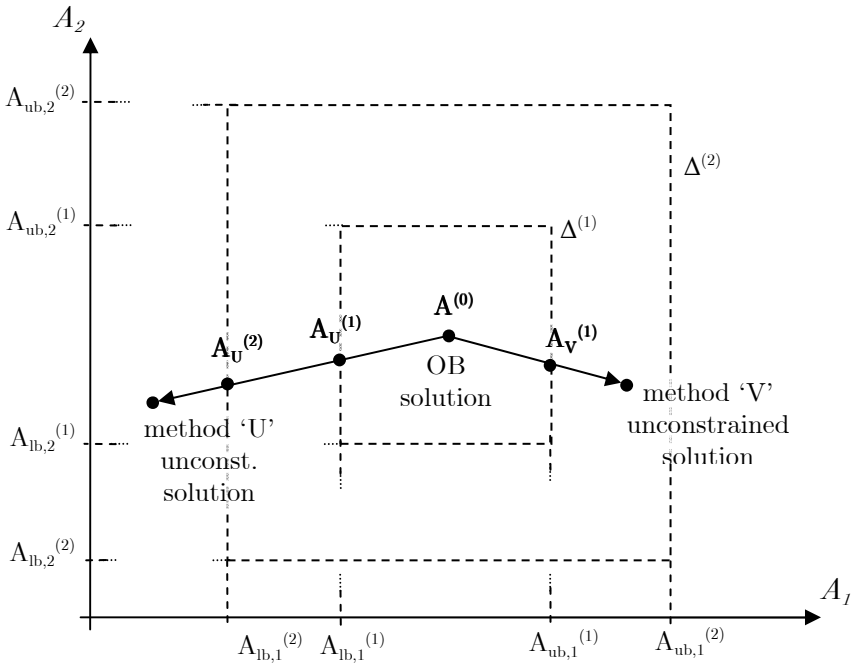


Figure 3.12: Example of the evolution of solutions in a constrained solution space.

where $A_i^{(0)}$ is the offered traffic in cell i in a reference solution, $\mathbf{A}^{(0)}$. For instance, $\Delta \to \infty$ means that no constraints are considered, while $\Delta \to 0$ causes that all strategies lead to the same solution, $A_i = A_i^{(0)}$. Higher values of Δ enlarge the solution space for offered traffic and, then, a better solution can be found. For simplicity, it is again assumed that fresh calls are uniformly distributed in the scenario, i.e., $\lambda_{f_i} = \lambda_T/N$, and the reference solution, $\mathbf{A}^{(0)}$, is the optimal traffic distribution in the unconstrained case. Consequently, the solution space always includes the optimal point (i.e., the best network performance). Figure 3.12 illustrates an example of how the solution of two methods (method 'U' and 'V') evolves in a constrained scenario when constraints are changed. For simplicity, a two cell scenario is depicted, where offered traffic is a 2-D vector, i.e., $\mathbf{A}=[A_1\ A_2]$. Superscripts indicate consecutive step in Δ values, e.g., $\mathbf{A_V^{(1)}}$ defines the offered traffic solution for method 'V' when $\Delta^{(1)}$ value is applied to constraints, (3.28)-(3.29). Each Δ value defines an upper and lower bound in the figure, and so is the squared solution space. $\Delta^{(0)}$=0 forces one point as the only feasible solution, i.e., $\mathbf{A^{(0)}}$.

Figure 3.13 shows the network capacity of methods compared to that of OB without constraints, $C_{m,const}$, as traffic constraints become looser. It is observed that, for $\Delta \approx 0$ (i.e., tight constraints), all methods lead to the same solution, $\mathbf{A}^{(0)}$, and therefore have the same performance (i.e., $C_{LB,const} = C_{BPB,const} = C_{BTB,const} = C_{OB,const}$). This shared performance is the optimal, i.e. $C_{m,const}$=1 because the only possible solution point, $\mathbf{A}^{(0)}$, is defined as the unconstrained optimal point. In contrast, for $\Delta \gg 0$ (i.e., loose constraints), each method leads to a different solution. Specifically, $C_{LB,const} < C_{BTB,const} \leq C_{BPB,const} < C_{OB,const}$. Note that, due to the way bounds are defined in (3.28)-(3.29), the feasible solution space is always centered at the optimal solution in the unconstrained case. Therefore, the optimal solution to the constrained problem is always the same as for the unconstrained case, and $C_{OB,const}$=1 regardless of the value of Δ. As Δ increases,

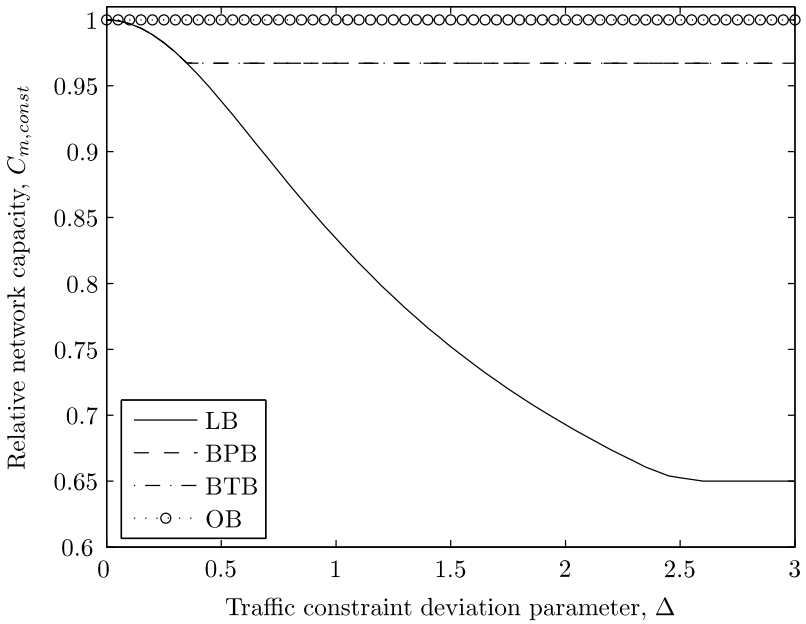


Figure 3.13: Evolution of relative network capacity achieved by methods with varying constraints.

traffic constraints become inactive and heuristic methods can reach their own balance conditions. Once each heuristic method reaches its balance condition, further increments of Δ do not have any influence (i.e., $\Delta > 0.3$ for BTB and BPB, and $\Delta > 2.6$ for LB). Note that, when constraints do not apply, the constrained solution become identical to Scenario 2 and, then, $C_{LB,const}$=0.65 and $C_{BTB,const}$=$C_{BPB,const}$=0.987, respectively.

Similarly to Scenario 2, due to uniform user spatial distribution ($\lambda_{f_i} = \lambda_T/N$), BTB and BPB methods show identical behaviour. From the Figure 3.13, it can be inferred that the LB solution is the one differing the most from the optimal distribution, as it needs the largest value of Δ to reach a constant capacity value. From the value of Δ, it can be deduced that the offered traffic in at least one cell in LB differs by 360% (=100·(1+2.6)) from that in the optimal traffic distribution.

Scenario 4

The last scenario corresponds to the geographical area served by a base station controller in a live GERAN network. The area consists of a mixture of omnidirectional and sectorised sites distributed over 579 km^2, comprising 117 cells and 313 transceivers. The dataset consists of site coordinates, antenna bearings, number of channels per cell, and number of fresh call attempts and carried traffic per cell during the BH of each 10 days. The number of channels per cell in the scenario varies from 6 to 44 (i.e. from 1 to 6 transceivers), and so do the offered and carried traffic per cell.

The main differences with previous scenarios are a) the uneven spatial distribution of users (and, consequently, of fresh calls), and, b) the consideration of uneven traffic bounds in cells. Thus, the spatial distribution of traffic demand due to fresh calls is not defined

anymore. Instead, it is defined relative to the total traffic demand in the scenario and maintained when increasing the latter to estimate network capacity. Thus,

$$\lambda_{f_i} = \mu \, A_T \, r_i \, , \tag{3.30}$$

where r_i is the ratio of traffic demand in cell i compared to the global traffic demand, which is derived from network measurements, and μ is also known from live measurements. Bounds on offered traffic are calculated on a cell basis based on geometric considerations. As already mentioned, the upper bound, A_{ub_i}, is the offered traffic in the coverage area of cell i, and, therefore, no more traffic can be carried by that cell. The lower bound, A_{lb_i}, is the offered traffic in the area only covered by cell i, and, then, no less traffic can carried by that cell. In this work, the coverage area of a cell is defined by a coverage radius, r_{cvg}, and a maximum angle off the antenna bearing, θ_{cvg} (i.e., half of the catchment angle). For simplicity, it is assumed that r_{cvg} is the same for all cells, $\theta_{cvg} = \frac{360^\circ}{2k}$ for k-sectorial sites and $\theta_{cvg} = 180^\circ$ for omnidirectional sites. In the analysis, r_{cvg} ranges from 1 to 20 km.

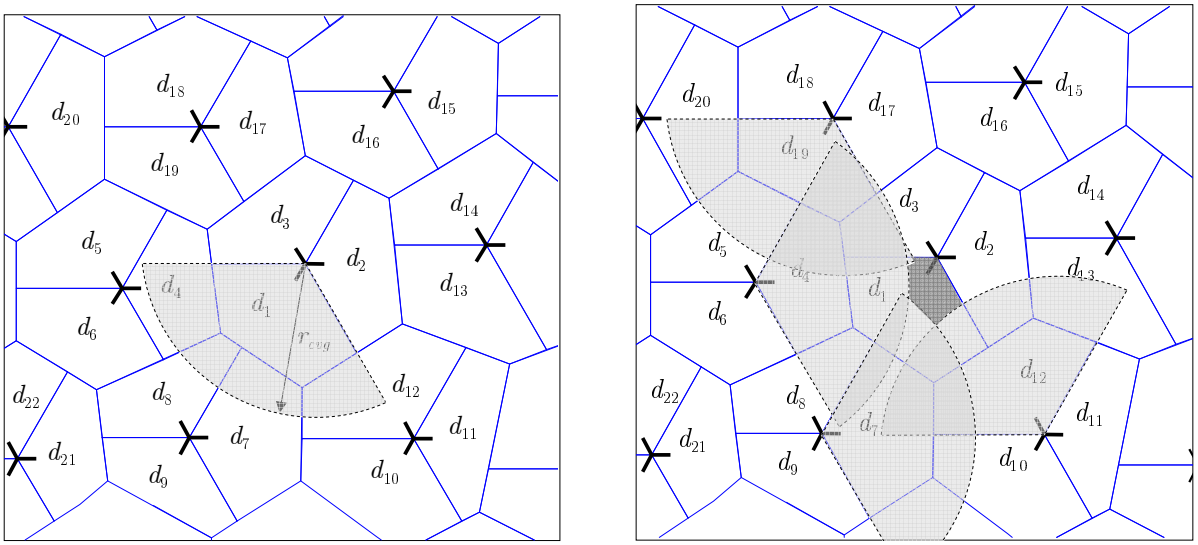
To compute offered traffic bounds, the spatial distribution of fresh traffic is needed. For this purpose, cell service areas are derived from network configuration. First, the dominance area of sites is computed from site coordinates by Voronoi tessellation, [84]. The dominance area of one site comprises all the points that site is the nearest one. For omnidirectional sites, cell service area is the site dominance area. In sectorised sites, cell service area is built by dividing the site dominance area into as many subareas as sectors based on antenna bearings. Finally, the spatial traffic distribution is built by mapping traffic measurements onto polygons representing cell service areas. Figure 3.14 illustrates an example of how traffic bounds are calculated in the scenario. In the figure, site location and antenna bearings are given by a symbol '—', while cell service and coverage areas are represented by solid and dashed lines, respectively. Figure 3.14(a) shows the maximum service area of a cell in light grey. Note that the shaded area coincides with the coverage area of the cell. Figure 3.14(b) shows the minimum service area of the cell in dark grey. From the figures, it can easily be deduced that

$$A_{ub_i} = A_T \left(r_i + \sum_{j \neq i} r_j \, \frac{a_i \cap s_j}{s_j} \right) \, , \tag{3.31}$$

$$A_{lb_i} = A_T \, r_i \left(s_i - \left(s_i \cap \left(\bigcup_{j \neq i} a_j \right) \right) \right) \, , \tag{3.32}$$

where a_i is the coverage area of cell i, s_i is the service area of cell i, and '$\cup$' and '$\cap$' operators are the union and intersection of areas. Note that both traffic bounds are defined relative to the total offered traffic in the scenario, A_T.

As in other scenarios, network capacity for each strategy is computed by gradually increasing A_T until GoS exceeds 2%. For each value of A_T, the fresh call arrival rate per



(a) Maximum service area (b) Minimum service area (dark grey)

Figure 3.14: Limits of cell service areas.

cell is calculated as in (3.30). Then, for each value of r_{cvg}, cell traffic bounds are computed as in (3.31)-(3.32). Figure 3.15 shows the relative capacity of methods, $C_{m,const}$, as r_{cvg} increases in the real scenario. In the figure, it is observed that OB is always the best method. For $r_{cvg} = 1\,\text{km}$, all methods perform the same due to the tight constraints. For the more realistic value of $r_{cvg} = 5\,\text{km}$, $C_{OB,const} = 0.925$, $C_{LB,const} = 0.912$, $C_{BPB,const} = 0.904$ and $C_{BTB,const} = 0.859$. Hence, in a real scenario, the optimal network capacity only decreases by 7.5% when considering traffic constraints. Even with these constraints, network capacity is increased by 2% when using OB instead of BPB. Similar results are obtained for larger values of cell coverage radius. Unexpectedly, LB performs better than BPB and BTB for small values of r_{cvg}. A more detailed analysis shows that, with $r_{cvg} < 5\,\text{km}$, the LB solution is close to the OB solution in this particular scenario. As r_{cvg} increases, LB and BPB become the worst methods, and BTB approaches to OB. Specifically, for $r_{cvg} \to \infty$, $C_{OB,const} = 1$, $C_{BTB,const} = 0.997$, $C_{LB,const} = 0.977$ and $C_{BPB,const} = 0.971$. Note that, in this case, BPB is the worst method.

Sensitivity analysis

Obviously, performance figures in Scenario 4 might vary depending on the spatial distribution of users and channels per cell (i.e., r_i and c_i). To quantify the impact of varying these parameters, and how representative are the capacity gain values obtained in the previous section, a sensitivity analysis has been carried out in the real scenario following a Montecarlo method. Thus, 100 realisations of user and channel distribution were generated by randomly selecting values for r_i and c_i, while ensuring that $\sum r_i = 1$ in each sample and maintaining the total number of channels in the scenario. As a result, the capacity gain of OB versus BPB varied from 2% to 21%, averaging 10%. This result shows that the value of 2% reported above for the real scenario can be considered a conservative value.

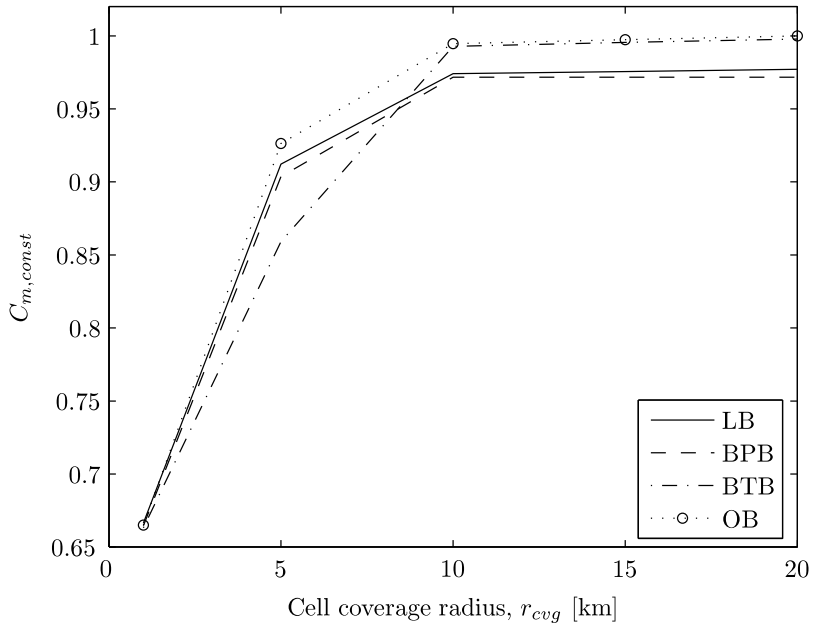


Figure 3.15: Relative capacity of methods in the real scenario.

3.5 Conclusions

In this chapter, the problem of finding the best traffic share between cells in an existing GERAN system has been studied. Two analytical teletraffic models of the network have been presented, as a result of different assumptions. Both models consider the network as a loss system consisting of multiple cells, but differ in the mechanism used to re-distribute traffic: call access control or handover. In both models, a closed-form expression for the optimality criterion has been derived, defining an optimal load balancing strategy. In addition to this optimal criterion, other three heuristic load sharing strategies have been defined. These four methods implement load sharing by equalising different network indicators (cell traffic load, call blocking probability, blocked traffic and the optimal indicator previously obtained). In all strategies, offered traffic on a cell basis is obtained by solving a classical optimisation problem for each load sharing approach.

Preliminary analysis shows that equalising blocking rates across the network, as operators currently do, is not the optimal strategy. A comprehensive analysis in different scenarios has shown:

a) Small differences in capacity gain between load sharing methods are encountered when the call access control is used as the load balancing mechanism if no user mobility is considered. Capacity gain increases when handover is used as the mechanism for load balancing.

b) Spatial constraints restrict the feasible solution space for all the methods. In a tightly constrained scenario most methods tend to perform similarly. As constraints get looser, each method performs differently.

c) Using the optimal sharing criterion instead of balancing blocking rates can increase

network capacity in a realistic scenario by 3%.

Such a 3% figure, albeit small, is not negligible in terms of operator revenues. More important, the benefit is obtained without changing network equipment, which is key in mature technologies such as GERAN. An additional sensitivity analysis has been performed changing the spatial distribution of users and channels per cell, following a Montecarlo approach. Such an analysis demonstrate that the 3% capacity gain can be seen as a conservative value.

The selection of a proper time scale for re-allocating traffic demand is an important issue. Any traffic sharing method based on the optimal criterion relies on robust estimates of traffic indicators, such as the average fresh offered and carried traffic. Time scale must therefore be large enough to get reliable measurements (i.e., hours). Thus, the proposed criterion is conceived to tune handover parameters by the network management system. Such a parameter tuning can only be performed at most on an hourly basis in current networks. It should be pointed out that an hourly measurement might not be valid for the following hours due to traffic fluctuations in a day. This drawback can be circumvented by tuning network parameters based on measurements at the same hour of the previous days, as in [55].

Chapter 4

Self-tuning of Inter-System Handover Parameters in Multi-Radio Access Networks

In this chapter, an auto-tuning scheme is proposed to adjust parameters in a standardised Inter-System HandOver (IS-HO) algorithm for traffic sharing in a heterogeneous network, including GSM and UMTS radio access technologies. The proposed auto-tuning scheme is presented first, based on a Fuzzy Logic Controller (FLC) modifying IS-HO parameters. A simulation platform has been built for the proposal assessment, including the main intra- and inter-technology functionalities. Results analyse the auto-tuning scheme capability for adapting the simulated network in a traffic changing scenario. Finally, a sensitivity analysis is then performed to quantify the effect of the internal FLC configuration on the network adaptation speed.

4.1 Introduction

Wireless communication networks are rapidly increasing in complexity due to the introduction of new services and technologies. As a result, many different and new Radio Access Technologies (RATs) can co-exist in the same geographical area. In such a complex scenario, a bunch of services is offered by different radio access networks, often owned by the same operator, and new promising technologies live with mature networks in continuous evolution.

Operators usually try to work jointly with these RATs seamlessly for the user. The user has a multi-RAT terminal and the User Equipment (UE) connects each service to each RAT transparently, [25]. New organisational tasks arise and new network entities have to take charge of them. That is the case of the JRRM entity. JRRM deals with

radio resources from every technology as a whole and acts coordinating the decisions for every RAT. Within the JRRM entity, Joint Admission Control (JAC) and IS-HO are the most representative algorithms. JAC and IS-HO in a heterogeneous scenario are the counterparts of AC and HO algorithms in single-RAT networks, i.e., JAC decides to which RAT a new service is assigned, and IS-HO decides if an ongoing service must be handed over towards another RAT. These new algorithms work in this joint scenario, but, at the same time, must manage radio resources, radio measurements and user mobility from every single RAT, where technical specifications are quite different. A joint management must work with very different radio resource from radio technologies, such as time slots in GSM, codes in UMTS or shared channels in WLAN.

In parallel to network evolution, new services are launched, causing a strong change on the properties of traffic demand. New user trends, population flows, cities/transportation deployment, or even economical conditions cause changes in the characteristics of mobile traffic demand. Such a variety in traffic characteristics easily causes traffic unbalance and, then, congestion problems in mobile networks which have to manage that changing traffic. Thus, congestion problems in a multi-RAT scenario come, as in single-RAT scenarios, from unequal spatial or temporal traffic patterns, but also due to different traffic nature, i.e., Quality of Service (QoS) requirements, burstiness or bandwidth allocation for the different services. So, one of the most demanding network features of multi-technology scenarios is the automatic adaptation capability, referred to as *Self-Organised* Networks (SONs), [85]. In this context, SONs can modify JRRM parameters, usually based on network Performance Indicators (PI), to cope with the heterogeneous and varying traffic.

The rest of this chapter is organised as follows. Section 4.2 outlines the load sharing problem in a multi-technology scenario and presents the state of the research and technology. Section 4.3 describes the multi-technology scenario considered here, i.e., the IS-HO algorithm to be optimised and the auto-tuning scheme. Section 4.4 presents the performance assessment based on simulations. Different configurations for the auto-tuning scheme will be tested later, and network performances will be compared. Finally, conclusions are presented in Section 4.5.

4.2 Problem Outline

In this section, the possibilities of a JRRM parameter auto-tuning scheme for traffic sharing strategies are described. Later, the state of the research and technology on this topic is detailed.

4.2.1 JRRM Auto-Tuning in a Multi-Radio Scenario

A new multi-technology scenario's complexity is usually reflected in more detailed technical specifications and management algorithms. It is crucial for network efficiency that all

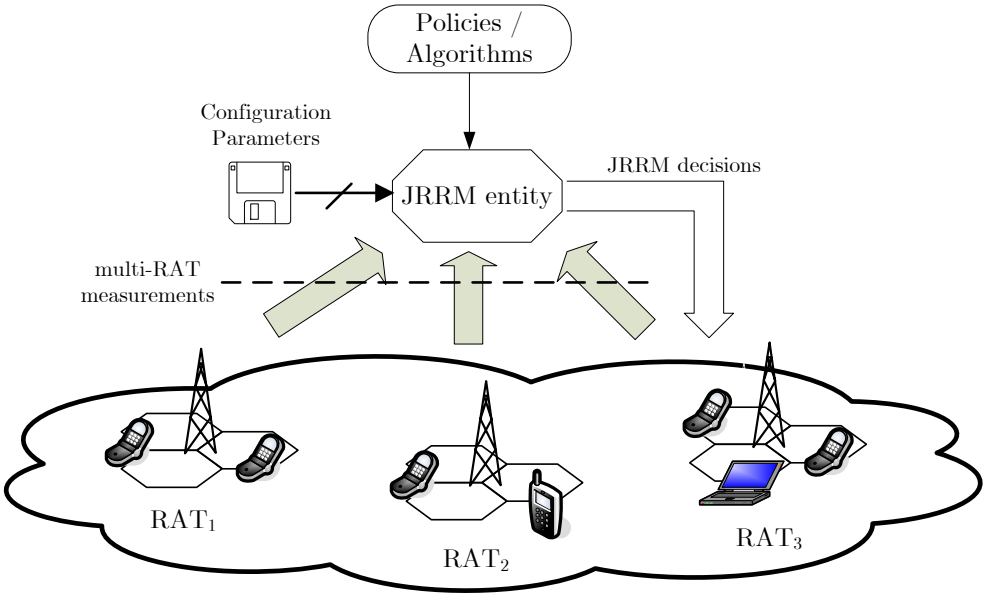


Figure 4.1: JRRM decision process.

network segments do not compete, but cooperate closely to cover user needs seamlessly and transparently. From the end-user point of view, there would be only one virtual network. As shown in Figure 4.1, the JRRM entity deals with this new heterogenous scenario by means of algorithms and policies that aim at integrating those distinct radio interfaces to support the different service data rate, traffic and user mobility requirements, [86].

Such algorithms have to manage radio resources in each individual technology, where technical specifications are quite different, as well as each RAT's measurements and indicators. As an example, $RxLEV$ in GSM and E_c/N_o in UMTS both reflect the pilot signal reception, but with a very different measure (i.e., signal level and signal-to-noise ratio, respectively), so they cannot be directly compared. Thus, the design of the JRRM entity is a very challenging task. Despite its complexity, advantages of a successful JRRM are very attractive:

a) Trunking gain by sharing resources from different RATs.

b) Extended coverage by joining different RAT service areas.

c) Service-user adequacy by choosing the best available radio technology to suit Quality-Of-Service (QoS) needs, [87].

The achievement of the previous advantages relies on a good mobility management in JRRM. For circuit-switched services, this is accomplished by JAC and IS-HO, as commented in the previous section. IS-HO and JAC algorithms could be considered as the two main JRRM mechanisms to get a seamless network. Specifically, IS-HO (also called *Vertical*-HO) plays a very special and important role in this scenario. Since not all the RATs have a global coverage, it is the basic mechanism to get a global coverage handling the user over distinct technologies.

Additionally, IS-HO is also used for load balancing when some RAT is congested while

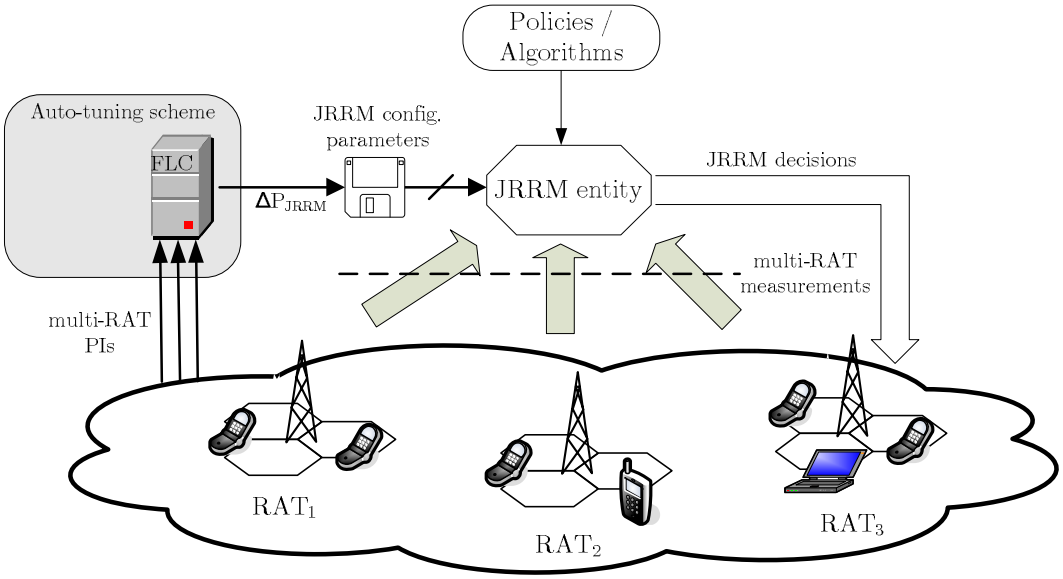


Figure 4.2: IS-HO parameters modification scheme.

others have free resources. With the aim of a simple JRRM design and ensuring network stability, operators usually configure JRRM with very simplistic decisions (e.g., service 'A' and 'C' are always carried through RAT_1, and service 'B' is always carried through RAT_2, the so-called traffic splitting policies), [25][88], so unbalanced traffic situations are not so rare. Through IS-HO parameters and thresholds settings, inter-technology traffic flows can be controlled and a better network performance could be achieved, [89]. In addition, as traffic conditions are quite changeable, network performance is optimised constantly by adapting those IS-HO parameters to traffic changes. A new traffic condition will lead to a new and, hopefully, optimum network performance.

The JRRM auto-tuning process must take decisions about parameter modifications based on network conditions, shown in Figure 4.2. The auto-tuning scheme has to deal with information from different technologies, which can be done easily by means of Fuzzy Logic techniques. Fuzzy Logic is also capable of translating human knowledge into rules, expressed as a set of IF/THEN rules, [90], for network parameter modifications. Thus, operator expertise can be automatically applied for complex problems. In the problem considered here, FLC collects statistics from different technologies and translates those indicators into network states and parameter changes (ΔP_{JRRM} in the figure). For instance, FLC might detect that call blocking rate is high in one RAT, deduce that this particular RAT is loaded and decide to decrease IS-HO margins to other underused RATs. Thus, FLC controllers adjust IS-HO parameters applying human reasoning to reach better network traffic sharing via improved configuration.

4.2.2 State of Research

The design of JRRM algorithms has received considerable attention in the literature. In a first stage, references focused on the definition of network topology and JRRM entities, so that the mobile network could implement JRRM capabilities, [25][91]. 3GPP has also

defined different grades of cooperation between technologies in a multi-RAT scenario, including Wireless Local Area Network (WLAN), [92], and [93] also established different working schemes for JRRM depending on that cooperation.

In a later stage, many JRRM schemes have been proposed. First proposals describe JRRM schemes and self-tuning algorithms changes based on some specific rules previously defined, so they are usually categorised as *policy-based* algorithms. Most studies deal with IS-HO, JAC and RAT selection algorithms. So, [94][88] make some performance analysis for IS-HO and RAT selection approaches in a simplistic scenario over simulation platforms. While [94] focuses on the analysis of the standard GSM-UMTS IS-HO procedure, [88] proposes different load balance JRRM schemes for a better network performance through an adequate configuration of load thresholds. A different approach consists in modelling multi-technology scenarios with Markov chains, showing the advantages of a JRRM approach, [26][95]. More sophisticated, [96][27][86] introduce more refined network, traffic, terminal and mobility characteristics in their performance analysis platforms for JAC and IS-HO algorithms. An additional step in JRRM schemes focuses the performance analysis over some specific parameter configurations. So, [89] evaluates the impact of IS-HO timing parameters (i.e., time-to-trigger) changes on the global network performance. Additionally, [97] describes the influence of RAT coverages over the global JRRM performance.

As an advantage, policy-based algorithms are easy to implement and control. However, they experience strong limitations, specially coming from the comparison of different technology parameters. FLC-based schemes overcome these limitations, and they are the most used approach for the auto-tuning of network parameters, [90]. Fuzzy Logic's popularity comes from its capacity of translating human knowledge into rules, which can be automatically applied to a specific problem. Additionally, FLCs can successfully deal with information of very dissimilar nature, and, so, fuzzy decision making algorithms have usually been proposed for mobile network scenarios.

About single network scenarios, 3G RATs has been the main platform for FLC performance assessment, due to the complexity and flexibility of UMTS RRM techniques, specially soft-handover, [98][99][100], power control, [101][102][103], or admission control techniques, [104]. In GERAN, [12] modifies handover margins and signal-level constraints based on network statistics for traffic sharing. About multi-technology scenarios, FLC has been used for JRRM decisions (i.e., non-policy based RRM techniques), [105][106], and JAC and IS-HO parameter modifications, [104][107][108].

With the aim of adapting the network to the traffic changing conditions, fuzzy logic auto-tuning schemes must be flexible and must include some mechanisms to perform differently in each new situation, [109]. When network traffic conditions change quickly, it is interesting to analyse how fast and precise the FLC is reaching the new network configuration. FLC internal settings, such as inference rules or membership functions, have an influence on how the JRRM entity (and, consequently, network performance) adapts to the new scenario. Figure 4.3 shows the addition of this FLC setting procedure.

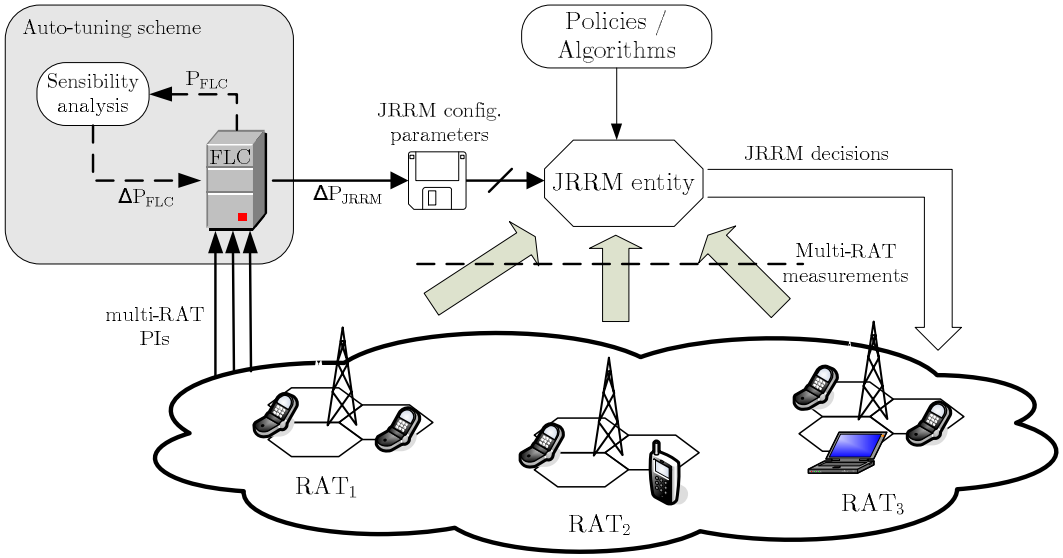


Figure 4.3: FLC and IS-HO parameters modification scheme.

Some parameters in the auto-tuning scheme are modified. Since network performance is directly affected by the particular FLC setting, different FLC configurations lead to different adaptation rates and, then, network performance indicators. A better FLC setting would reach the optimal network performance faster (i.e., in few iterations) and without network performance oscillations.

Different proposals detail several adaptative FLC schemes with that aim in a multi-technology scenario. [110] and [111] present an adaptative fuzzy scheme for JRRM in a WLAN/GSM/UMTS scenario, considering economical and user preferences in their decisions, and [112] and [113] focus on the admission stage, performing centralised and distributed schemes, respectively. With special interest for this thesis, [114] implement a load sharing mechanism through a changing fuzzy scheme in a WLAN/UMTS scenario. This work defines a load balancing algorithm and analyses performance improvements when some IS-HO parameters are modified by a FLC.

In this chapter, several auto-tuning schemes are proposed for a standard GSM/UMTS IS-HO algorithm, [115]. The considered heterogeneous scenario comprises GSM and UMTS access technologies, but could easily be extended to other RATs. As a contribution from this thesis, quality and level parameters from that IS-HO algorithm are modified in a FLC-based scheme with load balance purposes. A very high network sensitivity to those parameters is expected, which is the main reason for such a selection. The proposed scheme is tested in a joint network simulation platform. This work follows a similar methodology to that in [12], which describes a single-technology scenario. Model accuracy is also required, so the simulation platform includes most up-to-date network features. Additionally, this thesis looks for the best (i.e., the fastest) FLC configuration. So, strongly unbalanced scenarios are configured and FLC adaptation is mainly tested over the simulation platform.

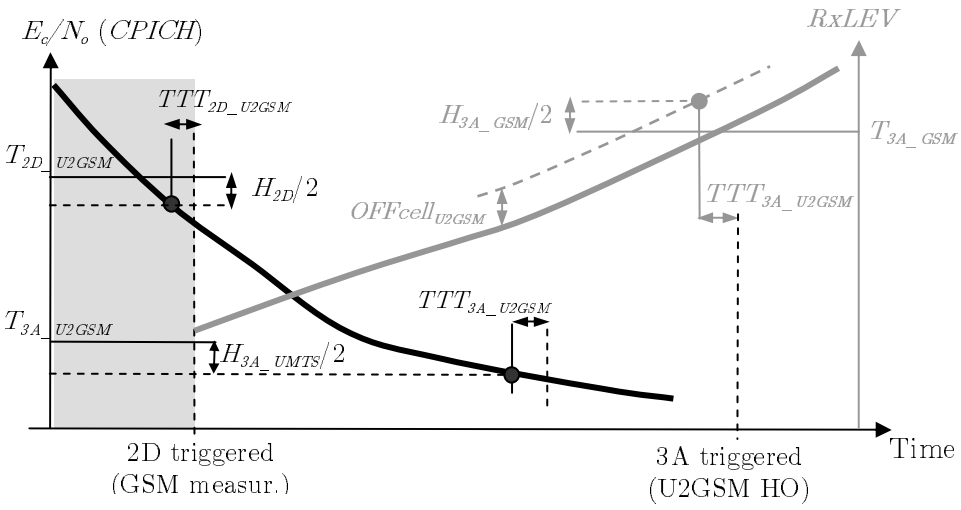


Figure 4.4: UMTS to GSM IS-HO algorithm.

4.3 Description of the Auto-Tuning Scheme

This section outlines the proposed IS-HO auto-tuning scheme for load sharing in a multi-RAT network. Firstly, the standardised IS-HO algorithm under optimization is presented, identifying its main parameters. Secondly, the FLC for tuning IS-HO parameters is described.

4.3.1 IS-HO Algorithm Description

A detailed description of the HO algorithm from UMTS to GSM (U2GSM) is given in 3GPP standards, [115]. Similarly to intra-RAT handovers, an IS-HO occurs after two criteria are met:

1. a low signal quality/level is experienced at the origin RAT, and
2. the target RAT has enough signal quality and signal level.

Figure 4.4 summarises the IS-HO process for a user moving from a UMTS cell to a GSM cell. The bold line in the figure represents the Signal to Noise-Interference ratio for the Common PIlot CHannel (CPICH) in UMTS, $CPICH\ E_c/N_o$, and the gray line plots the Received signal level, $RxLEV$, in GSM. On the time axis, two events are clearly identified. Event 2D starts the collection of GSM measurements, and event 3A starts the HO process. Additional hysteresis levels (H_{2D} and H_{3A} and temporal windows (TTT_{3A_U2GSM} and TTT_{2D_U2GSM}) are defined. Event 3A is triggered when both a) UMTS connection quality is below some threshold, T_{3A_U2GSM}, and, b) GSM signal level is above a similar threshold, T_{3A_GSM}.

These two conditions triggering the event 3D can be formulated as

$$\frac{E_c}{N_o}(i) < T_{3A\_U2GSM}(i) - \frac{H_{3A\_UMTS}(i)}{2} \quad \text{, and} \tag{4.1}$$

$$RxLEV(j) + OFFcell_{U2GSM}(i,j) > T_{3A\_GSM}(j) + \frac{H_{3A\_GSM}(j)}{2} \; . \tag{4.2}$$

where i and j are the origin and destination cells, T_{3A_U2GSM} and T_{3A_GSM} are signal-quality (UMTS $CPICH\ E_c/N_o$) and signal-level (GSM $RxLEV$) thresholds, H_{3A_UMTS} and H_{3A_GSM} are hysteresis parameters to avoid instabilities, and $OFFcell_{U2GSM}(i,j)$ is an offset term to bias IS-HO decisions in favour of any cell in the destination RAT. Both equations must be fulfilled for TTT_{3A_U2GSM} seconds. All terms in (4.1) and (4.2) are expressed in decibels, and defined on a cell basis except $OFFcell_{U2GSM}(i,j)$, which is defined on a per-adjacency basis.

As already shown in [94], T_{3A_U2GSM} has a strong influence on inter-RAT (i.e., IS-HO) call flow intensity. Generally speaking, T_{3A_U2GSM} manages the overall call flow between RATs, with no control on the final destination cell. Note that only calls satisfying (4.1) will be evaluated by (4.2). A large T_{3A_U2GSM} value makes more calls to become candidates to trigger an IS-HO (i.e., more calls, even with acceptable UMTS signal quality, could be redirected to GSM). Subsequently, $OFFcell_{U2GSM}(i,j)$ controls the final destination for calls fulfilling (4.2), where j is any neighbour cell of cell i.

In current vendor equipment, the IS-HO algorithm from GSM to UMTS (GSM2U) may have slight differences from its U2GSM counterpart. For simplicity, a symmetric algorithm has been assumed in this work, since a direct translation can easily be made by a proper setting of existing parameters.

From the previous explanation, it can easily be deduced that all previous IS-HO parameters in both directions (i.e., thresholds T_{3A_U2GSM} and T_{3A_GSM2U}, and offset parameters $OFFcell_{U2GSM}$ and $OFFcell_{GSM2U}$) can be used to perform load sharing between RATs, and will thus be the main focus of the auto-tuning process.

4.3.2 Auto-Tuning Scheme

The auto-tuning algorithm presented in this section is implemented by a Fuzzy Logic Controller (FLC), [90]. The proposed scheme adjusts thresholds (i.e., T_{3A_U2GSM} and T_{3A_GSM2U}) on a per-cell basis and offset parameters (i.e., $OFFcell_{U2GSM}$ and $OFFcell_{GSM2U}$) on a per-adjacency basis. As shown in Figure 4.5, U2GSM FLCs compute the increments in $OFFcell_{U2GSM}$ and T_{3A_U2GSM} for each cell (i.e., $\Delta OFFcell_{U2GSM}$ and ΔT_{3A_U2GSM}) from past congestion statistics in both UMTS and GSM. GSM2U FLCs compute $\Delta OFFcell_{GSM2U}$ and ΔT_{3A_GSM2U} in a similar way. For brevity, the following explanation is restricted to only one direction in IS-HO, i.e., from UMTS to GSM.

The structure of FLCs is depicted in Figure 4.5. A FLC can be divided into three main blocks: fuzzifier, inference engine and defuzzifier. Congestion rates in the uplink

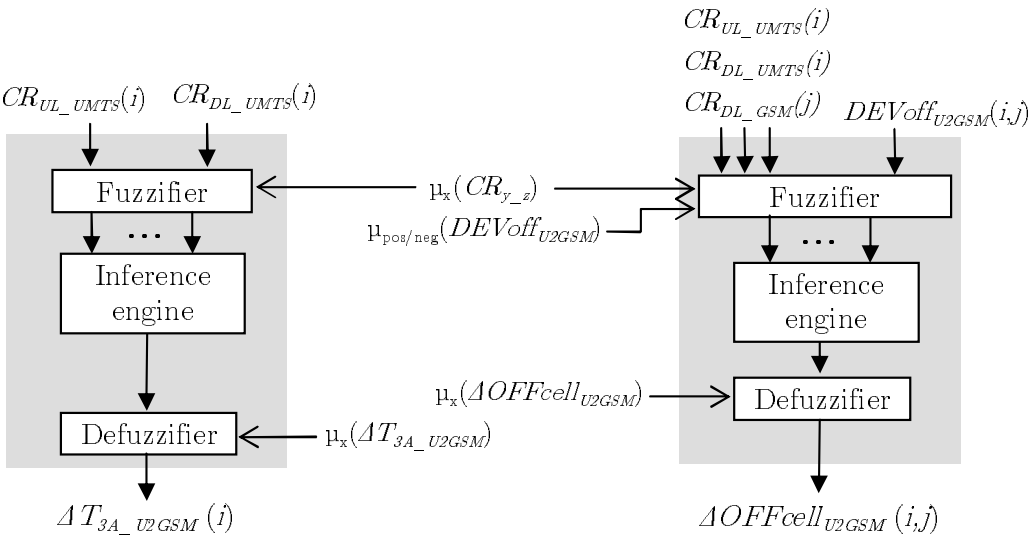


Figure 4.5: U2GSM fuzzy controller diagrams.

and downlink for both RATs are used as inputs. They are denoted as $CR_{y\_z}(i)$, where $y \in \{UL, DL\}$ and $z \in \{GSM, UMTS\}$. For UMTS, CR is the percentage of time during which new calls cannot be accepted, i.e., the time during which all channel codes are assigned, or maximum DL signal power or maximum UL interference power is reached at the base station. For GSM, CR is the percentage of time during which all Time SLots (TSLs) are occupied. To tune $OFFcell_{U2GSM}$, an additional input, $DEVoff_{U2GSM}$, indicates the deviation of the current offset value from the default setting as

$$DEVoff_{U2GSM}(i,j) = OFFcell_{U2GSM}(i,j) - OFFcell^{(0)}_{U2GSM}(i,j)\,, \tag{4.3}$$

where $OFFcell^{(0)}_{U2GSM}(i,j)$ is the original (default) value of the offset parameter, before any modification by the FLC.

For simplicity, all FLCs are implemented based on the Takagi-Sugeno approach, [90]. In the fuzzifier, FLC inputs (i.e., network performance indicators) are classified according to some so-called linguistic terms. The fuzzyfier translates input values into a value in the range $[0, 1]$ indicating the degree of membership to a linguistic term, $x \in \{L(low), M(medium), H(high)\}$, according to several input membership functions, μ_x. For instance, $\mu_{low}(CR_{UL\_GSM})$ function indicates how low is the uplink CR in GSM valued between '0' (i.e., CR_{UL_GSM} is not low at all) and '1' (i.e., it is definitely low). For simplicity, the selected input membership function are triangular or trapezoidal, as shown in Figure 4.6(a). It should be pointed out that CR membership functions are similar for GSM or UMTS and uplink or downlink, i.e.,

$$\mu_x(CR_{y\_GSM}) = \mu_x(CR_{y\_UMTS})\,, \text{and} \tag{4.4}$$
$$\mu_x(CR_{DL\_z}) = \mu_x(CR_{UL\_z})\,. \tag{4.5}$$

In the inference engine, a set of IF-THEN rules define the mapping of the input to

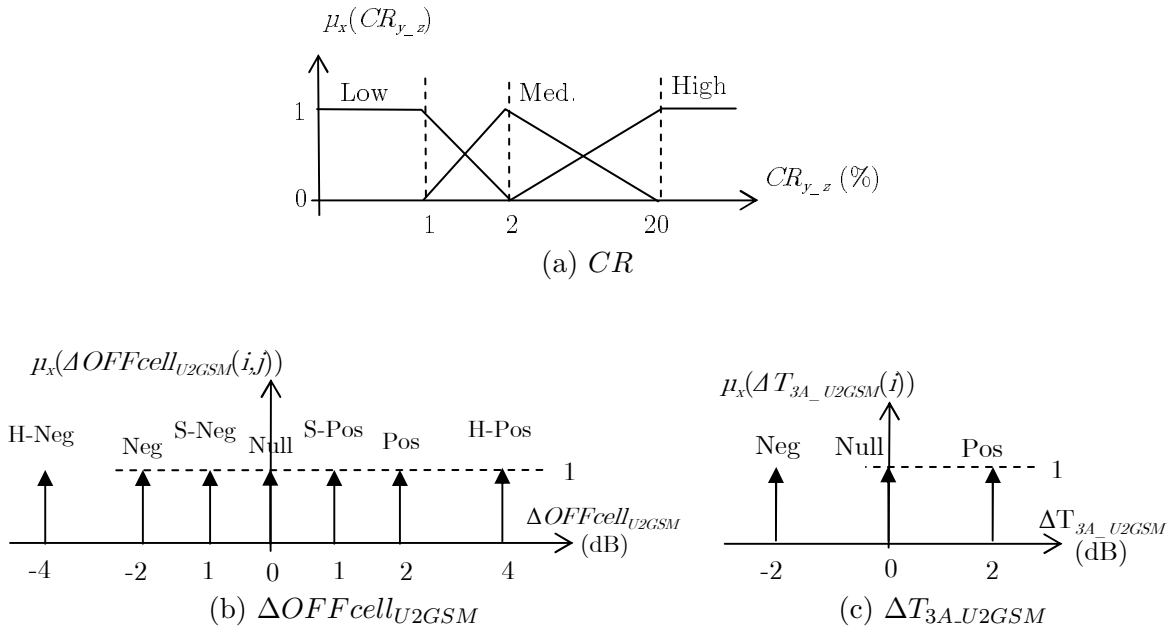


Figure 4.6: Input and output membership functions for U2GSM FLCs.

the output in linguistic terms. Table 4.1 describes the set of rules in the U2GSM tuning process. Briefly, the parameter $\Delta T_{3A\_U2GSM}(i)$ is positive (i.e., $T_{3A\_U2GSM}(i)$ quality threshold in (4.1) increases) when $CR_{UL\_UMTS}(i)$ or $CR_{DL\_UMTS}(i)$ are large, making the U2GSM IS-HO easier. For $\Delta OFFcell_{U2GSM}$, FLC rules point that $\Delta OFFcell_{U2GSM}$ is positive (negative) when CR_{y_UMTS} is more (less) congested than CR_{DL_GSM}. Moreover, when all the system experiences a low congestion (first two rules in $\Delta OFFcell_{U2GSM}$ FLC), $\Delta OFFcell_{U2GSM}$ is modified so that $OFFcell_{U2GSM}$ recovers its original value (i.e., $\Delta OFFcell_{U2GSM}$ is positive if $DEVoff_{U2GSM}$ is negative, and viceversa). FLC rules for GSM2U parameters (not shown) are similar.

Finally, the defuzzifier obtains a crisp output value by aggregating all rules. As shown in Figure 4.6(b), the output membership functions for $\Delta OFFcell_{U2GSM}$ are constants. The output membership functions for ΔT_{3A_U2GSM} are similar, as shown in Figure 4.6(c), but only with 'negative', 'null' and 'positive' values (-2, 0 and 2 dBs, respectively). The centre-of-gravity method is applied here to compute the final value of the output.

To avoid network instabilities due to excessive parameter changes, T_{3A_U2GSM} and $OFFcell_{U2GSM}$ values are restricted to a limited variation interval. Thus, the value of one generic parameter P for the next iteration can be expressed as

$$P^{(u+1)}(i) = min\{\ max\{\ P^{(u)}(i) + \Delta P^{(u)}(i), P_{min}(i)\ \}, P_{max}(i)\ \} \tag{4.6}$$

where P is the generic parameter, $[P_{min}\ P_{max}]$ is the allowed variation interval, and u is the current iteration in the optimisation process. This is aligned to usual operator policies, which avoid, if possible, large changes in network configuration for safety reasons.

To avoid unnecessary IS-HOs, T_{3A_U2GSM} and $OFFcell_{U2GSM}$ are not always modified. On the contrary, changes proposed by the FLC controlling $T_{3A\_U2GSM}(i)$ are only

$CR_{UL\_UMTS}(i)$	$CR_{DL\_UMTS}(i)$	$\Delta T_{3A\_U2GSM}(i)$
L	L	Neg
L	M	Null
M	L | M	Null
H	-	Pos
-	H	Pos

$CR_{UL\_UMTS}(i)$	$CR_{DL\_UMTS}(i)$	$CR_{DL\_GSM}(j)$	$DEVoff_{U2GSM}(i,j)$	$\Delta OFFcell_{U2GSM}(i,j)$
L	L	L	Pos	S-Neg
L	L	L	Neg	S-Pos
L	L	H	-	H-Neg
H	-	L	-	H-Pos
-	H	L	-	H-Pos
H	-	M	-	Pos
-	H	M	-	Pos
L | M	L | M	H	-	Neg
L | M	L | M	L	-	Pos
L | M	L | M	M	-	Null
H	-	H	-	Null
-	H	H	-	Null

"|" : Logical OR

Table 4.1: U2GSM fuzzy logic controller rules.

implemented in the scenario when the average value $OFFcell_{U2GSM}(i,j)$ for all j cells is close to its variation limits (i.e., [-6 6] dB). As some situations of unbalanced traffic can be managed by only changing $OFFcell_{U2GSM}$, this approach tries to avoid unnecessary T_{3A_U2GSM} modifications and, therefore, unnecessary IS-HOs. Thus, the network signalling load is only increased when needed.

4.4 Auto-tuning Performance Assessment

This section presents the assessment of the proposed auto-tuning scheme. The simulation set-up is described first and performance results are presented later. Results are divided into a first part, which considers an initial FLC configuration, and a second part, where FLC internal settings are modified to check network sensitivity respecting to changes in the FLC.

4.4.1 Simulator Set-Up

A dynamic system-level simulator of a GSM-UMTS network has been developed in MATLAB® as the main assesment platform. Figure 4.7 shows the simulator structure. The multi-technology platform has been built from two original single technology simulators. Traffic generation, cell and RAT re/selection and admission control processes from each separated technology were joined in each single multi-technology module. An IS-HO module from and towards both radio accesses is also included in the simulator. The simulation platform also includes the main intra-RAT functionalities (e.g., power control, direct retry,

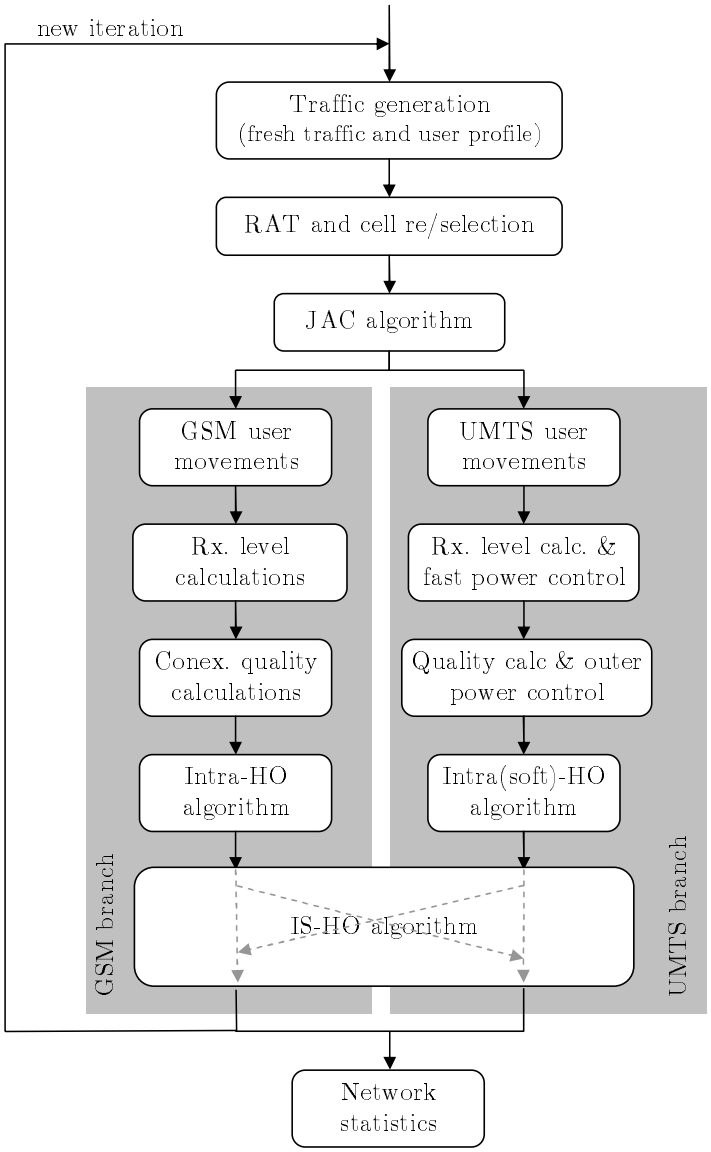


Figure 4.7: Multi-technology Simulator Scheme.

intra-HO or call dropping).

The simulation scenario models a macro-cellular environment where full overlap between GSM and UMTS coverage areas exists. The layout, shown in Figure 4.8, consists of 19 tri-sectorised sites evenly distributed in the scenario. Thus, every site has 3 GSM cells and 3 UMTS cells (i.e., GSM and UMTS cells are co-sited). Table 4.2 summarises the models and default parameters in the simulator, which have been widely used in the literature.

To check FLC auto-tuning capability, a varying traffic demand is used in the experiments. Figure 4.9 shows the offered traffic temporal distribution (consisting of circuit-switched voice calls) configured in the simulations. Initially, GSM and UMTS traffic sources are configured to result in a strongly unbalanced scenario, where GSM is congested and UMTS is underused. Thus, it is expected that parameter changes performed by FLC manage to relieve congestion in most GSM cells. At some instant (i.e., the 20^{th} iteration), the congestion situation is reversed to check the capability of the network to

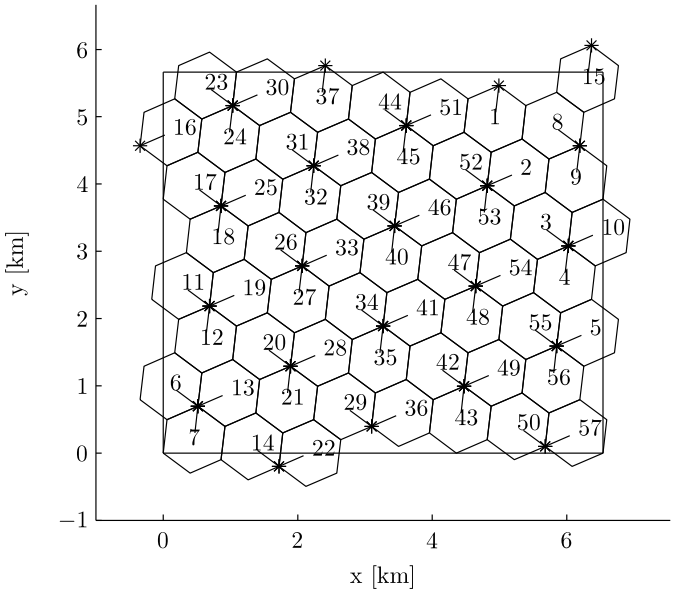


Figure 4.8: Simulation Scenario.

Scenario	TU3, MACRO, cell radius 0.5 km, 57 UMTS cells + 57 GSM cells, wrap-around	
Propagation model	Okumura-Hata with wrap-around, correlated log-normal slow fading, $\sigma_{SF} = 6\,dB$	
Mobility model	Constant direction and speed (3 km/h)	
Service model	Speech, MCD=100 s, activity factor $\alpha = 0.5$	
Spatial traffic distr.	Uniform	
BS model	Tri-sectorized antenna, $EIRP_{max} = 43\,dBm$, 1 TRX (GSM), 1 channel code tree (UMTS)	
Adjacency plan	Symmetrical adjacencies, 32 per cell	
JRRM parameters	T_{3A_U2GSM}	-28 dB
	T_{3A_GSM2U}	-100 dBm ([-110 . . . -47])
	$OFFcell$	0 dB ([-6 . . . 6])
Time resolution	480 ms (GSM), 100 ms (UMTS)	
Netw. simulated time	1 h (per optimisation step), 80 h (total)	

Table 4.2: Simulation parameters.

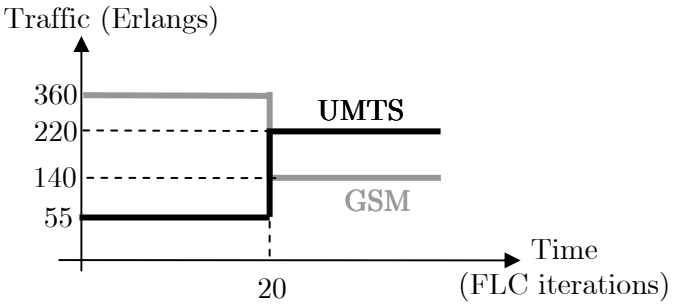


Figure 4.9: Offered Traffic Temporal Distribution.

adapt to changes in the traffic distribution (e.g., population movements, new premises). Depending on the adaptation rate, network indicators will reach equilibrium using many or fewer optimisation steps (i.e., FLC iterations). For clarity, these two periods are hereafter referred to as first and second stage.

Assessment is based on three overall network performance indicators: (a) single-RAT Blocked Call Rate (BCR_{GSM} and BCR_{UMTS}) as a network capacity indicator; (b) IS-HO ratio (i.e., ratio of total IS-HOs, U2GSM and GSM2U, to global carried calls) as a signalling load indicator; and (c) global blocking rate, BCR_{total} defined as

$$BCR_{total} = \frac{blocked_{GSM} + blocked_{UMTS}}{offered_{GSM} + offered_{UMTS}} \quad , \tag{4.7}$$

where $blocked_z$ and $offered_z$ refer to the number of blocked and offered calls for each radio access technology z.

4.4.2 Performance Results

Main results for the FLC-based auto-tuning scheme are described in this section. A first part analyses the results for a default FLC setting. With the aim of a faster network adaptation, additional FLC settings are tested and compared to the initial configuration.

Preliminary results with default FLC settings

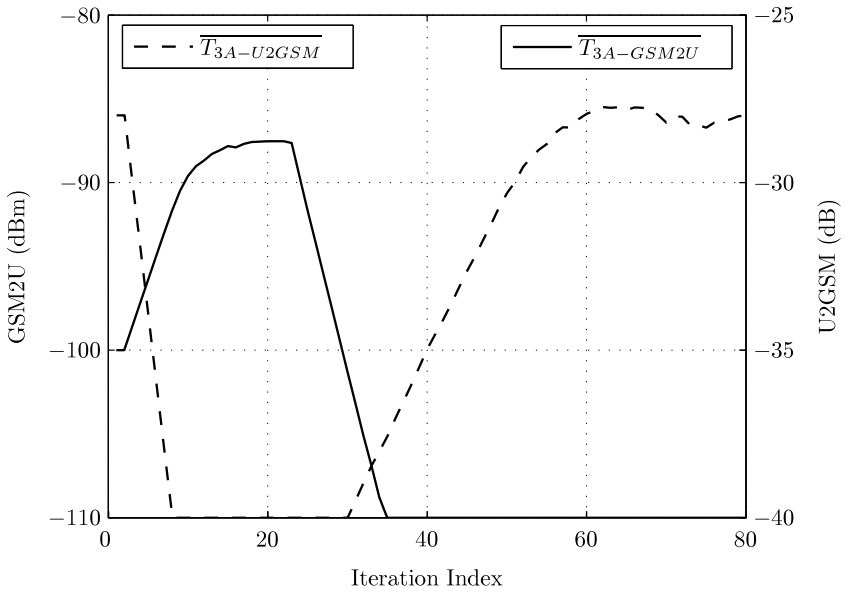
Multiple iterations have been simulated under the traffic conditions described in section 4.4.1. Since traffic spatial distribution is uniform in the scenario, $T_{3A\_U2GSM}(i)$ and $OFFcell_{U2GSM}(i,j)$ cell averages are statistically representative of all cells in the scenario. Such averages are defined as

$$\overline{T_{3A\_U2GSM}} = \sum_i \frac{T_{3A\_U2GSM}(i)}{N_{cell}} \quad , \text{ and} \tag{4.8}$$

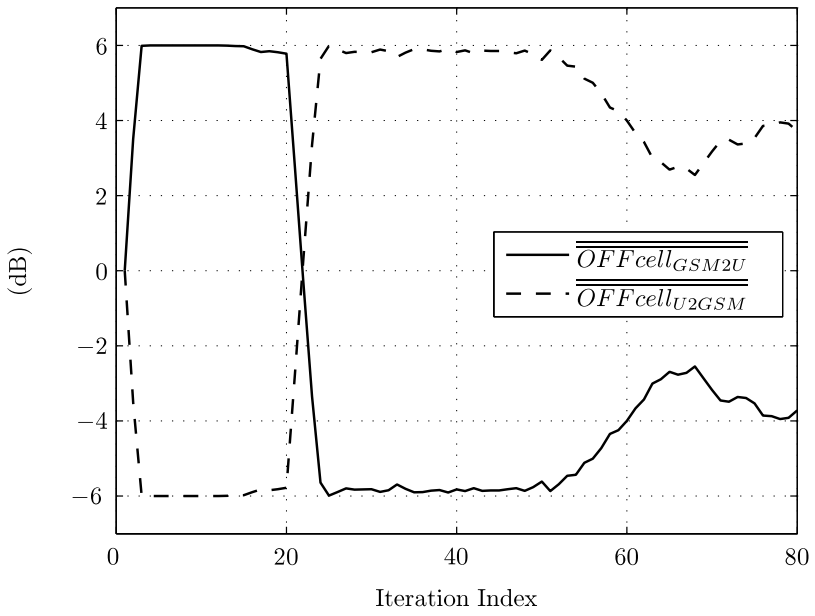
$$\overline{\overline{OFFcell_{U2GSM}}} = \sum_i \left(\frac{\frac{\sum_j OFFcell_{U2GSM}(i,j)}{N_{adj}(i)}}{N_{cell}} \right) \quad , \tag{4.9}$$

where N_{cells} is the number of cells in the origin RAT (i.e., 57), and $N_{adj}(i)$ in (4.9) represents the number of adjacent cells in the destination RAT for cell i. Similar equations can be defined for GSM2U statistics.

Figure 4.10 shows the evolution of parameters across iterations. For this purpose, the figure shows the values of indicators (4.8) and (4.9) for both U2GSM and GSM2U HOs. It should be pointed out that confidence intervals for these averages are negligible, and are thus not shown. In the figure, it is observed that, in the first stage, when GSM is overloaded, FLC favours GSM2U IS-HO by increasing $T_{3A\_GSM2U}(i)$ and $OFFcell_{GSM2U}(i,j)$. Similarly, FLC progressively reduces U2GSM parameters to avoid flow of users from UMTS to GSM. As a result, T_{3A_U2GSM} becomes highly negative (i.e., -45 dB) at the



a) $\overline{T_{3A}}$ parameter evolution



b) $OFFcell$ parameter evolution

Figure 4.10: Simulation Scenario.

end of this 1^{st} stage. This situation is maintained until the 2^{nd} stage, when the traffic distribution changes.

Once traffic distribution changes in the 2^{nd} stage, FLCs change network parameters to cope with congestion in UMTS. Thus, T_{3A_GSM2U} is restricted to avoid GSM2U HOs, while T_{3A_U2GSM} is relaxed to favor U2GSM user flow. It is worth noting that changes in T_{3A_U2GSM} and T_{3A_GSM2U} start some iterations after traffic change in the 20^{th} iteration. As explained before, only when $OFFcell_{U2GSM}$ and $OFFcell_{GSM2U}$ parameters are close to their limit values (i.e., around the 5^{th} and 25^{th} iterations in Figure 4.10), FLC changes in T_{3A_U2GSM} and T_{3A_GSM2U} are allowed. Parameter changes find the equilibrium values in both stages when CR_{GSM} and CR_{UMTS} are balanced, as FLC rules were defined.

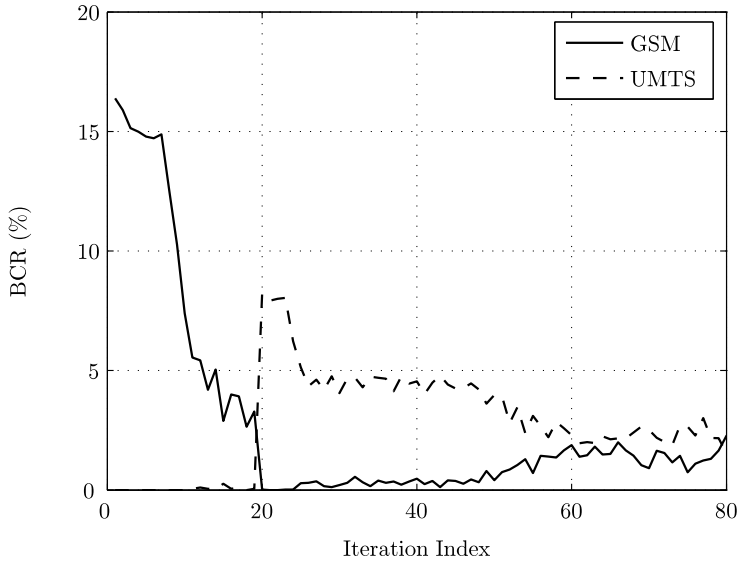


Figure 4.11: Blocked Call Rate (BCR) evolution.

Figure 4.11 shows the Blocked Call Rate (BCR) in both technologies across iterations. As expected, changes performed by the FLC progressively reduce blocking differences between technologies in both simulation stages. In the figure, it is observed that, by sharing load between RATs, BCR in GSM is reduced in first stage up to 12% in absolute terms (i.e., from 16% to 4%), while keeping BCR in UMTS almost unaltered. As a result of congestion relief, network carried traffic increases in, approximately, 15% (not shown in the figure). These high figures are obtained due to the strongly unbalanced traffic scenario.

In the second stage, the initial load imbalance between technologies is not so severe and BCRs are not as high as in the first stage. Therefore, the rules fired in the FLC inference engine suggest more subtle parameters changes. As a result, convergence to the equilibrium is slower and balance of BCRs between RATs is only reached after 35-40 iterations. In spite of the FLC capability to equalise blocking, it is observed in both Figure 4.10 and Figure 4.11 that the convergence to equilibrium is somewhat slow.

Note that the starting value for T_{3A_U2GSM} in the 2^{nd} stage is far away from values that can have an influence on IS-HO call flow. Such a lack of sensitivity is clearly observed in Figure 4.11, where BCR values remain unchanged from 22^{nd} to 42^{nd} iteration, regardless of changes in T_{3A_U2GSM} shown in Figure 4.10. It is not until the 43^{rd} iteration that FLC manages to bring T_{3A_U2GSM} to values that affect the U2GSM IS-HO flow. It can thus be concluded that, for this scenario, network performance sensitivity to T_{3A_U2GSM} parameter is high above -32 dBs, but low (or even null) for smaller values. Hence, FLC should only modify T_{3A_U2GSM} within a range of values where network performance is responsive. Thus, periods of a non-optimal network configuration are shortened.

Figure 4.12 evaluates the influence of the tuning process on network signalling load by showing the IS-HO ratio. In the first simulation stage, a high IS-HO rate is experienced to balance traffic (up to 35% in 19^{th} iteration). In the second stage, less IS-HOs are needed

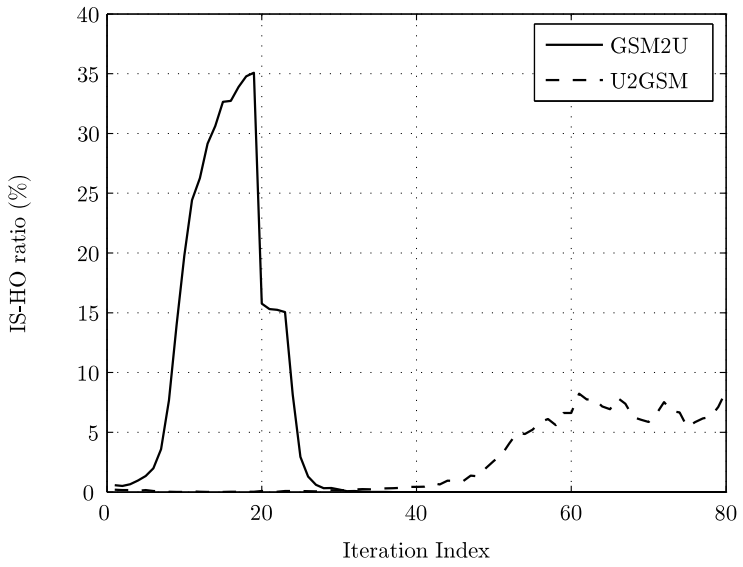


Figure 4.12: U2GSM and GSM2U IS-HO ratios.

Configurations	$\lvert\Delta T_{3A}\rvert$	T_{3A_U2GSM} (dB)		T_{3A_GSM2U} (dB)	
		Initial	Range	Initial	Range
S_0	2 dB	-28	[-45 0]	-100	no limit
S_1	1 dB	-28	[-40 0]	-95	[-98 -74]
S_2	2 dB				
S_3	4 dB				
S_4	6 dB				

Table 4.3: FLC internal parameter configurations.

and handover ratios are, then, lower (7-8%).

Yet not shown in the figures, enough call quality is always ensured at any iteration in both radio access technologies. More specifically, the probability of experiencing a Frame Error Rate (FER) larger than 5% in GSM is less than 0.01 (i.e., 1% of simulation time). Likewise, the probability of experiencing a Block Error Rate (BLER) in UMTS larger than 5% is below 0.001.

Results with optimised FLC internal settings

As stated above, the FLC with the original settings performs too slow and the convergence to the equilibrium is reached after too many iterations. The following analysis quantifies the benefit of adjusting internal FLC parameters properly. Two FLC internal parameters are modified: the bounds for T_{3A_U2GSM}, used in (4.6), and the magnitude of steps ΔT_{3A_U2GSM}, shown in Figure 4.6(c). Different configurations are tested, which are summarised in Table 4.3.

In the table, configuration S_0 is defined as a benchmark, since it is the original FLC configuration used so far. S_1-S_4 differs in ΔT_{3A} parameter (i.e., step magnitude), ranging from 1 to 6 dB in absolute value (i.e, $+\lvert\Delta T_{3A}\rvert$ when FLC decides a positive variation

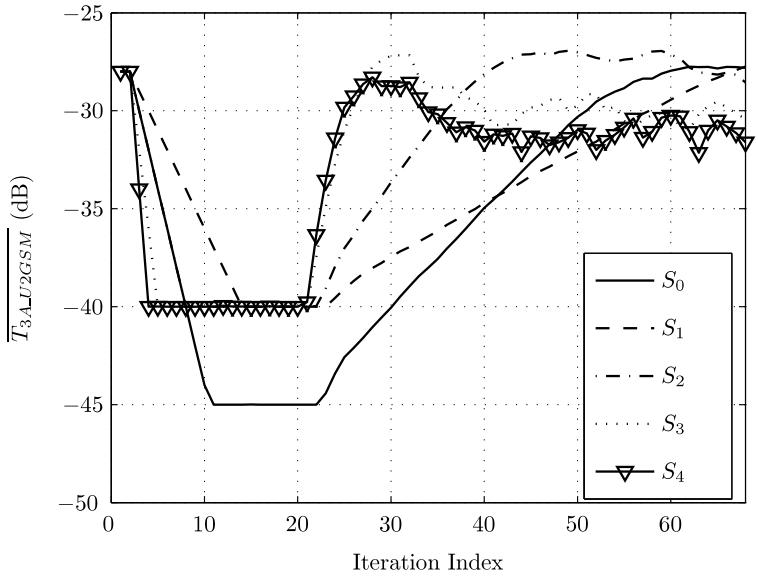


Figure 4.13: Evolution of IS-HO threshold with different FLC configurations.

and $-|\Delta T_{3A}|$ when negative). Those configurations are equally applied to U2GSM and GSM2U parameters. Large ΔT_{3A} values help to reach the desired T_{3A_U2GSM} or T_{3A_GSM2U} value in fewer iterations, but could lead to oscillations of IS-HO parameters around their equilibrium. Hence, there is a trade-off between speed and precision in reaching the final IS-HO parameter values. In addition, configurations S_1-S_4 modify the initial value and minimum bound for both IS-HO thresholds, T_{3A_U2GSM} and T_{3A_GSM2U}.

Figure 4.13 shows, for every configuration in Table 4.3, $\overline{T_{3A\_U2GSM}}$ evolution across iterations. As expected, IS-HO threshold values in curves $S_1 - S_4$ are limited to -40 dB, in contrast to -45 dB for S_0. Thus, when the 2^{nd} stage starts (i.e., the 20^{th} iteration), T_{3A_U2GSM} starts to increase from a higher value with S_1-S_4 configurations. Moreover, the different slopes in Figure 4.13 correspond to different $|\Delta T_{3A}|$ values defined in Table 4.3. A larger value of $|\Delta T_{3A}|$ (e.g., 6 dB in S_4) speeds up the convergence process, but, at the same time, it causes oscillations in $\overline{T_{3A\_U2GSM}}$ evolution. A similar behaviour is observed with $\overline{T_{3A\_GSM2U}}$ variations (not shown).

Changes in FLC settings have a strong impact on the number of iterations needed to reach the load balance situation (i.e., $BCR_{GSM}{\approx}BCR_{UMTS}$ in Figure 4.11). This temporal behaviour becomes especially important when IS-HO parameters must adapt to abrupt changes of traffic demand. Figure 4.14 shows BCR_{total} statistics. In the 1^{st} stage (1^{st}-20^{th} iteration), S_3 and S_4 configurations reach the balance situation (i.e., $BCR_{total} \approx 2\%$) around 10 iterations before S_0 or S_1. S_2 performance falls in between the two pairs. A similar trend is observed in the 2^{nd} stage, where $BCR_{total}{\approx}2\%$ is reached in the 33^{rd} and 28^{th} iterations for S_3 and S_4 , respectively. In contrast, S_0 and S_1 make $\overline{T_{3A\_U2GSM}}$ change later and slowly (as observed in Figure 4.13) and, thus, BCR_{total} also evolves slowly.

From Figures 4.13 and 4.14, it can be concluded that all FLC configurations achieve the right IS-HO parameter value and a stable BCR, provided that enough iterations are simulated for each traffic scenario. However, some configurations provide faster network

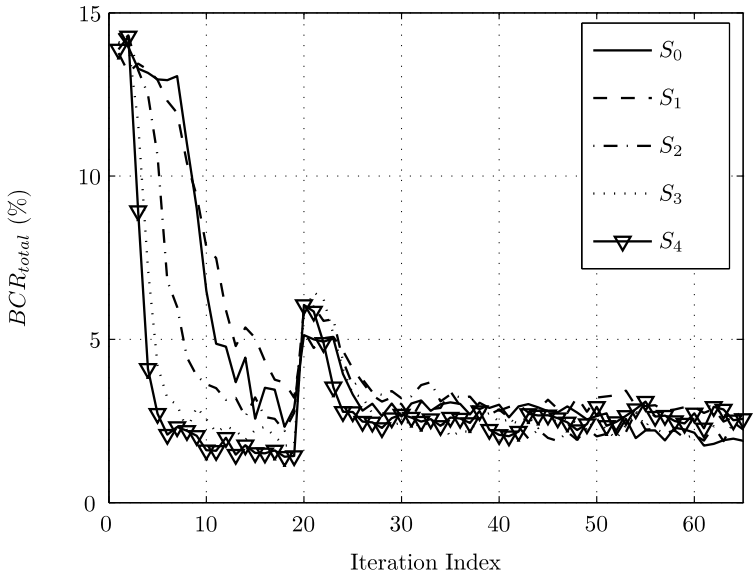


Figure 4.14: Total blocked call rate with between configurations.

FLC configuration	S_0	S_1	S_2	S_3	S_4
$CumBCR_{S_c}(\%)$	7.91	8.39	6.08	4.47	3.71

Table 4.4: BCR cumulative value for 1^{st} stage.

adaptation than others, which translates into a better network performance temporarily (e.g., in 5^{th} iteration, $BCR_{total} \approx 13\%$ and 3% for S_0 and S_4 respectively, Figure 4.14). Therefore, it is worthwhile to assess the global network performance along the whole adaptation process (i.e., transient regime) and not only in equilibrium (i.e., steady state). For this purpose, a global figure of merit, $CumBCR_{S_c}$, is defined as

$$CumBCR_{S_c} = \frac{\sum_n blocked_{GSM,n,c} + blocked_{UMTS,n,c}}{\sum_n offered_{GSM,n,c} + offered_{UMTS,n,c}} \quad , \tag{4.10}$$

where the terms $offered_z$ and $blocked_z$ refers again to the number of blocked and offered calls to RAT z, n indicates the number of the iteration where statistic is collected, and S_c is the FLC configuration in Table 4.3. Table 4.4 shows the results for all FLC configurations during the 1^{st} stage ($n \in \{1 \dots 20\}$, i.e., $CumBCR_{S_c}$ calculates the blocked calls temporal average in a entire stage before traffic changes.

In the table, it is observed that S_3 and S_4 show the lowest CumBCR (i.e., more than half of that of the default settings, S_0). S_1 behaves worse than S_0, due to a smaller $|\Delta T_{3A}|$=1 dB, and, therefore, a slower network adaptation. Simulation S_2, despite of the same $|\Delta T_{3A}|$, presents a lower $CumBCR$ respecting S_0. The difference between S_0 and S_2 is in the range of IS-HO parameters, which is adapted to network sensibility for S_2 FLC configuration.

4.5 Conclusions

In this chapter, a FLC-based auto-tuning scheme has been proposed for a standard IS-HO algorithm in a multi-technology scenario. FLCs perform traffic sharing between technologies by re-directing calls between technologies through changes in IS-HO parameters. Parameters to be tuned have been a) the level and quality thresholds in the origin RAT to start the IS-HO process, and b) offset margins to find out the destination cell in the final technology. Performance assessment is carried out by implementing FLC and IS-HO algorithms in a dynamic system-level network simulator including GSM and UMTS radio access technologies with the main intra- and inter-RAT RRM algorithms.

To check FLC adaptation capabilities, a strongly unbalanced and changing traffic scenario has been simulated. Preliminary results have shown that load sharing between technologies can be performed effectively by tuning IS-HO parameters. As a result, BCR can be reduced by a factor of 4 in the extreme scenario considered. Obviously, such a large reduction comes from the specific traffic scenario, where one technology is congested while the other is underused, which might not be the case in a more realistic scenario. Nonetheless, the proposed scheme is a cost-effective means to increase network capacity, since it does not require changing network equipment. The price to be paid is a significant increase in network signaling load due to more IS-HOs. Such a negative effect can be counteracted by jointly tuning JAC and IS-HO parameters to ensure that users camp in the technology where the IS-HO would send them.

Although the FLC is able to modify network parameters to re-distribute traffic between technologies, its speed of response is dependent on its internal settings. A sensitivity analysis has been performed to quantify the effect of several internal parameters in the FLC on the speed of response. Several FLC configurations have been tested. Experiments have shown that adjusting tuning ranges and step magnitudes for the controlled IS-HO parameters has a strong impact on performance. A significant BCR reduction (up to six-fold in some iterations) can be achieved in the transient regime when larger steps and adequate parameter ranges are configured. As a global figure of merit, time-averaged BCR can be reduced by more than a half by setting FLC properly, although it is expected that a more realistic scenario (i.e., with a weaker unbalance) would result smaller gain figures.

Chapter 5

Summary Conclusions

This final chapter summarises the major findings of this thesis. A first section highlights the main contributions of this work. A second section describes possible future work in the topics covered. The last section enumerates a list of tangible results from this thesis.

5.1 Main Contributions

This thesis has dealt with very different topics, namely teletraffic models for dedicated signalling channels in GERAN and traffic sharing techniques in GERAN and multi-technology networks. In the following paragraphs, the main contributions are presented separately for each topic.

a) Teletraffic model for dedicated signalling channel in GERAN

- A complete analysis over dedicated signalling data extracted from a live network has been detailed. Such analysis has allowed to detect the main faults in existing models for dedicated signalling data and channel dimensioning.
- A teletraffic model with correlated arrivals and retrial for dedicated signalling channels in GERAN has been defined. The inclusion of correlation and retrial characteristics, specially the first one, supply a significant accuracy. Correlation characteristic has been found to come from location management procedures in signalling channels and user group movements.
- A model parameter estimation procedure is defined for the new model. This procedure is based on formulating the problem as a least square problem, trying to fit with live network data.

b) Optimal traffic sharing in GERAN

- A teletraffic model for traffic sharing has been proposed, introducing user mobility characteristics. Traffic sharing is implemented through the modification of HO margins, leading to cell service area modifications.
- An optimal analytical indicator for load balancing in GERAN has been extracted to solve localised congestion problems due to spatial concentration of traffic demand.
- A realistic scenario has been constructed from live network data. The new model and optimal load balance indicator have been tested in that scenario. Through a sensitivity analysis, it has been shown how the capacity gain with the optimal indicator depends on the geographical conditions of the problem.

c) Self-tuning of IS-HO Parameters in a GSM/UMTS scenario

- A FLC-based auto-tuning scheme for network parameter modifications in an heterogeneous scenario is proposed. The proposed scheme modifies IS-HO algorithm parameters with the aim of balancing the total traffic load in the network.
- Several FLC configurations have been also proposed, with the aim of speeding up the optimisation process. FLC configurations differ in modification steps and adequate ranges for IS-HO parameters to be modified.
- As a platform for the assessment of this proposal, a dynamic network simulator has been constructed, including both GSM and UMTS technologies. Main functionalities have been included in the network simulator, both intra- and inter-RAT capabilities. Moreover, this simulator has been a basic tool in some research projects where this thesis has been involved in.

Most of the work in this thesis has been developed in GERAN networks. In this thesis, it has been said that GERAN technology is already in a mature stage, so it is a very suitable scenario to test different optimisation strategies, specially if new proposals for optimisation do not imply equipment changes. This is the case of the different contributions enunciated in this work. Additionally, most of the strategies and algorithms proposed here can be translated to other newer technologies, as it will be presented in the next section.

5.2 Future Work

Several issues remain unexplored once this thesis is finished. Due to their interest, some of those are described below.

a) Performance Analysis of Signalling Channels in Other RATs

The proposed queueing model introduced the effect of retrials and time correlated arrivals in signalling data. However, queueing system models in this thesis have been conceived for the structure of SDCCH in GERAN. An interesting issue is how to extend the signalling performance analysis to other radio access technologies. While signalling channels in GERAN are based on a TDMA/FDMA scheme, their counterparts in UMTS and LTE (i.e., Dedicated Control CHannel, DCCH) use CDMA and OFDMA/TDMA schemes, respectively, [116][117]. Therefore, queueing models must be adapted for an adequate translation to these newer radio technologies.

In the literature, several attempts have been made to extend queueing models for TDMA/FDMA to CDMA and OFDMA for user traffic channels. For CDMA systems, the proposed model can be upgraded with state-dependent blocking probabilities to reflect that cell capacity depends on neighbour cell interference dynamically, [67][68]. A similar approach can be used in OFDMA-TDMA systems, where adaptive modulation and coding cause that the bandwidth allocated to each user is not deterministic, but dependent on channel conditions, [69]. Such an approach applied to user traffic channels could be also used for signalling channels. It can be argued that the distinction between signalling and user traffic resources in UMTS and LTE is not as clear as in GERAN. Nonetheless, it is expected that a minimum share of cell capacity is reserved for signalling purposes in these networks, as currently done by GERAN operators.

More important, this work has proved that new LU requests experience temporal correlation characteristics in many cells of a live GERAN system. Such a behaviour is expected to be the same in UMTS and LTE, since:

a) idle user mobility does not depend on the radio access technology, so retrial and correlation characteristics in traffic are expected to be maintained,

b) UMTS and LTE networks are also divided into location, routing and tracking areas, and, consequently, some kind of location update procedure is expected to be configured in those RATs, and

c) location management procedures in UMTS and LTE are quite similar to those in GERAN, [118].

With the steady decrease in cell size, it is expected that signalling due to mobility management is a major traffic component in future mobile communication networks (as has been shown for GERAN). The proposed methodology in this thesis can help UMTS and LTE operators to re-allocate cell signalling resources by detecting correlation between arrivals.

b) Optimal Load Sharing in UMTS and LTE

An optimal load balancing criterion for traffic sharing in GERAN has been formulated in this thesis. An analytical expression for this indicator has been defined and every cell in the network should equalise that indicator network wide. The proposed system model from which this expression was extracted has been conceived for voice traffic in TDMA/FDMA systems. An important issue is how to extend the presented analytical framework to other services and other radio access technologies.

In a first step, the model can be extended to multiple services by the multi-rate Erlang loss model. In a multi-service scenario several traffic flows are configured, and each source demands a different amount of radio resources. This model presents blocking probabilities depending on the system state, as usual, which is determined by the incoming rate of each service. A typical state-transition diagram can be also configured, and several effective methods exist to compute the blocking probability and its derivatives in this new multi-service scenario, [119][120]. Such a model is insensitive to the service time distribution and can consider a mixture of not only poissonian but also smoother or more bursty traffic, [121]. However, services still have full accessibility to resources and no queueing in this model is considered, which is rarely the case of interactive and background packet-data services. Network operators usually define some specific radio resource reservations and packet scheduling for non-real time services. So, as a second step, the goal of minimising blocked traffic can be substituted by that of minimising delay probability for these scheduled services (i.e., change Erlang-B by Erlang-C formula in (3.12)).

Regarding technology extensions, the multi-rate Erlang loss model with state-dependent blocking probabilities can reflect cell capacity dependence on neighbour cell interference for CDMA systems (as already suggested to extend signalling traffic models in this technology). The extension to OFDM-TDMA systems requires studying the dependence between user bandwidth and channel conditions. In all these models, a new optimal balance equation, equivalent to (3.17), could be derived by the approach in Appendix B. However, a closed-form expression of the indicator to be balanced might be difficult to obtain.

c) Self-tuning Joint Radio Resource Management Algorithms

An FLC-based auto-tuning scheme has been proposed in this thesis for IS-HO parameters. The aim was to balance the load between technologies to maximise the trunking gain. The price to be paid is an increase of inter-system signalling traffic load in the form of handovers.

Other Joint RRM algorithms can be also tuned for load balancing between technologies. In particular, JAC parameters can be configured to assign more incoming connections to RATs with free resources, [27]. Alternatively, Inter-System Cell Reselection (IS-CR) algorithm can be also configured to derive users in idle mode to any other RAT, [122][123]. It is worth noting that, when a user is re-directed by JAC, an excessively long

admission delays can occur due to complicated signalling procedures (e.g., measurements from other radio technologies are required during the process). By contrast, IS-CR user reallocations take place when no data connection is ongoing, nor it is expected to start in a short time. Thus, a approach similar to that used in the FLC scheme in this thesis can be adapted to optimise IS-CR parameters. The main difference is that IS-CR parameters are defined on a per-cell basis, whereas IS-HO parameters are defined on a per-adjacency basis.

More interesting, it is expected that a joint management of several auto-tuning schemes carry extra benefits compared to modifications from separated and uncoordinated controllers, [124]. Then, several parameters from distinct JRRM algorithms could be jointly modified (e.g., parameters from JAC, IS-CR and IS-HO algorithms) from a higher level controller. Since several parameters from different algorithms are being modified simultaneously, high-level policies are first stated (e.g., "reduce blocking") and, then, translated to lower level actions ("increase parameter 1 from IS-HO algorithm and decrease parameter 2 from JAC algorithm"). To avoid contradictory actions when different parameters are modified, an important coordination and design effort is needed for the actions to be taken over the different JRRM algorithms.

5.3 List of Contributions

The following list enumerates the main results of this thesis.

Articles

[I] S. Luna Ramírez, M. Toril, M. Fernández Navarro and V. Wille , "Optimal Traffic Sharing in GERAN," *Wireless Personal Communications*, Springer. Published online (Nov. 17^{th}, 2009), DOI 10.1007/s11277-009-9861-6.

[II] S. Luna Ramírez, M. Toril and V. Wille, "Performance Analysis of Dedicated Signalling Channels in GERAN by retrial Queues," *Wireless Personal Communications*, Springer. Published online (Feb. 24^{th}, 2010), DOI 10.1007/s11277-010-9939-1.

Conferences and Workshops

[III] R. Barco, S. Luna Ramírez and M. Fernández Navarro "Optimisation and Troubleshooting of Heterogeneous Mobile Communication Networks," 3^{rd} *Workshop Trends in Radio Resource Management*, Barcelona (Spain), November, 2007.

[IV] S. Luna Ramírez, M. Toril, F. Ruiz and M. Fernández Navarro, "Adjustment of a Fuzzy Logic Controller for IS-HO parameters in a heterogenous scenario," in *Proc.*

IEEE 14th Mediterranean Electrotechnnical Conference (MELECON'2008), May, 2008, pp. 29–34.

[V] S. Luna Ramírez, M. Toril, F. Ruiz and M. Fernández Navarro, "Inter-system Handover Parameter Auto-Tuning in a Joint-RRM Scenario," in *Proc. IEEE 67th Vehicular Technology Conference* (Spring VTC'08), May, 2008, pp. 2641–2645.

[VI] S. Luna Ramírez, M. Toril, and V. Wille, "Balance de Carga Óptimo en GERAN," in *Proc. XXIV Simposium Nacional de la Unión Científica Internacional de Radio* (URSI 2009), Santander (Spain), September, 2009.

[VII] S. Luna Ramírez, M. Toril and V. Wille, "Performance Analysis of Dedicated Signalling Channels in GERAN," TD(09)805 in *Proc. of the 8th Management Committe Meeting* COST 2100 action, Valencia (Spain), May, 2009.

[VIII] S. Luna Ramírez, M. Toril, V. Wille and M. Fernández Navarro "Optimal Traffic Sharing in GERAN," TD(10)10022 in *Proc. of the 10th Management Committe Meeting*, COST 2100 action, Athens (Greece), Februar, 2010.

[IX] I. de la Bandera, S. Luna-Ramírez, R. Barco, M. Toril, F. Ruiz and M. Fernández-Navarro "Inter-system Cell Reselection Parameter Auto-Tuning in a Joint-RRM Scenario," in *Fifth International Conference on Broadband and Biomedical Communications* (IB2COM 2010), Málaga (Spain), December, 2010 (*accepted*).

[II, VII] are devoted to traffic modelling in dedicated signalling channels in GERAN. Similarly, [I, VI, VIII] are devoted to the definition of analytical teletraffic models and an optimal criterion for the traffic sharing problem in GERAN, while [III-V] describe an IS-HO parameter tuning algorithm in an heterogeneous environment, and [IX] describes a similar parameter auto-tuning scheme for the inter-system cell reselection algorithm in a similar scenario. The author has been the primary author of all the contributions except [III,IX], being there a key contributor. In [III], the author wrote the section dedicated to parameter optimisation in heterogeneous networks, and the sections defining the scenario description and performance analysis were his main contributions in [IX].

Several research projects have been involved in these contributions. [III-V] were developed in the frame of two different projects: the european project "*GANDALF: Monitoring and self-tuning of RRM parameters in a multisystem network*" (CELTIC-EUREKA european initiative), awarded with *Celtic Excellence Award*, and the grant TIC-4052, "*Técnicas adaptativas de gestión de recursos radio en redes B3G*", from the Junta de Andalucía. [II, VII] were presented in the frame of the TEC2008-06216 grant ("*Optimización de la estructura de redes de acceso radio heterogéneas mediante partición de grafos*") from the Spanish Ministry of Science and Technology, and [VIII,IX] was developed under an additional national grant (TEC2009-13413, "*Optimización automática de redes de comunicaciones móviles heterogéneas*").

Appendix A

Gaver Method in Retrial Queues

Chapter 2 introduced two retrial models for the SDCCH: Retrial Model (RM) and Retrial Model with Correlated Arrivals (RMCA). In such retrial queues, Gaver, Jacobs and Latouche's method, [42], can be used to compute the steady-state probabilities efficiently. This appendix outlines Gaver's method and its adaptation to retrial queues in this thesis. Similar to Chapter 2, the adaptation to RM is presented first and the adaption to RMCA is discussed later. This appendix follows a formulation and methodology similar to that in [35].

A.1 Retrial Model (RM)

In this section, Gaver's method is applied to the retrial system described in section 2.3.1 and depicted in Figure 2.8. For clarity, the state transition diagram is presented again in Figure A.1.

The total arrival rate for services with and without retrials are represented by λ_r and λ_{nr}, respectively. Other parameters are the service rate, μ, the retrial rate, α, the number of sub-channels, N, and the size of the orbit, M. The state of the system (i, j) is described by the number of busy SDCCH sub-channels, i, and the number of users in the orbit, j. In this appendix, the state of the system will be summarised by a single value, u, in the form

$$u = (i + j \cdot N) + 1 \qquad u \in \{1, 2, \ldots, N_s\}, \tag{A.1}$$

where N_s is the total number of states in RM (i.e., $N_s = (N + 1)(M + 1)$, as deduced from Figure A.1). Thus, the states of the system are ordered by a correlative numeration from top to down and left to right.

Teletraffic performance indicators for RM, (2.4)–(2.6), are obtained by computing the stationary distribution of the Markov chain describing system dynamics. This is achieved by solving the system of linear equations

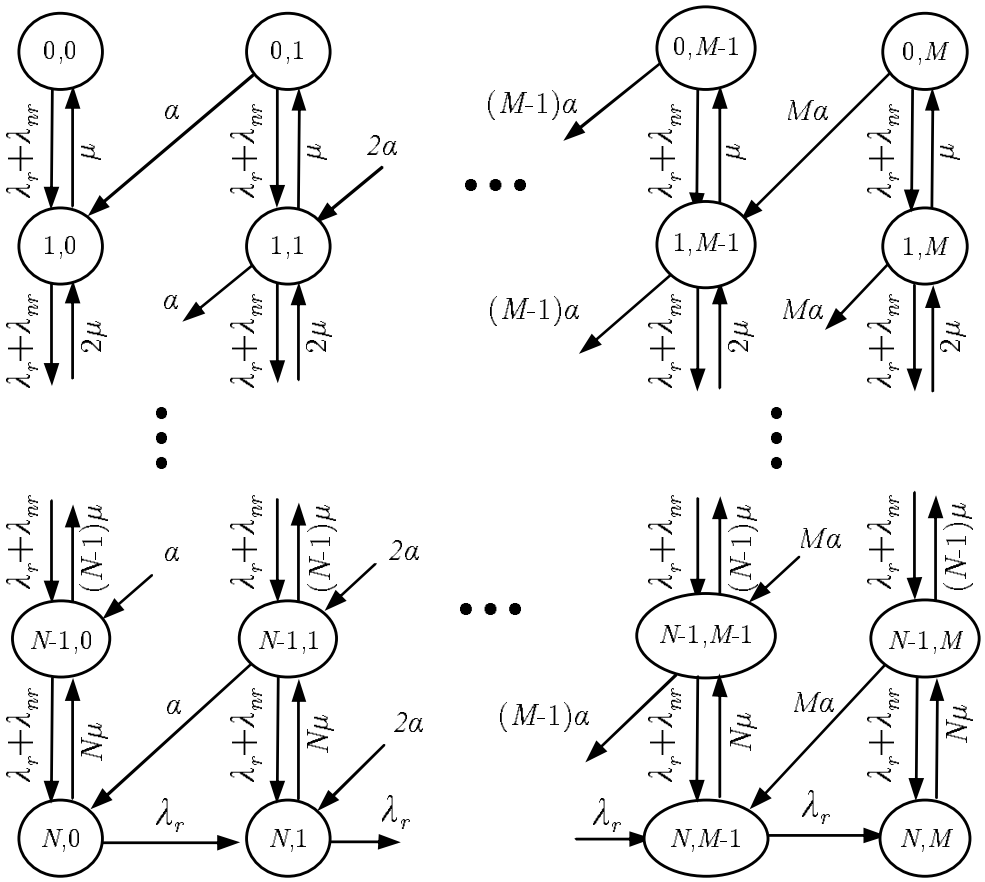


Figure A.1: State transition diagram of the retrial model.

$$\begin{aligned} \overline{\Pi}\,\mathbf{Q} &= 0 \quad , \\ \overline{\Pi}\,\overline{e} &= 1 \quad , \text{ and} \\ \overline{\Pi} &\geq 0 \quad , \end{aligned} \tag{A.2}$$

where $\overline{\Pi}$ is the steady-state probability vector, $\mathbf{Q}$ is the infinitesimal generator matrix including the Transition Rates (TR) between states, and $\overline{e}$ is a column vector of ones, i.e.,

$$\overline{\Pi} = [\pi(1)\ \pi(2)\ \ldots\ \pi(N_s-1)\ \pi(N_s)] \quad , \qquad \overline{e} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \\ 1 \end{bmatrix} \quad \text{and}$$

$$\mathbf{Q} = \begin{bmatrix} TR_{1,1} & TR_{1,2} & \cdots & TR_{1,N_s-1} & TR_{1,N_s} \\ TR_{2,1} & TR_{2,2} & \cdots & TR_{2,N_s-1} & TR_{2,N_s} \\ \ldots & \ldots & \ddots & \ldots & \ldots \\ TR_{N_s-1,1} & TR_{N_s-1,2} & \cdots & TR_{N_s-1,N_s-1} & TR_{N_s-1,N_s} \\ TR_{N_s,1} & TR_{N_s,2} & \cdots & TR_{N_s,N_s-1} & TR_{N_s,N_s} \end{bmatrix} \tag{A.3}$$

where $\pi(u)$ is the steady-state probability for the u-state. Additionally, $TR_{u,v}$ in $\mathbf{Q}$ matrix

values the transition rate from state u to state v. As an example, $TR_{1,2} = \lambda_r + \lambda_{nr}$ and $TR_{2,1} = \mu$ in RM, Figure A.1.

To solve (A.2), any classical method for lineal equation systems (e.g., Gauss-Jordan or Gaussian elimination) can be used. Alternatively, Gaver's method is a computationally efficient method to solve the linear equation system, provided that $\mathbf{Q}$ matrix has a tri-diagonal structure, i.e., $\mathbf{Q}$ can be expressed as

$$\mathbf{Q} = \begin{bmatrix} \mathbf{D_0} & \mathbf{U_0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{L_1} & \mathbf{D_1} & \mathbf{U_1} & \cdots & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{L_2} & \mathbf{D_2} & \cdots & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{D_{M-2}} & \mathbf{U_{M-2}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{L_{M-1}} & \mathbf{D_{M-1}} & \mathbf{U_{M-1}} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{L_M} & \mathbf{D_M} \end{bmatrix} \tag{A.4}$$

where $\mathbf{D_m}$, $\mathbf{L_m}$ and $\mathbf{U_m}$ are square matrices of $(N+1)$ dimensions and $m \in \{0, \ldots, M\}$. Vector $\overline{\Pi}$ can also be divided into $M+1$ sub-vectors with $N+1$ elements each. Such a $\mathbf{Q}$ structure can be achieved for RM model in Figure A.1 if these sub-matrices and vector are defined as

$$\overline{\Pi} = [\overline{\pi}_0\ \overline{\pi}_1\ \ldots\ \overline{\pi}_{M-1}\ \overline{\pi}_M] \quad ,$$

$$\mathbf{L_m} = \begin{bmatrix} 0 & m\alpha & 0 & \cdots & 0 \\ 0 & 0 & m\alpha & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & m\alpha \\ 0 & 0 & 0 & \cdots & 0 \end{bmatrix} \quad , \quad \mathbf{U_m} = \begin{bmatrix} 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 0 & \lambda_r \end{bmatrix} \quad \text{and}$$

$$\mathbf{D_m} = \begin{bmatrix} D_m^{(u_1)} & \lambda_r + \lambda_{nr} & 0 & \cdots & 0 & 0 & 0 \\ \mu & D_m^{(u_2)} & \lambda_r + \lambda_{nr} & \cdots & 0 & 0 & 0 \\ 0 & 2\mu & D_m^{(u_3)} & \ddots & 0 & 0 & 0 \\ \vdots & & \ddots & \ddots & & \vdots & \vdots \\ 0 & 0 & 0 & & D_m^{(u_{N-1})} & \lambda_r + \lambda_{nr} & 0 \\ 0 & 0 & 0 & \cdots & (N-1)\mu & D_m^{(u_N)} & \lambda_r + \lambda_{nr} \\ 0 & 0 & 0 & \cdots & 0 & N\mu & D_m^{(u_{N+1})} \end{bmatrix} . \tag{A.5}$$

where $\overline{\pi}_m$ sub-vectors contain $N+1$ consecutive state probabilities, $\pi(u)$, starting at $u = m \cdot (M+1) + 1$ until $u = m \cdot (M+1) + (N+1)$. The diagonal elements in $\mathbf{D_m}$ sub-matrices, $D_m^{(u)}$, are also located at the global $\mathbf{Q}$ matrix diagonal, (A.4). To fulfill the first equation in (A.2), $D_m^{(u)}$ values are calculated for each row as the negative value of the sum of all the values in the row as

$$D_m^{(u)} = -\sum_{v=1}^{N_s} \mathbf{Q_{u,v}} = -\sum_{v=1}^{N_s} TR_{u,v} \qquad \forall\, v \neq u \quad . \tag{A.6}$$

The previous formulation of $\mathbf{Q}$ allows to use specialised algorithms for solving (A.2). Once sub-matrices are defined, the following steps must be taken:

1. A first initialisation step configures the intermediate matrix $\mathbf{C_M}$:

$$\mathbf{C_M} = \mathbf{D_M} \tag{A.7}$$

2. Additional $\mathbf{C_m}$ matrices are obtained by successive iterative calculations:

$$\mathbf{C_m} = \mathbf{D_m} + \mathbf{U_m} * (-\mathbf{C_{m+1}}^{-1}) * \mathbf{L_{m+1}} \qquad \forall\, m = M-1, \ldots, 0\ . \tag{A.8}$$

3. Once all $\mathbf{C_m}$ matrices are known, state probabilities can be obtained. This will be done by $\overline{\pi}_m$ sub-vectors. First:

$$\overline{\pi}_0 \mathbf{C_0} = 0 \tag{A.9}$$

4. The others probability vectors are obtained by a direct recursion:

$$\overline{\pi}_m = \overline{\pi}_{m-1} \mathbf{L_{m-1}} (-\mathbf{C_l})^{-1} \qquad \text{for } \ m = 1, \ldots, M \tag{A.10}$$

5. As (A.2) indicates, $\overline{\Pi} \cdot \overline{e} = 1$, so a final normalisation is applied, i.e.,

$$\sum_{u=1}^{N_s} \pi(k) = 1 \quad . \tag{A.11}$$

Once (A.7)–(A.11) are finished, vector $\overline{\Pi}$ is constructed by concatenating $\overline{\pi}_l$ vectors. From state probabilities value, queueing performance indicators can be calculated for RM model.

The numerical complexity and stability of Gaver's approach is analysed in [42]. Numerical stability is ensured for square block matrixes and positive values in $\mathbf{Q}$, which is the case for RM. The complexity of solving (A.2) is $O(MN^3)$, where M is the size of the orbit and N the number of sub-channels. This complexity is similar to other block gaussian elimination methods (e.g., [125][126]) and much lower than classical Gauss-Jordan techniques, whose complexity is $O((MN)^3)$, [127].

A.2 Retrial Model with Correlated Arrivals (RMCA)

In this section, Gaver's method is applied to the Retrial Model with Correlated Arrivals described in section 2.3.2, and depicted in Figure 2.9. Again, for clarity, the state transition diagram is presented in Figure A.2.

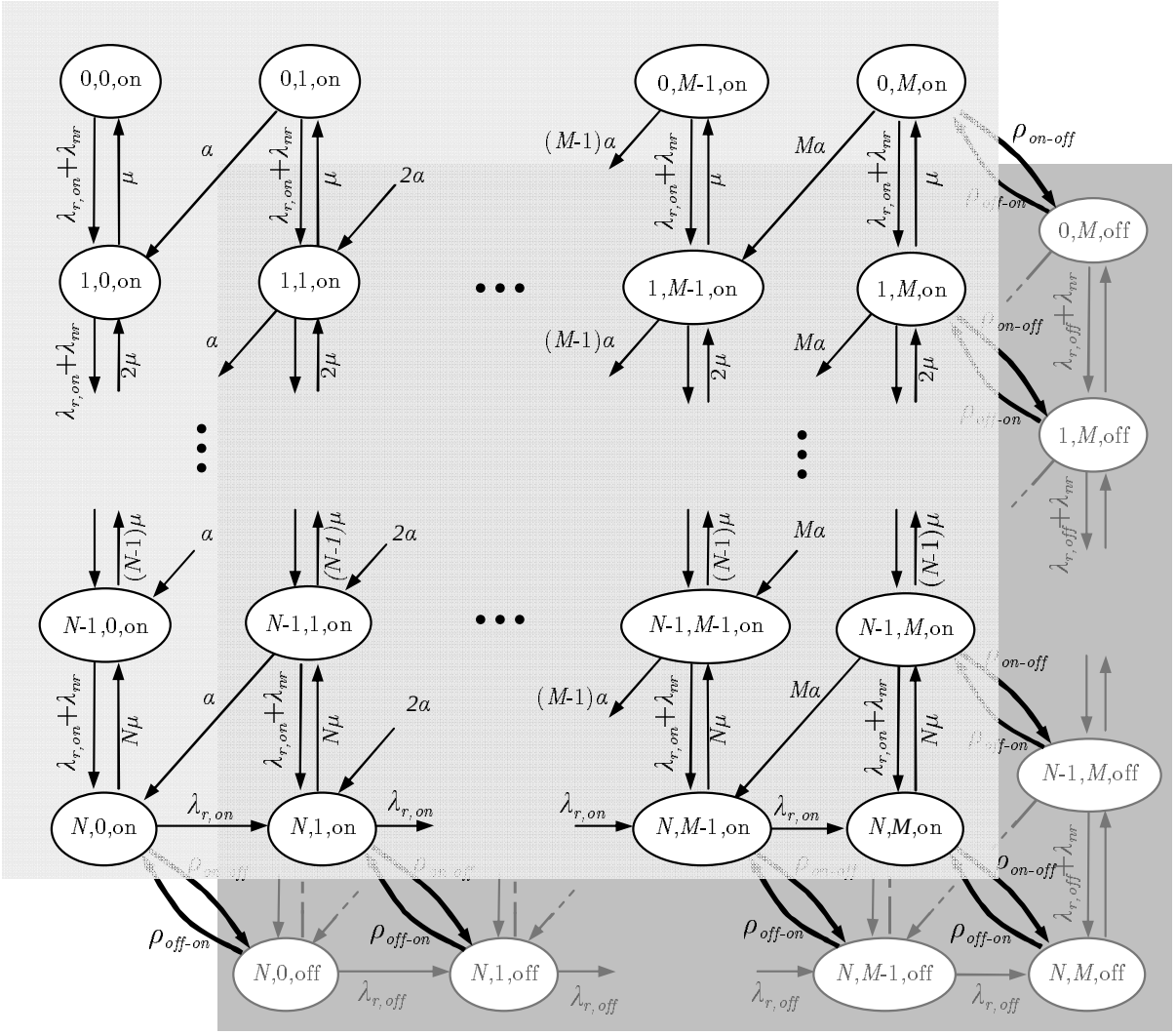


Figure A.2: State transition diagram for RMCA.

Figure A.2 presents the tri-dimensional state transition diagram, considering two states for the retrial traffic intensity, denoted as *on* and *off*. Thus, the state of the system (i, j, k) is described by the number of busy SDCCH sub-channels, i, the number of requests waiting for re-attempt, j, and the retrial activity state, k. The state of the system will be summarised by a single value, u, defined as

$$u = (N \cdot j + 2 \cdot i + w) + 1 \quad , \tag{A.12}$$

$$\text{where} \quad w = \begin{cases} 0 & \text{if } k = on \\ 1 & \text{if } k = off \end{cases} \quad , \tag{A.13}$$

N is the number of sub-channels, $i \in \{0, 1, \ldots, N\}$ and $j \in \{0, 1, \ldots, M\}$. Consequently, $u \in \{1, 2, \ldots, N_s\}$, where N_s is the number of states in the system, computed as $N_s = 2 \cdot (N+1)(M+1)$. As indicated by (A.12), the states of the system are numbered through the third dimension first and column-wise after, as shown in Figure A.2.

Similarly to RM, teletraffic performance indicators for RMCA, (2.8)–(2.10), are ob-

tained by computing first the stationary distribution of the Markov chain describing system dynamics. The system of linear equations described in (A.2) must be solved again. Matrix $\mathbf{Q}$ can again be defined by sub-matrices, as in (A.4). In RMCA, however, $\overline{\Pi}$, $\mathbf{D_m}$, $\mathbf{L_m}$ and $\mathbf{U_m}$ differ from the RM case, (A.5). In RMCA case, vector and sub-matrices are defined as

$$\overline{\Pi} = [\overline{\pi}_0\ \overline{\pi}_1\ \ldots\ \overline{\pi}_{M-1}\ \overline{\pi}_M] \quad ,$$

$$\mathbf{L_m} = \begin{bmatrix} 0 & 0 & m\alpha & 0 & \cdots & 0 \\ 0 & 0 & 0 & m\alpha & & 0 \\ \vdots & \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & & \ddots & m\alpha \\ 0 & 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 0 & \cdots & 0 \end{bmatrix} \quad , \quad \mathbf{U_m} = \begin{bmatrix} 0 & \cdots & 0 & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & \cdots & 0 & 0 & 0 \\ 0 & \cdots & 0 & \lambda_{r,on} & 0 \\ 0 & \cdots & 0 & 0 & \lambda_{r,off} \end{bmatrix} \quad \text{and}$$

$$\mathbf{D_m} = \begin{bmatrix} D_m^{(u_1)} & \rho_{on-off} & \lambda_{T,on} & 0 & 0 & \cdots & & 0 \\ \rho_{off-on} & D_m^{(u_2)} & 0 & \lambda_{T,off} & 0 & \cdots & & 0 \\ \mu & 0 & D_m^{(u_3)} & \rho_{on-off} & \lambda_{T,on} & \cdots & & 0 \\ 0 & \mu & \rho_{off-on} & D_m^{(u_4)} & 0 & \ddots & & 0 \\ 0 & 0 & 2\mu & 0 & D_m^{(u_5)} & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & 2\mu & \rho_{off-on} & \ddots & \ddots & \lambda_{T,off} \\ \vdots & \vdots & \vdots & & \ddots & \ddots & \ddots & 0 \\ 0 & 0 & 0 & \cdots & & N\mu & \rho_{off-on} & D_m^{(u_{2(N+1)})} \end{bmatrix} \quad (A.14)$$

$\mathbf{D_m}$, $\mathbf{L_m}$ and $\mathbf{U_m}$ are square matrices of $2 \cdot (N+1)$ dimension, and $m \in \{0, \ldots, M\}$. The steady-state probability vector, $\overline{\Pi}$, contains N_s elements, and it is divided into $(M+1)$ sub-vectors, $2 \cdot (N+1)$ elements long each. A $\overline{\pi}_m$ sub-vector contains $2 \cdot (N+1)$ consecutive steady-state probabilities, $\pi(u)$, starting at $u = 2m \cdot (M+1)+1$ until $u = 2 \cdot (m+1)(M+1)$. The diagonal elements in $\mathbf{D_m}$ sub-matrices, $D_m^{(u)}$, are also located at the global $\mathbf{Q}$ matrix diagonal, (A.4), and their values are calculated as in RM.

Once matrix $\mathbf{Q}$ is expressed as a tri-diagonal structure, Gaver's method can be applied as in (A.7)–(A.11). As a result, $\overline{\Pi}$ vector is calculated and model performance indicators can be obtained.

As in RM case, numerical stability is ensured for square block matrixes and positive values in $\mathbf{Q}$, which is the case for RMCA. The complexity of solving (2.3) is $O(M(2N)^3)$ for RMCA. As said for RM solution, this complexity is much lower than classical Gauss-Jordan techniques, whose complexity is $O((2MN)^3)$ for RMCA, [127].

Appendix B

Optimal Traffic Sharing Models

In this appendix, the optimality conditions for the two problem models described in Section 3.3 are derived. The naive model problem has been mainly extracted from [2]. In this work, the refined model problem is differently defined compared to [2], but a similar approach has been followed.

B.1 Naive Model

The traffic balance problem for the naive model shown in Section 3.3.1 is formulated as

$$\text{Minimise} \qquad \sum_{i=1}^{N} A_i \, E(A_i, c_i) \tag{B.1}$$

$$\text{subject to} \qquad \sum_{i=1}^{N} A_i = A_T \, , \tag{B.2}$$

$$A_i \geq 0 \qquad \forall \, i = 1 : N \, . \tag{B.3}$$

This problem has N independent variables $(A_1, A_2, \ldots, A_N)$, an objective function consisting of a sum of N non-linear terms, $A_i E(A_i, c_i)$, a linear equality constraint and N inequality constraints.

The convexity of the objective function in (B.1) with respect to A_i can be intuitively shown from the properties of the traffic overflowing term, $A_i E(A_i, c_i)$, which is known to be a convex function of A_i, [128]. Thus, the objective function consists of a sum of convex functions, which is also a convex function. Likewise, the feasible region defined by constraints (B.2) and (B.3) is a convex set[1], because it is the intersection of two convex sets. As both the objective function and the feasible region are convex, the problem is convex. Hence, any local minimum to the problem is a global minimum.

The problem can be re-formulated as an unconstrained optimisation problem. Firstly, it is assumed that constraint (B.3) is inactive at the optimum. Note that, once A_i is zero,

[1] In a *convex* set, the midpoint of any two points in the set is also a member of the set.

further decrements in the unconstrained problem have no effect on the overflowing term, $A_i E(A_i, c_i)$, but cause an increase of the other decision variables to maintain (B.2), which increases the value of the objective function. Hence, (B.3) can be eliminated without affecting the optimal solution. Secondly, (B.2) is eliminated by solving for one of the decision variables (e.g., A_N) and substituting in (B.1). As a result, the problem is re-formulated as

$$\text{Minimise} \qquad \sum_{i=1}^{N-1} A_i\, E(A_i, c_i) \;+\; \left(A_T - \sum_{i=1}^{N-1} A_i\right) E\left(A_T - \sum_{i=1}^{N-1} A_i,\; c_N\right). \tag{B.4}$$

In such an unconstrained problem, the optimal solution must satisfy the stationary condition

$$\nabla A_{bT} = \left(\frac{\partial A_{bT}}{\partial A_1}, \frac{\partial A_{bT}}{\partial A_2}, \cdots, \frac{\partial A_{bT}}{\partial A_{N-1}}\right) = 0 \tag{B.5}$$

(i.e., the gradient of the objective function in the optimum must be 0). The latter equation can be developed further by derivating (B.4) with respect to the decision variables, A_j. This operation results in a set of (N-1) equations

$$\begin{aligned}\frac{\partial A_{bT}}{\partial A_j} \;=\; & E(A_j, c_j) + A_j \frac{\partial E(A_j, c_j)}{\partial A_j} \;- E\left(A_T - \sum_{i=1}^{N-1} A_i,\; c_N\right) \\ & + \left(A_T - \sum_{i=1}^{N-1} A_i\right) \frac{\partial E\left(A_T - \sum_{i=1}^{N-1} A_i,\; c_N\right)}{\partial A_j} \;=\; 0 \qquad \forall\, j = 1 : (N-1),\end{aligned} \tag{B.6}$$

which can be re-written as

$$\begin{aligned}E(A_j, c_j) + A_j \frac{\partial E(A_j, c_j)}{\partial A_j} \;=\; & E\left(A_T - \sum_{i=1}^{N-1} A_i,\; c_N\right) \\ & - \left(A_T - \sum_{i=1}^{N-1} A_i\right) \frac{\partial E\left(A_T - \sum_{i=1}^{N-1} A_i,\; c_N\right)}{\partial A_j} \quad \forall\, j = 1 : (N-1).\end{aligned} \tag{B.7}$$

For symmetry reasons,

$$\frac{\partial E\left(A_T - \sum_{i=1}^{N-1} A_i,\; c_N\right)}{\partial A_j} = \frac{\partial E\left(A_T - \sum_{i=1}^{N-1} A_i,\; c_N\right)}{\partial A_k} \qquad \forall\, j, k \tag{B.8}$$

and the right-hand side of (B.7) is equal $\forall\, j = 1 : (N-1)$. Thus, the left-hand side of (B.7) is also equal $\forall\, j = 1 : (N-1)$ and the optimality conditions can be re-formulated as

$$E(A_i, c_i) + A_i \frac{\partial E(A_i, c_i)}{\partial A_i} = E(A_j, c_j) + A_j \frac{\partial E(A_j, c_j)}{\partial A_j} \qquad \forall\, i, j = 1 : N \tag{B.9}$$

and

$$\sum_{i=1}^{N} A_i = A_T \,. \tag{B.10}$$

Note that, in the latter equations, i and j have been extended to N for symmetry reasons (i.e., the solution should be the same, regardless of the eliminated decision variable). Likewise, (B.10) is needed to avoid the trivial solution of (B.9) $A_1 = A_2 = \cdots = A_N = 0$. (B.9) becomes in the optimal traffic balance condition for any solution point (i.e., A_i offered traffics) fulfilling (B.10).

From (B.9), it can be concluded that balancing the blocking probability, $E(A_i, c_i)$, would not lead to the optimal solution, unless the second terms on both sides of the equality were also equal in these conditions. To discard the latter, (B.9) can be developed by using the definition of the incremental blocking probability in (3.11). Thus, (B.9) is converted into

$$E(A_i, c_i)\,[1 + N_{fc}(A_i, c_i)] = E(A_j, c_j)\,[1 + N_{fc}(A_j, c_j)]\,. \tag{B.11}$$

It is well known that the average number of free channels (or, conversely, the average number of busy channels) is not the same for two cells with the same blocking probability but different number of channels. Hence, it is clear that forcing $E(A_i, c_i) = E(A_j, c_j)$ does not ensure that $N_{fc}(A_i, c_i) = N_{fc}(A_j, c_j)$, and it can be concluded that balancing the blocking probability does not lead to the optimal solution.

B.2 Refined Model

The traffic balance problem for the refined model, shown in Section 3.3.2, is formulated as

$$\text{Minimise} \qquad \sum_{i=1}^{N} \frac{\lambda_{f_i}}{\mu} E(A_i, c_i) \quad \text{or} \quad \sum_{i=1}^{N} \lambda_{f_i} E(A_i, c_i) \tag{B.12}$$

$$\text{subject to} \qquad \sum_{i=1}^{N} A_{fi} \left(1 - E(A_i, c_i)\right) = \sum_{i=1}^{N} A_i \left(1 - E(A_i, c_i)\right), \tag{B.13}$$

$$A_{lbi} \leq A_i \leq A_{ubi} \qquad \forall\, i = 1 : N, \tag{B.14}$$

where A_{lb_i} and A_{ub_i} are lower and upper bounds for the offered traffic in cell i. The naive model approach cannot be used to solve (B.12)-(B.14), since (B.14) cannot be eliminated as these constraints may be active at the optimum. Hence, the problem must be solved as an optimisation problem with inequality constraints, for which the *Karush-Kuhn-Tucker* (KKT) *multiplier method*, [129], can be used. The KKT method builds the Lagrangian function as a combination of the objective and constraints functions. For (B.12)-(B.14), the Lagrangian is

$$\begin{aligned} \Phi(\mathbf{A}, \phi, \mathbf{u}, \mathbf{z}) = &\sum_{i=1}^{N} A_{f_i} E(A_i, c_i) + \phi \sum_{i=1}^{N} (A_{f_i} - A_i)(1 - E(A_i, c_i)) \\ &+ \sum_{i=1}^{N} u_i (A_{lbi} - A_i) + \sum_{i=1}^{N} z_i (A_i - A_{ubi}), \qquad u_i, z_i \geq 0\,, \end{aligned} \tag{B.15}$$

where ϕ, u_i and z_i are the Lagrange multipliers associated to (B.13) and (B.14), [130].

The Lagrangian has the property that its stationary points are potential solutions to the constrained problem. Consequently, the optimality conditions can be derived by setting the gradient of the Lagrangian equal to zero. In a problem with inequalities, these necessary conditions are referred to as *KKT conditions*. If the problem is convex, as the one considered here, KKT conditions are also sufficient for optimality. For (B.12)-(B.14), the KKT conditions are

$$A_{f_i} \frac{\partial E(A_i, c_i)}{\partial A_i} - \phi \left(1 - E(A_i, c_i) + (A_{f_i} - A_i) \frac{\partial E(A_i, c_i)}{\partial A_i}\right) - u_i + z_i = 0\,, \tag{B.16}$$

$$u_i (A_{lbi} - A_i) = 0\,, \tag{B.17}$$

$$z_i (A_i - A_{ubi}) = 0\,, \tag{B.18}$$

$$\sum_{i=1}^{N} (A_{f_i} - A_i)(1 - E(A_i, c_i)) = 0\,, \tag{B.19}$$

$$u_i, z_i \geq 0\,, \tag{B.20}$$

$\forall\, i = 1 : N$. In (B.16), it has been used that A_{f_i} does not depend on A_i when computing the Lagrangian partial derivative. Note that A_i is modified by tuning HOC settings, while A_{f_i} is fixed by AC parameters, which remain unchanged.

The solution to (B.16)-(B.20) is the optimal solution, since the problem is convex. Unfortunately, these set of equations does not give any information about the values of ϕ, u_i and z_i. Alternatively, (B.16), (B.17), (B.18) and (B.20) can be re-formulated in a more convenient way. For convenience, let β be defined as

$$\beta(A_i, A_{f_i}, c_i) = \frac{A_{fi}\frac{\partial E(A_i,c_i)}{\partial A_i}}{1 - E(A_i, c_i) + (A_{f_i} - A_i)\frac{\partial E(A_i,c_i)}{\partial A_i}} \; . \tag{B.21}$$

From (B.17) and (B.18), it can be deduced that u_i and z_i must be zero when A_i is different from A_{lb} and A_{ub}, respectively. Thus, the values of u_i and z_i reflect whether the inequality constraints (B.14) are active or not in the optimal solution. In addition, it can easily be deduced (although not shown here) that $\left(1 - E(A_i, c_i) + (A_{f_i} - A_i)\frac{\partial E(A_i,c_i)}{\partial A_i}\right) \geq 0$. Therefore, it follows from (B.16) and (B.21) that:

a) If $A_{lb} < A_i < A_{ub}$ then $u_i = z_i = 0$, and

$$\beta(A_i, A_{fi}, c_i) = \phi. \tag{B.22}$$

b) If $A_i = A_{lb}$ then $u_i \geq 0$, $z_i = 0$, and

$$\beta(A_i, A_{fi}, c_i) = \phi + \frac{u_i}{1 - E(A_i, c_i) + (A_{f_i} - A_i)\frac{\partial E(A_i,c_i)}{\partial A_i}} \geq \phi. \tag{B.23}$$

c) If $A_i = A_{ub}$ then $u_i = 0$, $z_i \geq 0$, and

$$\beta(A_i, A_{fi}, c_i) = \phi - \frac{z_i}{1 - E(A_i, c_i) + (A_{f_i} - A_i)\frac{\partial E(A_i,c_i)}{\partial A_i}} \leq \phi. \tag{B.24}$$

As ϕ is a constant, it can be deduced from (B.22) that

$$\beta(A_i, A_{fi}, c_i) = \beta(A_j, A_{fj}, c_j) \tag{B.25}$$

$\forall\, i, j$ where constraint (B.14) is inactive (i.e., $A_{lbi} < A_i < A_{ubi}$). Likewise, from (B.23) and (B.24), it follows that

$$\beta(A_u, A_{fu}, c_u)|_{A_u = A_{ubu}} \leq \beta(A_i, A_{fi}, c_i) \leq \beta(A_l, A_{fl}, c_l)|_{A_l = A_{lbl}} \tag{B.26}$$

$\forall\, l, u$ where constraint (B.14) is active due to the lower and upper bound, respectively. Thus, the KKT conditions in (B.16)-(B.20) can be substituted by (B.25), (B.26) and (B.19).

Appendix C

Summary (Spanish)

Este apéndice presenta un amplio resumen en español del trabajo realizado en esta tesis. Una primera sección introduce el escenario de trabajo en el que se enmarcan las contribuciones realizadas. Posteriormente, se presentan los objetivos de investigación y el estado actual de la investigación y la tecnología relacionadas con este trabajo. Por último, se enumeran las principales conclusiones, junto con una lista de publicaciones asociadas a esta tesis.

C.1 Introducción

En los últimos años, el éxito de los servicios de comunicaciones móviles ha provocado un crecimiento exponencial del tráfico en las redes de telefonía móvil. Dicho crecimiento aleja el estado actual de la red de las condiciones iniciales para las que fue diseñada. Al mismo tiempo, la red evoluciona tecnológicamente, incluyendo cada vez más funcionalidades y haciéndose más compleja. Durante el diseño de una red móvil, se tienen en cuenta diversos modelos que incluyen previsiones en el incremento permanente de tráfico, pero, sobre todo, intentan reflejar de la manera más fiel posible el comportamiento de los distintos elementos de la red: canal de propagación, patrones de movimiento de usuario, peticiones de servicio, etc. El uso de buenos modelos permite que la red no sólo funcione adecuadamente, sino que lo haga con prestaciones óptimas No obstante, existen dos posibles fuentes de disfunción en este proceso de diseño: a) el empleo de modelos inadecuados, que no reflejen la realidad, y, b) la variación de las condiciones de diseño originales con las que se dimensionó la red.

En el primer caso, los modelos empleados para dimensionar la red móvil hacen uso de una serie de suposiciones sobre el tráfico de la red y la gestión de recursos radio, con la esperanza de que se ajusten al comportamiento real del sistema. Sin embargo, si algunas suposiciones son excesivamente restrictivas, el modelo de la red no será el adecuado, y, por tanto, la red diseñada estará lejos de su rendimiento óptimo. El segundo caso corresponde a un escenario con una red de comunicaciones móviles madura en su implantación. La red evoluciona tecnológicamente conforme el tiempo pasa, e igual hace el tráfico generado en la misma. Cuanto más lejos estemos en el tiempo respecto al momento de su diseño inicial, más probable es que el diseño de la red no sea adecuado para las características

actuales del tráfico.

Surgen así las estrategias de optimización de la red. El término optimización se utiliza aquí en un sentido amplio, e incluye todas aquellas técnicas que mejoran el rendimiento de la red existente. Estas estrategias pueden dividirse en técnicas de *replanificación* o de *reconfiguración*, según el objeto de modificación en la red. Por *replanificación* se entiende toda estrategia y método que cambie características más o menos estables en la red y, por ello, son procedimientos que se ejecutan con poca frecuencia. Las principales estrategias de re-planificación suelen centrarse en un nuevo reparto o incremento de los recursos radio que se ajuste a unas nuevas condiciones de trabajo de la red (bien por cambio de modelo o por evolución en la demanda de servicios). Por *reconfiguración* de redes se entienden aquellos métodos que buscan cambiar valores de ciertos parámetros de red, sin que haya que modificar ningún algoritmo o estructura estable de la misma. Dado que no modifican la estructura de la red, las estrategias de reconfiguración se ejecutan de manera más frecuente que las de replanificación.

En cualquiera de los escenarios para optimizar la red, es necesaria la construcción de modelos precisos para diseñar las estrategias a aplicar (p.ej., cómo redistribuir los recursos radio, o con qué criterio reconfigurar un parámetro). El modelado de redes de comunicaciones móviles puede dividirse en dos grandes categorías: a) modelado analítico basado en teoría de teletráfico, y b) modelado estadístico basado en simuladores. El primero hace uso de la teoría de teletráfico para resumir el funcionamiento de la red. Los modelos basados en teletráfico toman en cuenta un mayor número de simplificaciones a la hora de construir el modelo (p.ej., cómo se mueven los usuarios o con qué frecuencia solicitan servicios). A favor tienen que es posible obtener expresiones cerradas matemáticas de sus indicadores de rendimiento. Estas expresiones analíticas posibilitan la aplicación de técnicas clásicas de optimización matemática sobre el modelo del sistema, y, por tanto, permiten extraer estrategias óptimas de configuración de la red móvil real. En cambio, el modelado basado en simuladores refleja el funcionamiento de la red con mayor extensión, implementando un mayor número de funciones de la red y, por tanto, con mayor capacidad para modelar sistemas complejos como son las redes de comunicaciones móviles. Los simuladores de red, por contra, son de difícil construcción y manejo, y la obtención de resultados de indicadores de rendimiento necesita de mayor esfuerzo y tiempo, tanto de computación como de análisis de resultados.

Con el objetivo de reducir los costes de operación, los operadores tienden a automatizar cada vez más la gestión de la red y las estrategias de optimización de red. La automatización posibilita el diseño de algoritmos que cambien de manera autónoma alguna de las características de la red con el objetivo de optimizar su rendimiento (se habla entonces de redes auto-organizadas o auto-ajustables). Estos algoritmos generalmente modifican alguno(s) de los parámetros a los que la red es especialmente sensible en su funcionamiento. En la mayoría de los casos, los operadores se centran en parámetros de los algoritmos de gestión de recursos radio (*Radio Resource Management*, RRM) por su facilidad de control. En las redes heterogéneas, donde varias tecnologías radio conviven en un mismo área geográfica, las posibilidades de los esquemas de reconfiguración

automática de parámetros se multiplican. En estos nuevos escenarios, las entidades conjuntas de gestión de red deben manejar servicios, indicadores y medidas de muy distinta naturaleza. Esto provoca situaciones muy cambiantes en el tiempo que, por tanto, hacen muy necesarios los esquemas de auto-ajuste de parámetros.

Por lo general, el diseño de algoritmos de ajuste de parámetros de red intenta traducir el conocimiento del personal técnico del operador. Este tipo de diseño se basa en la experiencia adquirida durante años de trabajo, sin ninguna clase de prueba o demostración de que las técnicas empleadas sean las óptimas desde un punto de vista matemático. En cambio, el empleo de modelos de red (especialmente los analíticos) permite la obtención de estrategias óptimas de ajuste de parámetros a través del uso de técnicas de optimización. Se consigue así un rendimiento óptimo del sistema.

En esta tesis se trabaja en varios de los escenarios comentados anteriormente, tratando así de dar una perspectiva amplia. Por un lado, se construyen modelos de teletráfico y modelos basados en simulador; por otro lado, se definen estrategias tanto de replanificación como de configuración automática de parámetros; se diseñan estrategias óptimas y se adoptan otras basadas en la experiencia del operador; y, por último, las diversas técnicas empleadas se aplican en redes de una o múltiples tecnologías.

C.2 Objetivos

El principal objetivo de esta tesis es la elaboración de técnicas de replanificación y auto-ajuste de parámetros mediante modelos analíticos y de simulación para mejorar el rendimiento de una red de comunicaciones móviles. Este objetivo general se desglosa en los siguientes objetivos más específicos:

a) Elaborar un modelo analítico del tráfico en canales de señalización dedicados en GERAN (*GSM-EDGE Radio Access Network*), que permita desarrollar estrategias de replanificación de los canales radio de la red.

b) A partir de modelos de teletráfico, definir un criterio óptimo de reparto de carga de tráfico entre celdas de una red GERAN mediante la modificación de parámetros RRM para solucionar problemas locales de congestión en la red.

c) Diseñar un algoritmo heurístico de reparto de carga de tráfico en una red de múltiples tecnologías a través de la modificación de parámetros RRM conjunto (*Joint RRM*, JRRM) a verificar mediante modelos basados en simulador.

C.3 Estado Actual

En esta sección se describe el estado actual de la investigación y la tecnología en cada uno de los campos abordados. Para mayor claridad, los distintos objetivos de esta tesis

se tratan de forma separada.

a) Modelado de canales de señalización dedicados en GERAN

El análisis del rendimiento de los canales de señalización dedicados en tecnología GERAN (conocidos como *Stand-alone Dedicated Control CHannels*, SDCCH) se realiza con modelos de teletráfico. La teoría de teletráfico aplica la teoría de la probabilidad a los sistemas de telecomunicaciones. Inicialmente, el principal campo de aplicación del teletráfico fue el modelado de tráfico telefónico, junto con el diseño y dimensionamiento de redes de telefonía fija. La ecuación de Erlang B, (2.1), que es su principal exponente, permite estimar el número de canales necesarios para obtener un cierto grado de servicio, bajo ciertas suposiciones sobre el tráfico y la red que modela (tasa de llegadas según distribución de Poisson, no existe correlación ni reintentos en el tráfico, el tiempo de servicio posee distribución exponencial y el número de usuarios es infinito).

En el ámbito de las redes de comunicaciones móviles, existen numerosas referencias vinculadas al modelado de teletráfico. Una referencia clásica en el modelado de redes celulares es el trabajo de Hong y Rappaport, [16]. El modelo de red celular que proponen incluye tráfico proveniente de nuevas llamadas y traspasos. En ese trabajo los autores realizan un análisis de varios esquemas de prioridad para usuarios con llamadas en curso, suponiendo un sistema de pérdidas (esto es, un sistema donde las llamadas con petición de acceso denegada se pierden, sin entrar en estado de espera). [17] extiende el modelo de Hong y Rappaport con una distribución generalizada del tiempo de permanencia en una celda.

De igual manera, [8][37][9] analizan el fenómeno del reintento en llamadas y analizan su impacto en el rendimiento de las redes. Es conveniente destacar la referencia [8], en donde se define un modelo con reintentos. Los usuarios pueden entrar en una órbita, entendiendo como tal el estado en el que entra un usuario que va a volver a solicitar en breve espacio de tiempo el servicio que anteriormente le fue denegado. Siguiendo en el modelado de tráfico, [43][65] tratan el problema de llamadas múltiples y correladas en la entrada de nuevas conexiones por traspaso. Las ideas expuestas en estas referencias pueden ser trasladadas al problema de correlación entre peticiones de actualización de localización en el canal de señalización, Figura 2.5.

Debido a las especiales características que posee el tráfico de señalización, la mayoría de los modelos de teletráfico usados para el tráfico de datos tienen importantes carencias cuando se trasladan al tráfico de señalización. Por ello, el uso de estos modelos para el dimensionamiento de los canales de señalización requiere ser validado, lo que no se ha realizado hasta la fecha. La principal diferencia está en la existencia de reintentos y la correlación temporal entre intentos. Un gran número de referencias analizan el fenómeno del reintento en redes fijas e inalámbricas. En [28] ó [29] se presenta un análisis exhaustivo de colas de reintento, en el que las llamadas que inicialmente no son aceptadas entran en un estado de espera también denominado órbita, aunque referencias anteriores como [30] y [31] analizan el efecto de los reintentos en escenarios particulares. En [32] ó [33] se análiza el rendimiento de un sistema clásico de colas con múltiples servidores, reintento, peticiones

de servicio según Poisson y tiempo entre intentos según una distribución exponencial.

Para facilitar el análisis del sistema, es interesante encontrar expresiones analíticas que describan el funcionamiento de los sistemas con reintento. Estas soluciones analíticas sólo son posibles cuando el número de servidores es bajo, [29], lo cual no es habitual en los canales SDCCH (los datos del presente trabajo muestran un 79% de las celdas con 7 o más servidores, cursando el 83% del tráfico). Por esta razón, se requieren métodos numéricos eficientes a la hora de solucionar las ecuaciones que describen el funcionamiento del sistema. Una de las técnicas más difundidas y analizadas consiste en limitar el tamaño de la órbita a un número determinado de usuarios, asumiendo que el comportamiento de este sistema limitado apenas cambia con respecto a la eliminación de dicha restricción. Estos métodos de *cola truncada* se estudian ampliamente en [34] y [35].

En el contexto de redes de comunicaciones móviles, otras referencias incluyen características adicionales en los sistemas que proponen, como considerar traspasos, [8][37][38], reintentos por parte del terminal y usuario, [7][39], así como distintas distribuciones de probabilidades para modelar el tiempo entre reintentos, el tiempo de servicio o la llegada de nuevos usuarios, [9][38]. En [36], el análisis de rendimiento introduce la correlación temporal entre nuevas peticiones de servicio en colas con reintento, característica que es útil para el trabajo en este apartado de la tesis.

Aun así, hasta la fecha no se conocen medidas del tráfico de sealización sobre redes GERAN reales que demuestren la validez de los modelos usados para su dimensionamiento, que habitualmente incluyen suposiciones muy restrictivas. En esta tesis se formulan modelos de tráfico de señalización que consideran reintento y correlación. Este trabajo utiliza los conceptos presentados en [8] acerca de reintento de llamadas, para ser ampliados posteriormente con características de la correlación entre usuarios.

b) Reparto de tráfico óptimo en GERAN

El problema del reparto de tráfico en las redes móviles abarca dos grandes áreas en la literatura: el modelado de teletráfico y la optimización de parámetros de red sobre modelos. Como en el apartado anterior, [16] es una referencia básica en los modelos de redes de teletráfico. Hong y Rappaport proponen un modelo de red con prioridad para conexiones de traspaso y nuevas llamadas según Poisson y para tecnologías FDMA/TDMA (*Frequency/Time Division Multiple Access*). A partir de dicho modelo, se extraen diversas ecuaciones de indicadores que se usan para analizar el rendimiento del sistema reflejado en dichas ecuaciones. Sucesivas referencias han extendido el modelo de Hong y Rappaport añadiendo características como tiempos de espera en conexiones entrantes, [61], distintos flujos de tráfico, [62][63], o distribuciones de probabilidad más generales de los tiempos de permanencia en una celda, [64]. En todas estas referencias, la metodología es similar: primero se extraen indicadores de rendimiento del modelo, y después, a través de esos indicadores, se analiza dicho sistema modelado.

A partir de ahí, se incluyen más capacidades a los modelos de redes de comunicación móviles, como la correlación entre peticiones de servicio y los reintentos de llamada, [43][8],

o el tráfico multi-servicio. Con el éxito de las redes móviles, aparecen nuevas estructuras de red que también se ven reflejadas en los correspondientes modelos. Así aparecen modelos de red multicapa, o red jerárquica, donde conviven macroceldas y micro- o picoceldas como solución a problemas locales de congestión permanente. [21][24] proponen modelos para el análisis y estudio de estas redes jerárquicas. Con la aparición de nuevas tecnologías radio, en la bibliografía aparecen modelos de red para CDMA (*Code Division Multiple Access*), [67][68], y OFDMA (*Orthogonal Frequency Division Multiple Access*), [69].

Las referencias anteriores proponen diversos esquemas de red mayoritariamente con el objetivo de evaluar el rendimiento de distintas propuestas RRM o la sensibilidad de la red ante ciertos parámetros. Sin embargo, no es habitual encontrar una metodología que use el modelo de la red móvil como herramienta para optimizar la propia configuración de la red. En [70], se diseña una red móvil multi-capa y sobre ese modelo se define un problema de optimización clásico para maximizar el rendimiento de la red. En dicho problema de optimización las variables de decisión son los tamaños de celda y el número de canales por capa. Con una metodología similar, [71] define un esquema de reserva de recursos radio en un escenario multi-servicio, y consigue la configuración óptima del número mínimo de canales para cada flujo de tráfico. Más cercano al trabajo desarrollado en esta tesis, [57][58] formulan el problema del reparto de carga entre celdas adyacentes como un problema de optimización, donde se minimiza el ratio de bloqueo global de la red y las variables de decisión son los márgenes de traspaso. En [72] se define un modelo analítico de WCDMA (*Wideband CDMA*) para minimizar la interferencia en el enlace descendente a través de las orientaciones de las antenas. No obstante, ninguna de las referencias anteriores proponen expresiones analíticas cerradas como solución óptima para el problema del reparto de tráfico.

La principal contribución en este apartado es la definición de un criterio óptimo para el reparto de tráfico. A diferencia de los criterios conocidos de reparto de carga, que son de naturaleza heurística, en este trabajo se define un modelo de red y se extraen indicadores de rendimiento globales analíticos. A partir de estos indicadores de rendimiento, y siguiendo un proceso de optimización clásico, se ha definido un indicador de balance entre celdas óptimo.

c) Ajuste de parámetros de traspaso en un escenario conjunto GSM/UMTS

Dentro de los escenarios multi-tecnología (también llamados *heterogéneos*), el diseño de algoritmos JRRM ha sido tratado y analizado de forma intensa. En una primera fase, la literatura se centra en la definición de topologías y entidades de red necesarias para poder ejecutar funcionalidades JRRM, [25][91]. La organización 3GPP define distintos grados de cooperación en un escenario heterogéneo incluyendo redes de acceso local inalámbricas (*Wireless Local Access Network*, WLAN), [92]; y [93] establece distintos esquemas JRRM dependiendo de dicha cooperación. Posteriormente a estas primeras definiciones, han surgido multitud de propuestas de algoritmos JRRM. Los principales algoritmos tratados en la bibliografía son el traspaso entre sistemas (*Inter-System HandOver*, IS-HO), el control de admisión conjunto (*Joint Admission Control*, JAC) y la selección de tecnología.

Así, [94][88][131] definen un escenario con características simples de multi-tecnología sobre el que analizan el rendimiento de métodos IS-HO y selección de tecnología. Algo más elaboradas, [96][27][86] introducen características de movilidad, terminal, tráfico y red más avanzadas para la evaluación de algoritmos JAC e IS-HO. [89] introduce el análisis del impacto de ciertos parámetros de algoritmos JRRM, concretamente los contadores temporales en el algoritmo IS-HO, sobre el rendimiento global de la red. Otras referencias extraen modelos analíticos del escenario heterogéneo, basados en cadenas de Markov, para evaluar las ventajas del uso de algoritmos JRRM, [26][95].

Las referencias anteriores implementan esquemas JRRM basados en reglas, los cuales son fáciles de definir e implementar, pero poseen cierta rigidez a la hora de adaptarse a las condiciones cambiantes de la red. Los esquemas basados en controladores de lógica difusa (*Fuzzy Logic Controllers*, FLC) implementan métodos de auto-ajuste de parámetros en la red, superando las limitaciones de los esquemas basados en reglas, [90]. El éxito de este esquema difuso radica en la facilidad que provee a la hora de trasladar el conocimiento humano en reglas, y así tratar de manera automática problemas complejos que no poseen formulación analítica, como es es habitual en redes de comunicaciones móviles. Además, los FLCs pueden manejar, comparar y tomar decisiones a partir de información de muy distinta naturaleza, como la proveniente de redes de acceso distintas. Dentro de los escenarios de tecnología simple, los FLCs han sido ampliamente usados en tecnologías de acceso 3G debido a la complejidad y flexibilidad de las técnicas RRM en UMTS, especialmente el *soft-handover*, [98][99][100], control de potencia, [101][102][103], o el control de admisión, [104]. En escenarios heterogéneos, el esquema FLC es usado para decisiones en algoritmos JRRM, esto es, para algoritmo no basados en reglas, [105][106], y para la modificación de parámetros en algoritmos JAC y IS-HO, [104][107][108].

Para que el rendimiento de la red móvil no se resienta de manera significativa ante las condiciones cambiantes del tráfico, los FLCs deben incluir mecanismos que los hagan adaptarse ante cada nueva situación, [109]. En escenarios heterogéneos, [110][111] describen un esquema basado en lógica difusa para JRRM que tiene en cuenta consideraciones económicas y preferencias del usuario para la toma de decisiones. [112][113] usan esquemas difusos adaptativos con un enfoque centralizado y distribuido, respectivamente. Con especial interés para esta tesis, [114] implementa un mecanismo de balance de carga con un esquema basado en FLC con características cambiantes en un escenario WLAN/UMTS. Dicha referencia analiza las mejoras en el rendimiento de la red a través de modificaciones en la configuración del controlador que rige el algoritmo IS-HO.

Las contribuciones anteriores usan esquemas basados en FLC para la modificación de parámetros de red móvil. En esta tesis, se importa este esquema a un escenario multi-radio para modificar parámetros de calidad y nivel en el algoritmo de IS-HO con propósitos de balance de carga. Adicionalmente con respecto a otros trabajo, se desea encontrar la mejor configuración (esto es, la de adaptación más rápida) del propio FLC. Por esta razón se han construido escenarios con distribuciones muy desiguales de tráfico, espacial y temporalmente, y probar de esta manera la capacidad de adaptación del FLC. Los esquemas propuestos se prueban en un simulador de red conjunta. Con la intención

de obtener una alta exactitud, y a diferencia de muchas de las referencias anteriores, la plataforma de simulación incluye gran parte de las características de red.

C.4 Resultados

En esta sección se presentan los principales resultados alcanzados en el trabajo desarrollado en esta tesis. Como en anteriores secciones, los resultados se enumerarán en distintos apartados según el problema abordado.

a) Modelado de tráfico en canales de señalización dedicados en GERAN

En este problema, se han propuestos dos modelos distintos de teletráfico que caracterizan el tráfico de señalización en canales dedicados de red GERAN. Estos modelos mejoran modelos clásicos considerando características específicas del tráfico de señalización, que no han sido consideradas hasta la fecha.

El primer modelo, denominado Modelo con Reintento (*Retrial Model*, RM), representado en la Figura 2.8, introduce reintentos en las peticiones de servicio. A diferencia del tráfico de datos, donde es asumido como válido un modelo sin reintentos, [6], los mecanismos de reintento automático incorporados en los propios terminales móviles generan pruebas sucesivas para el establecimiento de conexión en el tráfico de señalización si existe congestión. Se hace necesaria entonces la incorporación del reintento al modelado de tráfico de señalización.

Se ha considerado que todos los servicios de señalización ejecutan reintentos, exceptuando las peticiones GH (*GHost Seizure*), que expiran sin reintento al ser provocadas por el comportamiento errático del canal radio. Los principales parámetros de este modelo son las tasas de peticiones por servicio, tiempo de servicio, tasa de reintento y número de canales de señalización. Todos estos datos se ajustan celda a celda según los datos reales en una red GERAN. El tamaño de la órbita es lo suficientemente grande como para considerar despreciable la probabilidad de que se encuentre llena. A partir del diagrama de estados del modelo RM, se extraen las expresiones analíticas de sus indicadores de rendimiento: tráfico cursado, ratio de congestión y ratio de bloqueo.

Un segundo modelo, denominado Modelo con Reintentos y Correlación Temporal (*Retrial Model with Correlated Arrivals*, RMCA), y representado en la Figura 2.9, introduce la característica de la correlación temporal entre usuarios. Dicha característica se modela mediante un proceso conmutado de Poisson que alterna el estado del tráfico entre estado *on* y *off*, con intensidades de tráfico distintas. Se intenta modelar así el hecho de que en algunas celdas los mensajes de actualización de localización (*Location Update*, LU) se concentran en cortos periodos de tiempo debido al movimiento en grupo de los usuarios de la red.

En este nuevo modelo RMCA aparecen varios parámetros nuevos respecto al RM, que

caracterizan el comportamiento de la conmutación entre los estados *on* y *off*; concretamente, el tiempo medio de permanencia y distribución estadística del tráfico en cada uno de los dos estados. Según el valor de dichos parámetros, la correlación temporal entre nuevas peticiones RMCA se hace más o menos acusada, variando enormemente el rendimiento de la red. Los valores de estos nuevos parámetros no están disponibles en forma de medidas, ya que ésta últimas sólo reflejan promedios temporales recopilados cadahora, no distinguiendo entre periodos conmutados. Se hace necesario, por tanto, un proceso de estimación y ajuste de estos parámetros en el modelo RMCA celda a celda, basándose en las estadísticas disponibles en el Sistema de Gestión de Red (*Network Management System*, NMS). Para la caracterización de los parámetros de correlación, se define un problema clásico de ajuste de mínimos cuadrados , donde se minimiza la diferencia entre los indicadores de rendimiento del modelo RMCA (de nuevo, tráfico cursado, ratio de congestión y ratio de bloqueo) y esos mismos indicadores medidos en la red real por cada celda. Las variables de decisión en el problema de optimización son los parámetros de correlación a estimar del modelo. Con esta estrategia se obtiene un modelo adaptado a cada celda; es decir, existe un modelo distinto para cada celda donde únicamente varían los valores de los parámetros de correlación temporal.

Durante el proceso de ajuste de parámetros de correlación en el modelo RMCA, se utiliza un método iterativo para resolver el problema de minimización, lo que requiere calcular las probabilidades de los estados en el sistema de colas con órbita repetidas veces. Para agilizar el cálculo, se ha implementado un método eficiente de resolución de sistemas de ecuaciones: el método Gaver, [42]. Este método de resolución de sistemas de ecuaciones es aplicable si las matrices tienen estructura tri-diagonal a bloques. Teniendo cuidado en la nomenclatura de los estados, las probabilidades de cada estado en el modelo RMCA pueden expresarse como un sistema de ecuaciones lineales que tiene dicha estructura tri-diagonal. El Apéndice A detalla el proceso de resolución.

Para comparar los distintos modelos de tráfico de señalización (el utilizado actual por los operadores, basado en la fórmula de Erlang B, y los dos propuestos en esta tesis, RM y RMCA) se definen dos indicadores de rendimiento que evalúan los errores en la estimación de cada modelo. Un primer estimador, ($\overline{SSE}$, (2.20)), contabiliza dichos errores en la estimación según una perspectiva más académica, en la cual todos los errores son igualmente importantes. Un segundo estimador, ($NSAE_{brgs}$, (2.19)), contabiliza dichos errores en la estimación según la perspectiva del operador, donde los errores más importantes son los que se traducen en una pérdida de ingresos).

El escenario para la evaluación de los distintos métodos es una red real GERAN con 1730 celdas, en las que se recogen datos de cada celda en su hora punta durante 8 días, dando lugar a más de 13000 medidas. De este conjunto de medidas se descartan las que se cursen un tráfico extremadamente bajo o las que muestren comportamientos anómalos ajenos al problema tratado en esta tesis. El conjunto de celdas considerado cubre un área geográfica de $120000\,km^2$. Un análisis preliminar de estas medidas recogidas muestra que existe un alto porcentaje de celdas, 19%, que muestran bloqueos inaceptables mayores del 1%, pese a que el 10% de las celdas posean canales sin usar (incluso en la hora punta,

cuando se obtuvieron las medidas). Esto indica que el modelo Erlang B usado para la planificación de los canales de señalización no se ajusta al comportamiento del tráfico en esos canales.

La formula de Erlang B no es capaz de predecir el comportamiento del tráfico en canales SDCCH. Las Figuras 2.14 y 2.15 muestran cómo las medidas de la red no coinciden con las estimaciones del modelo Erlang B. Más concretamente, al definir un intervalo de confianza del 95% alrededor de las medidas de probabilidad de bloqueo, el valor dado por la fórmula de Erlang B queda fuera de dicho intervalo en el 15% de las muestras. Más importante aún, ése 15% de las estimas erróneas corresponden a las muestras con mayor bloqueo y contienen un 26% del tráfico total. En otras palabras, las celdas que muestran más problemas y en las que el operador está más interesado son aquéllas en las que el modelo Erlang B más falla en sus predicciones.

Estos resultados eran esperados, puesto que el modelo de Erlang B no considera reintentos ni correlación entre llamadas. La Tabla 2.2 resume la evaluación de los distintos modelos en el escenario planteado a través de los indicadores $\overline{SSE}$ y $NSAE_{brgs}$ antes comentados. Los resultados están desglosados según el número de canales de señalización en la celda. El error al estimar el número de intentos bloqueados con Erlang B es de un 23%, 42% y 82% en las celdas con 3, 7 y 15 canales, respectivamente. El error se incrementa en celdas con más canales pues es en éstas donde el efecto de correlación es más acusado. El modelo RM reduce poco los errores de estimación, situándose cerca de los resultados de Erlang B. El fenómenos de reintentos, por tanto, justifica tan sólo una pequeña porción de bloqueos. Por contra, RMCA proporciona los menores errores de estimación para cualquier número de canales. En global, RMCA reduce un 63% y 77% los indicadores $\overline{SSE}$ y $NSAE_{brgs}$, respectivamente, respecto a las estimaciones proporcionadas por el modelo Erlang B.

La mejora introducida por las predicciones del modelo RMCA permiten al operador re-planificar recursos en la red a partir de estadísticos de red. La principal tarea consiste en identificar las celdas donde el número de canales SDCCH es innecesariamente alto, puesto que la situación contraria (esto es, excesivamente bajo) puede detectarse faćilmente a partir de las estadísticas de bloqueo de la red. El operador deberá observar si $CR \simeq BR$ y comprobar si coinciden con los valores de probabilidad de bloqueo dados por la fórmula de Erlang B, P_b. En aquellas celdas donde $CR < P_b < BR$ o $P_b < CR < BR$, se debe aplicar y ajustar el modelo RMCA a partir de las medidas de red y averiguar el tráfico ofrecido real, incluyendo las características de correlación y reintento, y su distribución temporal. Una vez realizado el ajuste, el modelo RMCA puede calcular el nuevo número de canales de señalización necesarios de forma precisa.

b) Reparto de tráfico óptimo en GERAN

En este apartado se han definido dos modelos de teletráfico para caracterizar una red GERAN. En dicha red se ajustan parámetros RRM para conseguir un reparto de tráfico óptimo, que consiga mínimizar el tráfico bloqueado. Un primer modelo simple, descrito en la Figura 3.6, no incluye las características de movilidad del usuario y realiza el reparto de

tráfico entre celdas a través del control de admisión. Con este modelo, el balance de carga entre celdas debe hacerse de manera que se iguale en cada celda el indicador obtenido en la ecuación (3.11). Dicho indicador difiere del ratio de bloqueo, sugiriendo que la técnica heurística de igualar el bloqueo entre celdas, adoptada hoy en día por los operadores, no es la óptima. Este modelo simple tiene limitaciones importantes al no considerar los efectos de movilidad del usuario. Para obtener resultados más realistas, se elabora el modelo refinado, descrito en la Figura 3.8, en el que el balance de carga se realiza modificando los márgenes de traspaso. Con una metodología similar al modelo simple, la condición de optimalidad se alcanza igualando en cada celda el indicador definido en (3.17). El Apéndice B describe el proceso matemático seguido para obtener los indicadores óptimos.

La evaluación de los distintos modelos y de las técnicas de balance de carga (tanto las técnicas óptimas como las heurísticas generalmente usadas por los operadores) se realiza en cuatro escenarios de complejidad creciente. Cada escenario añade una nueva característica, de manera que se puede observar su influencia sobre el rendimiento de la red. Una de las características incluidas en los escenarios más complejos consiste en las restricciones impuestas al mecanismo de reparto de tráfico. Dichas restricciones imponen límites en el tráfico que el mecanismo de reparto intenta ofrecer a cada celda. Se intenta reflejar de esta manera la situación real en la que no cualquier usuario puede asignarse a cualquier celda (debido al solapamiento parcial entre celdas). A continuación se ofrece una breve descripción de los escenarios. Los 3 primeros contienen 3 celdas GERAN con capacidad desigual (29, 6 y 6 canales, respectivamente, que corresponden a 4, 1 y 1 transceptor por celda).

- El *escenario 1* considera el modelo simple, esto es, usuarios estáticos, solapamiento total entre celdas y, por tanto, sin restricciones en el reparto de tráfico.
- El *escenario 2* considera el modelo refinado sin restricciones en el reparto de tráfico. El modelo considera ahora la movilidad del usuario, aunque aún se sigue suponiendo solapamiento total entre celdas. Por tanto, este escenario evalúa el impacto de introducir la movilidad del usuario. La distribución espacial de usuarios es uniforme.
- El *escenario 3* considera el modelo refinado con límites controlados en el tráfico ofrecido a cada celda. Se evalúa así el impacto de las restricciones en el tráfico ofrecido. Se siguen modelando las tres celdas GERAN con la distribución de canales antes descrita.
- El *escenario 4* extiende el análisis a un escenario construido a partir de datos reales. El escenario corresponde a las celdas servidas por un mismo controlador de estaciones base (*Base Station Controller*, BSC). El número de celdas es alto y el número de canales por celda es desigual. De igual manera, la distribución de usuarios no es uniforme. A diferencia de los escenarios anteriores, se analizan más de tres celdas y los límites en el reparto de tráfico se calculan a partir de consideraciones geográficas.

En cada uno de los escenarios se prueban 4 estrategias de reparto de tráfico. Todas ellas intentan ecualizar algún indicador de rendimiento. Los tres primeros métodos,

de naturaleza heurística, igualan respectivamente la carga en cada celda (*Load Balancing*, LB), la probabilidad de bloqueo (*Blocking Probability Balancing*, BPB) y el tráfico bloqueado (*Blocked Traffic Balancing*, BTB). El cuarto método es el óptimo (*Optimal Balancing*, OB) que considera el criterio de reparto óptimo en cada modelo de red, simple o refinado, definido en sus respectivas ecuaciones (3.11) y (3.17).

Los resultados se desglosan según los escenarios y técnicas evaluadas. En el escenario 1, el modelo OB consigue el menor tráfico bloqueado global (y, por tanto, el máximo tráfico cursado). Los métodos BPB y BTB consiguen un rendimiento muy parecido, haciendo que la red pierda, respectivamente, un 1% y un 2% de capacidad en términos de tráfico ofrecido para una probabilidad de bloqueo global del 2%. Las diferencias entre métodos se hacen mayores al considerar la movilidad del usuario y la modificación de márgenes de traspaso en el escenario 2. Aun sin restricciones en el tráfico ofrecido (las celdas tienen solapamiento total), el método OB sigue siendo el mejor y el LB el peor (con 35% de pérdida de capacidad). Los métodos BPB y BTB consiguen una capacidad un 3.3% menor que la óptima.

El escenario 3 introduce limitaciones en el espacio de soluciones alcanzable por las distintas técnicas de balance de carga. Para ello, se define un parámetro Δ que define un espacio de soluciones más pequeño o más amplio, siempre centrado en la solución óptima sin restricciones del método OB. Así, Δ=0 significa que todos los métodos de balance de carga tienen una única solución (la óptima), pues es el único punto posible. $\Delta \to \infty$ implica que no hay restricciones, y, por tanto, se alcanzará la misma solución que en el escenario 2. Puesto que el espacio de soluciones está centrado en el punto óptimo, la introducción de restricciones no afecta al método OB, consiguiendo siempre el 100% de la capacidad. Los demás métodos, no obstante, sí que evolucionan con el crecimiento del espacio de soluciones. Conforme el parámetro Δ va definiendo un espacio de soluciones más amplio (esto es, las restricciones espaciales del tráfico ofrecido a cada celda se hacen más relajadas), cada método busca su propia solución consiguiendo el balance del indicador respectivo en cada método. Se observa cómo el método LB es el que posee una solución más alejada del punto óptimo pues el tráfico ofrecido en al menos una celda difiere un 360% respecto a la distribución óptima de tráfico, Figura 3.13.

Por último, el escenario 4 verifica los distintos métodos de balance de carga en un escenario real,. Este escenario se corresponde con el área geográfica servida por una BSC en una red real GERAN. La BSC contiene 117 celdas, tanto omnidireccionales como sectorizadas, y 313 transceptores. Se dispone de la ubicación de las celdas, su número de canales (variando de 6 a 44 canales, esto es, de 1 a 6 transceptores), la orientación de las antenas y el número de intentos de llamada en la hora punta de los últimos 10 días. Las principales diferencias respecto a escenarios anteriores son: a) la distribución no uniforme de los usuarios (y, por tanto, de los intentos de llamada), y b) las restricciones de tráfico son ahora distintas celda a celda. Para el cálculo de los límites en cada celda se toman consideraciones geográficas, tal como se describe en la Figura 3.14. El tráfico máximo que una celda puede cursar es el que está bajo su área de cobertura. El tráfico mínimo es el tráfico dentro de la celda que no esté dentro del área de cobertura de ninguna otra celda.

En la Figura 3.15 se compara la capacidad de la red conseguida con cada método en el escenario 4, variando el radio de cobertura de las celdas. Así, para un radio de cobertura bajo (1 km.) los distintos métodos tienen un rendimiento muy parecido y muy alejado de la capacidad global de la red cuando no se aplican restricciones al tráfico ofrecido a cada celda. Esto es debido a que cuando el radio de cobertura es bajo apenas hay solapamiento entre celdas, y los mecanismos de reparto de tráfico encuentran poco, o ningún, margen para actuar. Para un radio más realista de 5 km. el método OB con restricciones pierde un 7.5% de capacidad respecto a la solución óptima sin restricciones. Más interesante aun, el método BPB tiene una capacidad un 2% menor que el OB en el escenario real.

En este escenario 4, la ganancia o pérdida de capacidad entre los distintos métodos depende del perfil de distribución espacial de canales y usuarios. Para valorar la representatividad de los números obtenidos, se ha hecho un análisis de sensibilidad frente a la distribución espacial de tráfico y recursos siguiendo un método de MonteCarlo. Dicho análisis recoge las variaciones en la ganancia del método OB frente al BPB para distintas distribuciones espaciales de canales y de usuarios por celda, manteniendo fijo el número global de canales y tráfico total en la red. Se han probado 100 escenarios distintos y se han conseguido mejoras que van desde el 2% hasta el 21%, siendo la media del 10%. Esto refleja que el valor del 2% del escenario 4 es un número claramente conservador.

c) Ajuste de parámetros de traspaso en un escenario conjunto GSM/UMTS

En esta última parte del trabajo, se ha definido un esquema de auto-ajuste de parámetros basado en un controlador de lógica difusa o FLC. Los parámetros que se modifican pertenecen al algoritmo de traspaso entre sistemas, IS-HO, dentro de un escenario multi-tecnología, y son a) los umbrales mínimos de nivel de señal y calidad en las tecnologías GSM y UMTS, que garantizan la calidad de la conexión tras el traspaso en ambas tecnologías, y b) los márgenes de traspaso que priorizan unas celdas frente a otras como receptoras de tráfico. Mientras que los primeros se definen a nivel de celda, los segundos se definen a nivel de adyacencia. El objetivo del ajuste es mejorar el rendimiento de la red conjunta mediante el balance de carga entre las distintas tecnologías radio.

Los indicadores de rendimiento de la red multi-tecnología sirven como entrada al FLC. El FLC recoge los indicadores de rendimiento y analiza la situación en la red con la ayuda de las reglas que tiene definidas, Tabla 4.1. Estas reglas traducen y automatizan los procedimientos de optimización de red basados en la experiencia del operador. Tras el análisis, decide las modificaciones adecuadas de los parámetros de IS-HO. Una vez modificados los parámetros, la red móvil modifica su comportamiento y, por tanto, experimentará valores distintos en sus indicadores de rendimiento. A este proceso de varios pasos se lo considera una iteración en la optimización de parámetros.

Para evaluar el rendimiento de esquema basado en FLC, y especialmente su capacidad de adaptación ante situaciones cambiantes, se ha definido un escenario de tráfico con dos fases, sucesivas en el tiempo. Una primera fase establece un tráfico muy descompensado entre GSM y UMTS, aunque uniforme espacialmente dentro de cada tecnología, como se refleja en la Figura 4.9. Posteriormente, en una segunda fase, la distribución

de tráfico cambia radicalmente, para así comprobar la capacidad de adaptación del FLC ante situaciones cambiantes.

COmo plataforma de pruebas para el esquema de modificación de parámetros, se ha modelado la red móvil multi-tecnología mediante un simulador conjunto GSM/UMTS a nivel de red. La Figura 4.7 muestra una estructura global de bloques del simulador GSM/UMTS. Dicho simulador ha sido construido a partir de dos simuladores independientes para cada tecnología radio. Sobre ese punto de partida de tecnologías separadas se ha unificado todo el proceso de generación de tráfico, incluyendo algoritmos de (re)selección de celda y el control de admisión, pasando a tener una perspectiva multi-tecnología. Además de la generación de tráfico conjunta, la funcionalidad multi-tecnología queda reflejada en el módulo de IS-HO, donde está implementado el algoritmo usado en esta parte del trabajo. Además, el simulador implementa las principales funcionalidades intra-sistema como control de potencia, redirección de conexión, traspaso entre celdas o caída de llamadas, por ejemplo. Los principales modelos empleados y la configuración del simulador están incluídos en la Tabla 4.2.

Como resultado, a través de los sucesivos lazos de optimización el FLC modifica adecuadamente los parámetros del IS-HO, como se aprecia en la Figura 4.10. Dichas modificaciones tienden a equilibrar el tráfico descompensado inicialmente. Así, la evolución de los parámetros tiende a favorecer el flujo de traspasos entre sistemas hacia aquella tecnología que tiene más recursos libres. Una vez que la distribución de tráfico cambia radicalmente en la segunda fase, el FLC muestra la tendencia contraria en la modificación de parámetros.

La tasa de bloqueo en cada tecnología tiende a igualarse con el paso de las iteraciones de optimización en ambas fases de simulación, como refleja la Figura 4.11. El balance de carga consigue en la primera fase reducir la tasa de bloqueo de GSM (*Blocking Call Rate*, BCR) en un 12%, mientras que el tráfico cursado en la red se incrementa en un 15%. En la segunda fase también se consigue reducir el BCR, aunque se consigue más tarde debido a que los valores de los parámetros de IS-HO se encuentran al inicio de la segunda fase muy alejados de la zona de mayor sensibilidad de la red. Por esta razón, el balance de carga se consigue después de 35-40 iteraciones. Si bien el FLC consigue el balance de carga, éste se produce de manera muy lenta.

En contrapartida a la reducción del bloqueo, el número de traspasos entre tecnologías se incrementa enormemente, lo que se aprecia en Figura 4.12. Se espera, por tanto, que el incremento del tráfico de señalización en la red sea importante. El balance de carga y la reducción consiguiente del bloqueo en la red no se hace a costa de reducir la calidad en las conexiones cursadas. Los ratios de error de bloque y trama en GSM y UMTS, respectivamente, se mantienen bajo umbrales suficientes.

A la luz de estos resultados, se deduce que la configuración original del FLC consigue el balance de carga entre tecnologías, junto con reducciones importantes del bloqueo, aunque necesita demasiado tiempo para alcanzar el equilibrio. Esto es debido principalmente a que la configuración del FLC presenta: a) unos márgenes de variación en los parámetros

de IS-HO demasiado amplios y alejados de la zona en la que la red conjunta muestra mayor sensibilidad a los cambios, y b) un paso de modificación muy pequeño, que causa que el ritmo de modificación sea ento. Para acelerar el balance de carga, se prueban diversas configuraciones del FLC en las que se restringe el margen y se aumenta el paso de variación de los parámetros del IS-HO, según la Tabla 4.3. Se crean así cuatro configuraciones del FLC que se comparan con la configuración inicial.

Las distintas configuraciones del FLC se comportan según lo esperado al observar las variaciones de los parámetros del IS-HO en la Figura 4.13. Valores más altos en el ritmo de variación de parámetros hacen que el proceso de convergencia se acelere, pero, al mismo tiempo, provoca oscilación en la evolución temporal de los parámetros. Además, las distintas configuraciones simuladas del FLC afectan significativamente a la evolución del ratio de bloqueo en cada tecnología. Algunas configuraciones del FLC consiguen alcanzar el balance de carga hasta 10 iteraciones antes en el proceso de optimización de parámetros, como se refleja en la Figura 4.14.

Para comparar de manera global las distintas configuraciones del FLC, se define un indicador que recoge el ratio de conexiones bloqueadas a lo largo de todo el proceso de optimización, incluyendo todas las iteraciones. Este ratio se reduce en un 4% para algunas configuraciones del FLC respecto a la original. Analizando los resultados de las distintas configuraciones, se observa cómo la reducción en el ratio global de bloqueo se consigue: a) acelerando el ritmo de variación de los parámetros de IS-HO (bajo ciertos límites para evitar la oscilación), y b) reduciendo el margen de valores posibles a la zona de máxima sensibilidad de la red conjunta.

C.5 Conclusiones

En esta tesis se han tratado distintos asuntos, como modelos de teletráfico para canales de señalización dedicados en GERAN y técnicas de reparto de tráfico tanto en GERAN como redes multi-tecnología. En esta sección se presentan las principales conclusiones de manera separada para cada asunto.

a) Modelo de teletráfico para canales de señalización dedicados en GERAN

- Un análisis completo sobre datos de señalización dedicados de una red real ha permitido detectar los principales fallos en los modelos usados actualmente por el operador para dimensionar los canales de señalización que cursan dicho tráfico.
- La adición de las características de reintento y correlación, especialmente la segunda, a un modelo de teletráfico para los canales de señalización dedicados en GERAN ha proporcionado una mayor exactitud al modelo cuando éste es comparado con los datos de red reales disponibles. El efecto de correlación proviene de los procedimientos de gestión de localización a través de canales de señalización dedicados cuando

existen movimientos de grupos de usuario.

- El nuevo modelo propuesto viene acompañado por un procedimiento para el ajuste de sus parámetros. Este procedimiento se basa en formular el ajuste de parámetros como un problema de mínimos cuadrados, intentando ajustar al máximo los indicadores de rendimiento del modelo a los datos reales de red.

b) Reparto de tráfico óptimo en GERAN

- Se ha obtenido un indicador óptimo de reparto de tráfico a partir de un modelo de teletráfico que incluye características de movilidad de usuario. El reparto de tráfico se realiza a través de la modificación de los márgenes de traspaso, causando variaciones en el área de servicio de las celdas.
- Los indicadores heurísticos usados por el operador para el reparto de tráfico no son óptimos y pueden tener pérdidas significativas de ganancia de tráfico en la red respecto al indicador propuesto en este trabajo.
- La ganancia obtenida por el método óptimo es dependiente de las condiciones geográficas del problema. Para el análisis del rendimiento se han construido escenarios a partir de datos de red real.

c) Auto-ajuste de parámetros de traspaso en un escenario GSM/UMTS

- El ajuste de parámetros de nivel y calidad en el algoritmo IS-HO basado en un esquema con FLC consigue de manera efectiva el balance de carga en escenario multi-tecnología con altos desequilibrios de tráfico.
- La configuración del FLC para la modificación de parámetros es muy influyente sobre el proceso de balance de carga, especialmente en cuanto al tiempo en alcanzar el equilibrio. Las modificaciones en la configuración del controlador para acelerar el proceso de balance de carga se centran en el paso y rango de modificación de parámetros.
- La plataforma para la evaluación de las propuestas ha sido un simulador de red dinámico conjunto GSM/UMTS que ha incluido las principales funciones inter- e intra-tecnología. Esta plataforma ha sido una herramienta básica en algunos de los proyectos de investigación en los que esta tesis ha estado involucrado.

La mayor parte del trabajo se ha desarrollado en redes GERAN. Este escenario de red, al estar en una fase madura, es muy adecuado para la prueba de diferentes estrategias de optimización, especialmente cuando las propuestas de mejora no implican cambios en la infraestructura de la red, siendo éste el caso de las distintas contribuciones aportadas en este trabajo. Además, la mayoría de algoritmos y estrategias presentadas pueden trasladarse a otras tecnologías más novedosas, tal como se describe en en el capítulo 5.

C.6 Lista de Publicaciones

A continuación se detalla una lista de las distintas contribuciones que ha originado esta tesis, así como los proyectos de investigación en los que su trabajo ha estado presente.

Artículos

[I] S. Luna Ramírez, M. Toril, M. Fernández Navarro y V. Wille, "Optimal Traffic Sharing in GERAN," *Wireless Personal Communications*, Springer. Publicado electrónicamente, DOI 10.1007/s11277-009-9861-6.

[II] S. Luna Ramírez, M. Toril y V. Wille, "Performance Analysis of Dedicated Signalling Channels in GERAN by retrial Queues," *Wireless Personal Communications*, Springer. Publicado electrónicamente, DOI 10.1007/s11277-010-9939-1.

Congresos y Jornadas de Trabajo

[III] R. Barco, S. Luna Ramírez y M. Fernández Navarro, "Optimisation and Troubleshooting of Heterogeneous Mobile Communication Networks," 3^{rd} *Workshop Trends in Radio Resource Management*, Barcelona (España), Noviembre, 2007.

[IV] S. Luna Ramírez, M. Toril, F. Ruiz y M. Fernández Navarro, "Adjustment of a Fuzzy Logic Controller for IS-HO parameters in a heterogenous scenario," en *Proc. IEEE 14th Mediterranean Electrotechnnical Conference* (MELECON'2008), Mayo, 2008, pp. 29–34.

[V] S. Luna Ramírez, M. Toril, F. Ruiz y M. Fernández Navarro, "Inter-system Handover Parameter Auto-Tuning in a Joint-RRM Scenario," en *Proc. IEEE 67th Vehicular Technology Conference* (Spring VTC'08), Mayo, 2008, pp. 2641–2645.

[VI] S. Luna Ramírez, M. Toril, y V. Wille, "Balance de Carga Óptimo en GERAN," en *Actas XXIV Simposium Nacional de la Unión Científica Internacional de Radio* (URSI 2009), Santander (España), Septiembre, 2009.

[VII] S. Luna Ramírez, M. Toril y V. Wille, "Performance Analysis of Dedicated Signalling Channels in GERAN," TD(09)805 en *Proc. of the 8th Management Committe Meeting*, acción COST 2100, Valencia (España), Mayo, 2009.

[VIII] S. Luna Ramírez, M. Toril, V. Wille y M. Fernández Navarro, "Optimal Traffic Sharing in GERAN," TD(10)10022 en *Proc. of the 10th Management Committe Meeting*, acción COST 2100, Atenas (Grecia), Febrero, 2010.

[IX] I. de la Bandera, S. Luna-Ramírez, R. Barco, M. Toril, F. Ruiz y M. Fernández-Navarro, "Inter-system Cell Reselection Parameter Auto-Tuning in a Joint-RRM Scenario," en *Fifth International Conference on Broadband and Biomedical Communications* (IB2COM 2010), Málaga (España), Diciembre, 2010 (*aceptado*).

[II, VII] tratan sobre el modelado de tráfico en canales de señalización dedicados en tecnología GERAN. [I, VI, VIII] se centran en la definición de modelos de tráfico analíticos y criterio óptimo para el problema del reparto de tráfico, también en tecnología GERAN. Por último, [III-V] describen el ajuste automático de parámetros en un entorno de red heterogénea, mientras que [IX] usa un esquema similar para el ajuste de parámetros del algoritmo de reselección de celda en un escenario similar. El autor de esta tesis ha sido el primer autor de todas las contribuciones relacionadas exceptuando [III,IX], en donde ha tenido una contribución importante. En [III], el autor ha sido el responsable de la sección dedicada a la optimización de parámetros en redes heterogéneas, mientras que en [IX] la contribución ha estado centrada en las secciones de descripción del escenario y análisis del rendimiento.

Todas estas contribuciones han surgido en el marco de diversos proyectos de investigación. [III-V] se desarrollaron a lo largo de dos proyectos distintos: el proyecto europeo "*GANDALF: Monitoring and self-tuning of RRM parameters in a multisystem network*" (en la iniciativa europea CELTIC-EUREKA), premiado con el *Celtic Excellence Award*, y el proyecto TIC-4052, "*Técnicas adaptativas de gestión de recursos radio en redes B3G*", becado por la Junta de Andalucía. [II, VII] están enmarcadas en el proyecto nacional TEC2008-06216 ("*Optimización de la estructura de redes de acceso radio heterogéneas mediante partición de grafos*") del Ministerio Español de Ciencia y Tecnología, así como [VIII, IX] se desarrollaron en un proyecto de similar categoría (TEC2009-13413, "*Optimización automática de redes de comunicaciones móviles heterogéneas*").

Bibliography

[1] V. B. Iversen, *Teletraffic engineering handbook*. COM depart., Techn. University of Denmark, 2002.

[2] M. Toril, "Self-tuning algorithms for the assignment of packets control units and handover parameters in GERAN," Ph.D. dissertation, University of Málaga, 2008.

[3] [Online]. Available: www.gsmworld.com

[4] M. Mouly and M. B. Pautet, *The GSM system for mobile communications*. Cell & Sys, 1992.

[5] T. Halonen, J. Melero, and J. Romero, *GSM, GPRS and EDGE performance: evolution Toward 3G/UMTS*. John Wiley & Sons, 2002.

[6] G. Tunnicliffe, A. Murch, A. Sathyendran, and P. Smith, "Analysis of traffic distribution in cellular networks," in *Proc. 48th IEEE Vehicular Technology Conference*, vol. 3, May 1998, pp. 1984–1988.

[7] E. Onur, H. Delic, C. Ersoy, and M. U. Caglayan, "Measurement-based replanning of cell capacities in GSM networks," *Computer Networks*, vol. 39, no. 6, pp. 749–767, 2002.

[8] P. Tran-Gia and M. Mandjes, "Modeling of customer retrial phenomenon in cellular mobile networks," *IEEE Journal on Selected Areas in Communications*, vol. 15, no. 8, pp. 1406–1414, Oct 1997.

[9] A. S. Alfa and W. Li, "PCS networks with correlated arrival process and retrial phenomenon," *IEEE Transactions on Wireless Communications*, vol. 1, no. 4, pp. 630–637, Oct 2002.

[10] U. Gotzner, A. Gamst, and R. Rathgeber, "Spatial traffic distribution in cellular networks," in *Proc. 48th IEEE Vehicular Technology Conference*, vol. 2, May 1998, pp. 1994–1998.

[11] S. Almeida, J. Queijo, and L. Correia, "Spatial and temporal traffic distribution models for GSM," in *Proc. 50th IEEE Vehicular Technology Conference*, vol. 1, May 1999, pp. 131–135.

[12] M. Toril and V. Wille, "Optimization of handover parameters for traffic sharing in GERAN," *Wireless Personal Communications*, vol. 47, no. 3, pp. 315–336, Nov 2008.

[13] S. Pedraza, V. Wille, M. Toril, R. Ferrer, and J. Escobar, "Dimensioning of signaling capacity on a cell basis in GSM/GPRS," in *Proc. 54th IEEE Vehicular Technology Conference*, vol. 1, April 2003, pp. 155–159.

[14] "3GPP TS 04.08 (v7.20.1), Mobile radio interface layer 3 specification; GSM-Phase2+, Release 98," Sep 2003.

[15] "3GPP TS 05.02 (v6.6.0), Multiplexing and multiple access on the radio path; GSM-Phase2+, Release 97," Nov 1999.

[16] 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*, vol. 35, no. 3, pp. 77–92, Aug 1986.

[17] Y. Fang and I. Chlamtac, "Teletraffic analysis and mobility modeling of PCS networks," *IEEE Transactions on Communications*, vol. 47, no. 7, pp. 1062–1071, July 1999.

[18] S. Rappaport, "Traffic performance of cellular communication systems with heterogeneous call and platform types," in *Conf. Record, 2nd Int. Conf. on Universal Personal Communications*, vol. 2, Oct 1993, pp. 690–695.

[19] B. Epstein and M. Schwartz, "Reservation strategies for multi-media traffic in a wireless environment," *Proc. IEEE 45th Vehicular Technology Conference*, vol. 1, pp. 165–169, vol.1, Jul 1995.

[20] Y. Fang, "Thinning schemes for call admission control in wireless networks," *IEEE Transactions on Computers*, vol. 52, no. 5, pp. 685–687, May 2003.

[21] S. Rappaport and L.-R. Hu, "Microcellular communication systems with hierarchical macrocell overlays: traffic performance models and analysis," *Proceedings of the IEEE*, vol. 82, no. 9, pp. 1383–1397, Sep 1994.

[22] X. Lagrange and P. Godlewski, "Teletraffic analysis of a hierarchical cellular network," in *Proc. IEEE 45th Vehicular Technology Conference*, vol. 2, Jul 1995, pp. 882–886 vol.2.

[23] B. Jabbari and W. Fuhrmann, "Teletraffic modeling and analysis of flexible hierarchical cellular networks with speed-sensitive handoff strategy," *IEEE Journal on Selected Areas in Communications*, vol. 15, no. 8, pp. 1539–1548, Oct 1997.

[24] P. Fitzpatrick, C. S. Lee, and B. Warfield, "Teletraffic performance of mobile radio networks with hierarchical cells and overflow," *IEEE Journal on Selected Areas in Communications*, vol. 15, no. 8, pp. 1549–1557, August 1997.

[25] J. Luo, R. Mukerjee, M. Dillinger, E. Mohyeldin, and E. Schulz, "Investigation of radio resource scheduling in WLANs coupled with 3G cellular network," *IEEE Communications Magazine*, vol. 41, no. 6, pp. 108–115, June 2003.

[26] J. Luo, M. Dillinger, E. Mohyeldin, and E. Schulz, "Gain analysis of joint radio resource management for reconfigurable terminals," *Multiradio Multimedia Communications*, February 2003.

[27] S. Luna-Ramírez, M. Toril, M. F. Navarro, R. Skehill, and S. McGrath, "Evaluation of policy-based admission control algorithms for a joint radio resource management environment," in *Proc. IEEE Mediterranean Electrotechnical Conference (MELECON)*, 2006, pp. 599–603.

[28] G. Falin and J. Templeton, *Retrial queues*. Chapman and Hall, 1997.

[29] J. R. Artalejo and A. Gómez-Corral, *Retrial Queueing Systems*. Springer, 2008.

[30] N. W. Macfadyen, "Statistical observation of repeated attempts in the arrival process," in *Proc. 9th International Teletraffic Congress (ITC)*, 1979.

[31] K. Liu, "Direct distance dialing: Call completion and customer retrial behavior," *Bell System Technical Journal*, vol. 59, pp. 295–311, 1980.

[32] M. Nesenbergs, "A hybrid of Erlang B and C formulas and its applications," *IEEE Transactions on Communications*, vol. 27, pp. 59–68, Oct 1979.

[33] M. Neuts and B. M. Rao, "Numerical investigation of a multiserver retrial model," *Queueing Systems*, vol. 7, no. 2, pp. 169–189, 1990.

[34] J. Artalejo and M. Pozo, "Numerical calculation of the stationary distribution of the main multiserver retrial queue," *Annals of Operations Research*, vol. 116, pp. 41–56, 2002.

[35] M. Domenech-Benlloch, J. Giménez-Guzmán, V. Pla, J. Martínez-Bauset, and V. Casares-Giner, "Generalized truncated methods for an efficient solution of retrial systems," *Mathematical Problems in Engineering*, vol. 2008, 2008.

[36] S. R. Chakravarthy, A. Krishnamoorthy, and V. C. Joshua, "Analysis of a multiserver retrial queue with search of customers from the orbit," *Performance Evaluation*, vol. 63, no. 8, pp. 776–798, 2006.

[37] M. A. Marsan, G. D. Carolis, E. Leonardi, R. L. Cigno, and M. Meo, "Efficient estimation of call blocking probabilities in cellular mobile telephony networks with customer retrials," *IEEE Journal on Selected Areas in Communications*, vol. 19, pp. 332–346, 2001.

[38] M. Domenech-Benlloch, J. Giménez-Guzmán, J. Martínez-Bauset, and V. Casares-Giner, "Efficient and accurate methodology for solving multiserver retrial systems," *IEE Electronics Letters*, vol. 41, no. 17, pp. 967–969, 2005.

[39] J. Giménez-Guzmán, M. Domenech-Benlloch, V. Pla, V. Casares-Giner, and J. Martínez-Bauset, "Analysis of cellular network with user redials and automatic handover retrials," *Lecture Notes in Computer Science, Next Generation Teletraffic and Wired/Wireless Advanced Networking*, vol. 4712/2007, pp. 210–222, 2007.

[40] R. Wilkinson, "Theories for toll traffic engineering in the U.S.A." *Bell System Technical Journal*, vol. 35, no. 2, pp. 421–514, 1956.

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

[42] D. Gaver, P. Jacobs, and G. Latouche, "Finite birth-and-death models in randomly changing environments," *Advances in Applied Probability*, vol. 16, no. 4, pp. 715–731, 1984.

[43] S. Rappaport, "The multiple-call hand-off problem in high-capacity cellular communications systems," in *Proc. IEEE 40th Vehicular Technology Conference (VTC)*, May 1990, pp. 287–294.

[44] K. Meier-Hellstern and W. Fischer, "The Markov-Modulated Poisson Process (MMPP) cookbook," *Performance Evaluation*, vol. 18, pp. 149–171, 1992.

[45] J. Nocedal and S. J. Wright, *Numerical Optimization*, 2nd ed. Springer, 2006.

[46] *Optimization Toolbox 4, User's Guide*, The MathWorks, 2008.

[47] D. Gross and C. M. Harris, *Fundamentals of queueing theory*, 3rd ed. Wiley, 1998.

[48] A. Andreadis, G. Benelli, G. Giambene, and B. Marzucchi, "Analysis of the WAP protocol over SMS in GSM networks," *Wireless Communications and Mobile Computing*, vol. 1, pp. 381–395, 2001.

[49] J. Karlsson and B. Eklund, "A cellular mobile telephone system with load sharing - An enhancement of directed retry," *IEEE Transactions on Communications*, vol. 37, no. 5, pp. 530–535, May 1989.

[50] M. Toril, R. Ferrer, S. Pedraza, V. Wille, and J. J. Escobar, "Optimization of half-rate codec assignment in GERAN," *Wireless Personal Communications*, vol. 34, no. 3, pp. 321 – 331, Aug 2005.

[51] J. Kojima and K. Mizoe, "Radio mobile communication system wherein probability of loss of calls is reduced without a surplus of base station equipment," U.S. Patent 4435840, Mar 1984.

[52] V. Wille, M. Toril, and R. Barco, "Impact of antenna downtilting on network performance in GERAN systems," *IEEE Communications Letters*, vol. 9, no. 7, pp. 598–600, Jul 2005.

[53] M. Toril, S. Pedraza, R. Ferrer, and V. Wille, "Optimization of signal level thresholds in mobile networks," in *Proc. 55th IEEE Vehicular Technology Conference (VTC)*, vol. 4, February 2002.

[54] N. Papaoulakis, D. Nikitopoulos, and S. Kyriazakos, "Practical radio resource management techniques for increased mobile network performance," in *12th IST Mobile and Wireless Communications Summit*, Jun 2003.

[55] V. Wille, S. Pedraza, M. Toril, R. Ferrer, and J. Escobar, "Trial results from adaptive hand-over boundary modification in GERAN," *Electronics Letters*, vol. 39, no. 4, pp. 405–407, Feb 2003.

[56] "Radio access network; Radio subsystem link control," 3rd Generation Partnership Project (3GPP), TR 45.008, Nov 2001.

[57] C. Chandra, T. Jeanes, and W. Leung, "Determination of optimal handover boundaries in a cellular network based on traffic distribution analysis of mobile measurement reports," in *Proc. IEEE 47th Vehicular Technology Conference (VTC)*, vol. 1, May 1997, pp. 305–309.

[58] J. Steuer and K. Jobmann, "The use of mobile positioning supported traffic density measurements to assist load balancing methods based on adaptive cell sizing," in *Proc. 13th IEEE Int. Symp. on Personal Indoor and Mobile Radio Communications (PIMRC)*, vol. 3, Jul 2002, pp. 339–343.

[59] J. Wigard, T. Nielsen, P. Michaelsen, and P. Morgensen, "On a handover algorithm in a PCS1900/GSM/DCS1800 network," in *Proc. IEEE 49th Vehicular Technology Conference (VTC)*, vol. 3, Jul 1999, pp. 2510–2514.

[60] A. Baier and K. Bandelow, "Traffic engineering and realistic network capacity in cellular radio networks with inhomogeneous traffic distribution," in *Proc. IEEE 47thVehicular Technology Conference*, vol. 2, May 1997, pp. 780 –784 vol.2.

[61] R. Guerin, "Queueing-blocking system with two arrival streams and guard channels," *IEEE Transactions on Communications*, vol. 36, no. 2, pp. 153–163, Feb 1988.

[62] S. S. Rappaport, "Blocking, hand-off and traffic performance for cellular communications with mixed platforms," *IEEE Journal on Selected Areas in Communications*, vol. 140, pp. 389–401, Oct 1993.

[63] S. Louvros, J. Pylarinos, and S. Kotsopoulos, "Mean waiting time analysis in finite storage queues for wireless cellular networks," *Wireless Personal Communications*, vol. 40, no. 2, pp. 145–155, Jan 2007.

[64] P. V. Orlik and S. S. Rappaport, "A model for teletraffic performance and channel holding time characterization in wireless cellular communication with general session and dwell time distributions," *IEEE Journal on Selected Areas in Communications*, vol. 16, no. 5, pp. 788–803, Jun 1998.

[65] W. Li and A. S. Alfa, "A PCS network with correlated arrival process and splitted-rating channels," *IEEE Journal on Selected Areas in Communications*, vol. 17, no. 7, pp. 1318–1325, Jul 1999.

[66] J. Wang, Q. Zeng, and D. Agrawal, "Performance analysis of a preemptive and priority reservation handoff algorithm for integrated service-based wireless mobile networks," *IEEE Transactions on Mobile Computing*, vol. 2, no. 1, pp. 65–75, Jan–Mar 2003.

[67] D. Staehle and A. Mäder, "An analytic approximation of the uplink capacity in a UMTS network with heterogeneous traffic," in *Proc. 18th International Teletraffic Congress (ITC)*, Sep 2003, pp. 81–91.

[68] V. Iversen, V. Benetis, N. Ha, V. Ha, and S. Stepanov, "Evaluation of multi-service CDMA networks with soft blocking," in *Proc. 16th International Teletraffic Congress (ITC)*, Sep 2004, pp. 212–216.

[69] H. Wang and V. Iversen, "Erlang capacity of multi-class TDMA systems with adaptive modulation and coding," in *Proc. IEEE International Conference on Communications (ICC)*, May 2008, pp. 115–119.

[70] A. Ganz, C. Krishna, D. Tang, and Z. Haas, "On optimal design of multitier wireless cellular systems," *IEEE Communications Magazine*, vol. 35, no. 2, pp. 88–93, Feb 1997.

[71] H. Chen, Q.-A. Zeng, and D. P. Agrawal, "A novel analytical model for optimal channel partitioning in the next generation integrated wireless and mobile networks," in *Proc. of the 5th ACM international workshop on Modeling analysis and simulation of wireless and mobile systems (MSWiM)*, Sep 2002, pp. 120–127.

[72] A. Eisenblatter and H.-F. Geerdes, "Capacity optimization for UMTS: bounds and benchmarks for interference reduction," in *Proc. IEEE 19th Int. Symp. on Personal, Indoor and Mobile Radio Communications (PIMRC)*, Sep 2008, pp. 607–610.

[73] V. Frost and B. Melamed, "Traffic modeling for telecommunications networks," *IEEE Communications Magazine*, vol. 32, no. 3, pp. 70–81, Mar 1994.

[74] D. Everitt, "Traffic engineering of the radio interface for cellular mobile networks," *Proceedings of the IEEE*, vol. 82, no. 9, pp. 1371–1382, Sep 1994.

[75] J. Evans and D. Everitt, "On the teletraffic capacity of CDMA cellular networks," *IEEE Transactions on Vehicular Technology*, vol. 48, no. 1, pp. 153–165, Jan 1999.

[76] D. L. Jagerman, "Some properties of the Erlang loss function," *Bell System Technical Journal*, vol. 53, no. 3, pp. 525–551, May 1974.

[77] G. Foschini, B. Gopinath, and Z. Miljanic, "Channel cost of mobility," *IEEE Transactions on Vehicular Technology*, vol. 42, no. 4, pp. 414–424, Nov 1993.

[78] T. Yum and K. Yeung, "Blocking and handoff performance analysis of directed retry in cellular mobile systems," *IEEE Transactions on Vehicular Technology*, vol. 44, no. 3, pp. 645–650, Aug 1995.

[79] H. Heredia-Ureta, F. Cruz-Perez, and L. Ortigoza-Guerrero, "Capacity optimization in multiservice mobile wireless networks with multiple fractional channel reservation," *IEEE Transactions on Vehicular Technology*, vol. 52, no. 6, pp. 1519–1539, Nov 2003.

[80] V. Pla, J. Martínez, and V. Casares-giner, "Efficient computation of optimal capacity in multiservice mobile wireless networks," in *Traffic and Performance Engineering for Heterogeneous Networks*, D. Kouvatsos, Ed. Aalborg, Denmark: River Publishers, 2009.

[81] F. Khan and D. Zeghlache, "Effect of cell residence time distribution on the performance of cellular mobile networks," in *Proc. IEEE 47th Vehicular Technology Conference (VTC)*, vol. 2, May 1997, pp. 949–953.

[82] P. V. Orlik and S. S. Rappaport, "On the handoff arrival process in cellular communications," *Wireless Networks*, vol. 7, no. 2, pp. 147–157, Mar/Apr 2001.

[83] S. Kourtis and R. Tafazolli, "Downlink shared channel: an effective way for delivering Internet services in UMTS," in *Proc. 3rd Int. Conf. 3G Mobile Communication Technologies*, May 2002, pp. 479–483.

[84] F. Aurenhammer, "Voronoi diagrams - A survey of a fundamental geometric data structure," *ACM Computing Surveys*, vol. 23, no. 3, pp. 345–405, 1991.

[85] "3GPP TS 32.500 (v8.0.0), Technical Specification group service and system aspects; Telecommunication management; Self-Organizing Networks (SON); Concepts and requirement, Release 8," Jun 2008.

[86] R. Skehill, M. Barry, W. Kent, M. O'Callaghan, N. Gawley, and S. McGrath, "The common RRM approach to admission control for converged heterogeneous wireless networks," *IEEE Wireless Communications*, vol. 14, pp. 48–56, 2007.

[87] V. Gazis, N. Alonistioti, and L. Merakos, "Toward a generic 'Always Best Connected' capability in integrated WLAN/UMTS cellular mobile networks (and beyond)," *IEEE Wireless Communications*, vol. 12, pp. 20–29, 2005.

[88] A. Tölli, P. Hakalin, and H. Holma, "Performance evaluation of common radio resource management (CRRM)," in *Proc. of IEEE International Conference on Communications*, vol. 5, 2002, pp. 3429–3433.

[89] C. Brunner, A. Garavaglia, M. Mittal, M. Narang, and J. Bautista, "Inter-system handover parameter optimization," in *Proc. of 64th IEEE Vehicular Technology Conference*, Sep 2006.

[90] T. Ross, *Fuzzy logic with engineering applications*. McGraw-Hill, 1995.

[91] J. Laiho, *Radio Network Planning and Optimisation for Umts*, A. Wacker and T. Novosad, Eds. New York, NY, USA: John Wiley & Sons, Inc., 2002.

[92] "3GPP TS 22.934 (v6.2.0), Technical Specification group service and system aspects; Feasibility study on 3GPP system to Wireless Local Area Network (WLAN) interworking Release 8," 2003.

[93] M. O'Callaghan, N. Gawley, M. Barry, and S. McGrath, "Admision control for heterogeneous networks," in *Proc. of 13th IST Mobile and Wireless Communications Summit*.

[94] M. Benson and H. Thomas, "Investigation of the UMTS to GSM handover procedure," in *Proc. 55th IEEE Vehicular Technology Conference*, vol. 4, 2002, pp. 1829–1833.

[95] N. Motte, R. Rmmler, D. Grandblaise, L. Elicegui, D. Bourse, and E. Seidel, "Joint radio resource management and QoS implications of software downloading for SDR terminals," in *Proc. of IST mobile & Wireless Telecommunications Summit*, 2002.

[96] N. Saravanan, N. Sreenivasulu, D. Jayaram, and A. Chockalingam, "Design and performance evaluation of an inter-system handover algorithm in UMTS/GSM networks," in *Proc. of IEEE TENCON*, 2005.

[97] L. Wang, H. Aghvami, N. Nafisi, O. Sallent, and J. Perez-Romero, "Voice capacity with coverage-based CRRM in a heterogeneous UMTS/GSM environment," in *Communications and Networking in China, 2007. CHINACOM '07. Second International Conference on*, aug. 2007, pp. 1085 –1089.

[98] B. Homnan, V. Kunsriruksakul, and W. Benjapolakul, "Adaptation of CDMA soft handoff thresholds using fuzzy inference system," in *Proc. IEEE International Conference on Personal Wireless Communications*, 2000, pp. 259 –263.

[99] B. Homnan and W. Benjapolakul, "QoS-controlling soft handoff based on simple step control and a fuzzy inference system with the gradient descent method," *IEEE Transactions on Vehicular Technology*, vol. 53, no. 3, pp. 820 – 834, may 2004.

[100] V. Kunsriraksakul, B. Homnan, and W. Benjapolakul, "Comparative evaluation of fixed and adaptive soft handoff parameters using fuzzy inference systems in CDMA mobile communication systems," in *Proc. IEEE 53rd Vehicular Technology Conference,*, vol. 2, 2001, pp. 1017 –1021 vol.2.

[101] Y.-L. Chen, Y.-S. Lin, J.-H. Wen, W.-M. Chang, and J. Liao, "Combined fuzzy-based rate and selective power control in multimedia CDMA cellular systems," in *Proc. IEEE 58th Vehicular Technology Conference,*, vol. 4, oct. 2003, pp. 2506 – 2510 Vol.4.

[102] C.-H. Jiang and J.-K. Lain, "Adaptive neuro-fuzzy power control and rate adaptation for multirate CDMA radio systems," in *Proc. IEEE International Conference on Networking, Sensing and Control,*, vol. 2, 2004, pp. 1307 – 1312 Vol.2.

[103] W. Panichpattanakul and W. Benjapolakul, "Fuzzy power control with weighting function in DS-CDMA cellular mobile communication system," in *Proc. 2003 International Symposium on Circuits and Systems*, vol. 5, may 2003, pp. V–785 – V–788 vol.5.

[104] J. Ye, X. Shen, and J.W.Mark, "Call admission control in wideband CDMA cellular networks by using fuzzy logic," *IEEE Transactions on mobile computing*, vol. 4, pp. 129–141, 2005.

[105] K. Murray and D. Pesch, "Intelligent network access and inter-system handover control in heterogeneous wireless networks for smart space environments," in *Proc. 1st International Symposium on Wireless Communication Systems,*, sept. 2004, pp. 66 – 70.

[106] W. Zhang, "Handover decision using fuzzy MADM in heterogeneous networks," in *Proc. IEEE Wireless Communications and Networking Conference,*, vol. 2, march 2004, pp. 653 – 658 Vol.2.

[107] Z. Altman and P. Stuckmann, "Planning, management and auto-tuning techniques for UMTS and heterogeneous radio acess networks," in *Proc. 9th IFIP/IEEE International Symposium onIntegrated Network Management,*, may 2005, pp. 790 – 790.

[108] S. Horrich, S. Jamaa, and P. Godlewski, "Adaptive vertical mobility decision in heterogeneous networks," in *Proc. 3rd International Conference on Wireless and Mobile Communications (ICWMC)*, 2007.

[109] H. Berenji and P. Khedkar, "Learning and tuning fuzzy logic controllers through reinforcements," *IEEE Transactions on Neural Networks*, vol. 3, no. 5, pp. 724 –740, sep 1992.

[110] R. Agustí, O. Salient, J. Pérez-Romero, and L. Giupponi, "A fuzzy-neural based approach for joint radio resource management in a beyond 3G," in *Proc. 1st International Conference on Quality of Service in Heterogeneous Wired/Wireless Networks (QSHINE)*, 2004.

[111] L. Giupponi, R. Agusti, J. Perez-Romero, and O. Sallent Roig, "A novel approach for joint radio resource management based on fuzzy neural methodology," *IEEE Transactions on Vehicular Technology*, vol. 57, no. 3, pp. 1789 –1805, may 2008.

[112] Y.-H. Chen, C.-J. Chang, and C. Y. Huang, "Fuzzy Q-learning admission control for WCDMA/WLAN heterogeneous networks with multimedia traffic," *IEEE Transactions on Mobile Computing*, vol. 8, no. 11, pp. 1469 –1479, nov. 2009.

[113] L. Saker, S. Ben Jemaa, and S. Elayoubi, "Q-learning for joint access decision in heterogeneous networks," in *Proc. IEEE Wireless Communications and Networking Conference*, april 2009, pp. 1 –5.

[114] R. Nasri, A. Samhat, and Z. Altman, "A new approach of UMTS-WLAN load balancing; algorithm and its dynamic optimization," in *Proc. IEEE International Symposium on a World of Wireless, Mobile and Multimedia Networks*, june 2007, pp. 1 –6.

[115] "3GPP TS 25.331 (v6.7.0), radio resource control (RRC); protocol specification, Release 6," Jun 2009.

[116] H. Holma and A. Toskala, *WCDMA for UMTS: Radio Access for Third Generation Mobile Communications*, 3rd ed. John Wiley & Sons, 2004.

[117] ——, *LTE for UMTS. OFDMA and SC-FDMA based radio access.* John Wiley & Sons, 2009.

[118] "3GPP TS 23.012 (v8.2.0), Technical specification group core network; Location management procedures, Release 8," Jun 2009.

[119] V. Iversen, S. N. Stepanov, and V. Kostrov, "The derivation of stable recursion for multi-service models," in *Proc. Next Generation Teletraffic and Wired-Wireless Advanced Networking (NEW2AN)*, Feb 2004, pp. 254–259.

[120] V. B. Iversen and S. N. Stepanov, "Derivatives of blocking probabilities for multi-service loss systems and their applications," in *Proc. Next Generation Teletraffic and Wired-Wireless Advanced Networking (NEW2AN)*, Sep 2007, pp. 260–268.

[121] L. Delbrouck, "On the steady-state distribution in a service facility carrying mixtures of traffic with different peakedness factors and capacity requirements," *IEEE Transactions on Communications*, vol. 31, no. 11, pp. 1209–1211, Nov 1983.

[122] A. Garavaglia, C. Brunner, D. Flore, M. Yang, and F. Pica, "Inter-system cell reselection parameter optimization in UMTS," in *Proc. IEEE 16th International Symposium on Personal, Indoor and Mobile Radio Communications*, vol. 3, Sep 2005, pp. 1636 –1640 Vol. 3.

[123] I. de la Bandera, S. Luna-Ramírez, R. Barco, M. Toril, F. Ruiz, and M. Fernández-Navarro, "Inter-system cell reselection parameter auto-tuning in a joint-RRM scenario," in *Proc. of Fifth International Conference on Broadband and Biomedical Communications (IB2COM)*, 2010 *(submitted)*.

[124] T. Jansen, M. Amirijoo, U. Turke, L. Jorguseski, K. Zetterberg, R. Nascimento, L. Schmelz, J. Turk, and I. Balan, "Embedding multiple self-organisation functionalities in future radio access networks," in *Proc. IEEE 69th Vehicular Technology Conference,*, april 2009, pp. 1 –5.

[125] W. Grassmann, M. Taksar, and D. Heyman, "Regenerative analysis and steady state distributions for markov chains," *Operations Research*, vol. 33, pp. 1107–1116, 1985.

[126] L. D. Servi, "Algorithmic solutions to two-dimensional birth-death processes with application to capacity planning," *Telecommunications Systems*, vol. 21, no. 2-4, pp. 205–212, 2002.

[127] R. L. Burden and D. J. Faires, *Numerical Analysis*. Brooks Cole, December 2004.

[128] K. Krishnan, "The convexity of loss rate in an Erlang loss system and sojourn in an Erlang delay system with respect to arrival and service rates," *IEEE Transactions on Communications*, vol. 38, no. 9, pp. 1314–1316, Sep 1990.

[129] S. Rao, *Engineering Optimization: Theory and Practice. 3^{rd} edition.* New York, USA: John Wiley & Sons, 1996.

[130] S. M. Stefanov, "Convex separable minimization subject to bounded variables," *Computational Optimization and Applications*, vol. 18, no. 1, pp. 27–48, Jan 2001.

[131] J. Luo, E. Mohyeldin, N. Motte, and M. Dillinger, "Performance investigations of ARMH in a reconfigurable environment," in *Proc. of 16th Workshop IST SCOUT*, 2003.