Departament de Teoria
del Senyal i Comunicacions

UNIVERSITAT POLITÈCNICA DE CATALUNYA

Contribution to Resource Management in Cellular Access Networks with Limited Backhaul Capacity

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

Hiram Galeana Zapién

Thesis Advisor:
Dr. Ramon Ferrús Ferré

Barcelona, January 2011

To the memory of my parents.
To my wife and daughter.

Summary

The radio interface of cellular mobile communication systems is normally considered as the only capacity limiting factor in the radio access network. However, as enhanced radio interfaces have been deployed, and mobile data and multimedia traffic increases, a growing concern is that the backhaul of the cellular network can become the bottleneck in certain deployment scenarios. In this context, the thesis focuses on the development of resource management techniques that consider a joint radio and backhaul resource management. This leads to a new paradigm where backhaul resources are considered not only at the network dimensioning stage but are included in the resource management problem.

Over such a basis, the first aim of the thesis is on evaluating capacity requirements in mobile backhaul networks that uses IP as a transport technology, attending to latest mobile architectural trends. In particular, it analyzes the impact of an IP-based transport solution on the capacity needed to meet QoS requirements in the mobile backhaul network. The evaluation is carried out in the context of the UMTS terrestrial access network, where a detailed characterization of the Iub interface is provided. The analysis of capacity requirements is conducted for two different scenarios: dedicated channels and high-speed channels. Afterwards, with the aim of fully exploit available resources in the radio access and backhaul, this thesis proposes a coordinated access resource management framework where the main idea is to incorporate transport network metrics within the resource management problem. In order to assess the benefits of the proposed resource management framework, this thesis concentrates on the evaluation of the base station (BS) assignment problem, as a strategy to distribute traffic among BSs based on both radio and backhaul loads. This problem is initially analyzed considering a generic radio access network by defining an analytical model based on Markov chains. This model allows us to compute the capacity gains that can be achieved by the proposed BS assignment strategy. Then, the analysis of the proposed BS assignment strategy is extended to cover specific radio access technologies. In particular, in the context of WCDMA networks this thesis develops a simulated-annealing-based BS assignment algorithm aimed to maximize a defined utility function, which reflects the availability of radio and transport resources. Lastly, this thesis tackles the design and evaluation of a backhaul-aware BS assignment algorithm for future OFDMA-based broadband cellular systems. In this case, the BS assignment problem is modeled as an optimization problem using a utility-based framework and resource cost functions, imposing constraints on both radio and backhaul resources, and mapped into a Multiple-Choice Multidimensional Knapsack Problem (MMKP). Then, a novel heuristic BS assignment algorithm is developed, evaluated and compared to classical schemes based exclusively on radio conditions. The conceived algorithm is based on the use of Lagrange’s multipliers and is aimed to simultaneously exploit both radio interface and backhaul load balancing.

Resumen

La interfaz radio de los sistemas de comunicaciones móviles es normalmente considerada como la única limitación de capacidad en la red de acceso radio. Sin embargo, a medida que se van desplegando nuevas y más eficientes interfaces radio, y de que el tráfico de datos y multimedia va en aumento, existe la creciente preocupación de que la infraestructura de transporte (backhaul) de la red celular pueda convertirse en el cuello de botella en algunos escenarios. En este contexto, la tesis se centra en el desarrollo de técnicas de gestión de recursos que consideran de manera conjunta la gestión de recursos en la interfaz radio y el backhaul. Esto conduce a un nuevo paradigma donde los recursos del backhaul se consideran no sólo en la etapa de dimensionamiento, sino que además son incluidos en la problemática de gestión de recursos.

Sobre esta base, el primer objetivo de la tesis consiste en evaluar los requerimientos de capacidad en las redes de acceso radio que usan IP como tecnología de transporte, de acuerdo a las recientes tendencias de la arquitectura de red. En particular, se analiza el impacto que tiene una solución de transporte basada en IP sobre la capacidad de transporte necesaria para satisfacer los requisitos de calidad de servicio en la red de acceso. La evaluación se realiza en el contexto de la red de acceso radio de UMTS, donde se proporciona una caracterización detallada de la interfaz Iub. El análisis de requerimientos de capacidad se lleva a cabo para dos diferentes escenarios: canales dedicados y canales de alta velocidad. Posteriormente, con el objetivo de aprovechar totalmente los recursos disponibles en el acceso radio y el backhaul, esta tesis propone un marco de gestión conjunta de recursos donde la idea principal consiste en incorporar las métricas de la red de transporte dentro del problema de gestión de recursos. A fin de evaluar los beneficios del marco de gestión de recursos propuesto, esta tesis se centra en la evaluación del problema de asignación de base, como estrategia para distribuir el tráfico entre las estaciones base en función de los niveles de carga tanto en la interfaz radio como en el backhaul. Este problema se analiza inicialmente considerando una red de acceso radio genérica, mediante la definición de un modelo analítico basado en cadenas de Markov. Dicho modelo permite calcular la ganancia de capacidad que puede alcanzar la estrategia de asignación de base propuesta. Posteriormente, el análisis de la estrategia propuesta se extiende considerando tecnologías específicas de acceso radio. En particular, en el contexto de redes WCDMA se desarrolla un algoritmo de asignación de base basado en simulated-annealing cuyo objetivo es maximizar una función de utilidad que refleja el grado de satisfacción de las asignaciones respecto los recursos radio y transporte. Finalmente, esta tesis aborda el diseño y evaluación de un algoritmo de asignación de base para los futuros sistemas de banda ancha basados en OFDMA. En este caso, el problema de asignación de base se modela como un problema de optimización mediante el uso de un marco de funciones de utilidad y funciones de coste de recursos. El problema planteado, que considera que existen restricciones de recursos tanto en la interfaz radio como en el backhaul, es mapeado a un problema de optimización conocido como Multiple-Choice Multidimensional Knapsack Problem (MMKP). Posteriormente, se desarrolla un algoritmo de asignación de base heurístico, el cual es evaluado y comparado con esquemas de asignación basados exclusivamente en criterios radio. El algoritmo concebido se basa en el uso de los multiplicadores de Lagrange y está diseñado para aprovechar de manera simultánea el balanceo de carga en la intefaz radio y el backhaul.

Acknowledgements

I would like to express my gratitude to my thesis advisor Dr. Ramon Ferrús, who has given me an excellent guidance and support throughout the realization of my Ph.D. education. This thesis would not be possible without his help at each and every step. I also want to thank the National Council of Science and Technology (CONACYT) of México for the financial support to study abroad. I would like to extend my gratitude to past and current colleagues at the Mobile Radio Communication Research Group, for all the good experiences shared during my stay in Barcelona. I want to thank to my family for their love and endless support. My deepest thank goes to my wife and best friend Araní for encouraging me to pursue this goal. Nothing I can say can do justice to how I feel about her patience, love and constant support.

Contents

List of Figures

List of Tables

List of Abbreviations

1G	First generation wireless systems
2G	Second generation wireless systems
3G	Third generation wireless systems
4G	Fourth generation wireless systems
3GPP	3rd Generation Partnership Project
AAL2	ATM Adaptation Layer 2
AC	Admission Control
aGW	Access Gateway
AF	Assured Forwarding
AMC	Adaptive Modulation and Coding
AMR	Adaptive Multi-Rate
AP	Access Point
AQM	Active Queue Management
ATM	Asynchronous Transfer Mode
B3G	Beyond 3G
BB	Bandwidth Broker
BE	Best Effort
BS	Base Station
BS	Bearer Selection (Chapter 4)
BS_CS	Best Server Cell Selection
CARM	Coordinated Access Resource Management
CC	Congestion Control
CCI	Co-Channel Interference
CDMA	Code Division Multiple Access
CIP	Composite IP
CRRM	Common Radio Resource Management
CN	Core Network
CoMP	Coordinated Multipoint Transmission
DCH	Dedicated Channel
DCH-FP	DCH Framing Protocol
DiffServ	Differentiated Services
DSL	Digital Subscriber Line
EDGE	Enhanced Data rates for GSM Evolution
EF	Expedited Forwarding
FDMA	Frequency Division Multiple Access
FP	Framing Protocol
GERAN	GSM EDGE Radio Access Network

GPRS	General Packet Radio System
GSM	Global System for Mobile Communications
GTP	GPRS Tunneling Protocol
H-ARQ	Hybrid ARQ
HSDPA	High-Speed Downlink Packet Access
HSUPA	High-Speed Uplink Packet Access
HS-DSCH	High-Speed Downlink Shared Channel
IETF	Internet Engineering Task Force
IntServ	Integrated Services
IP	Internet Protocol
IP-RAN	IP-based Radio Access Network
JRT	Joint Radio and Transport
LBR	Load Balancing Radio
LIPE	Lightweight IP Encapsulation
LTE	Long Term Evolution
MAC	Medium Access Control
MBAC	Measurement-Based Admission Control
MC	Mobility Control
MCS	Modulation and Coding Scheme
MIMO	Multiple-Input Multiple-Output
MMKP	Multiple-Choice Multidimensional Knapsack Problem
MPL	Minimum Path Loss
MPLS	Multi-Protocol Label Switching
MWIF	Mobile Wireless Internet Forum
Node B	3GPP term for a base station
OFDM	Orthogonal Frequency Division Multiplexing
OFDMA	Orthogonal Frequency Division Multiple Access
PBAC	Parameter-Based Admission Control
PDCP	Packet Data Convergence Protocol
PDH	Plesiochronous Digital Hierarchical
PDU	Packet Data Unit
PDR	Per-Domain Reservation
PHB	Per-Hop Behavior
PHR	Per-Hop Reservation
PPP	Point-to-Point-Protocol
QoS	Quality of Service
RAB	Radio Access Bearer
RACH	Random Access Channels
RAN	Radio Access Network
RAT	Radio Access Technology
RC	Route Control
RL	Radio Load
RLC	Radio Link Control
RMD	Resource Management in DiffServ
RNC	Radio Network Controller
RNL	Radio Network Layer
RoHC	Robust Header Compression
RPS	Radio Packet Scheduling
RP_CS	Radio Prioritized Cell Selection

RRM	Radio Resource Management
RS	RAT Selection
SA	Simulated Annealing
SDH	Synchronous Digital Hierarchy
SGSN	GPRS Super Node
SINR	Signal to Interference plus Noise Ratio
SNR	Signal Noise Ratio
SONET	Synchronous Optical Network
TB	Transport Block
TDM	Time Division Multiplexing
TDMA	Time Division Multiple Access
TFS	Transport Format Set
TNL	Transport Network Layer
TP_CS	Transport Prioritized Cell Selection
TTI	Transmission Time Interval
UE	User Equipment
UMTS	Universal Mobile Telecommunications System
UTRAN	UMTS Terrestrial Radio Access Network
VoIP	Voice over IP
WiMAX	Worldwide Interoperability for Microwave Access
WCDMA	Wideband Code Division Multiple Access

1 Introduction

1.1. Scope and Motivation

Technological advances have facilitated the growth of mobile communications in the last few years. Recently, the use of more spectral efficient air interface technologies (e.g., orthogonal frequency division multiple access, OFDMA) has enabled the step towards the future broadband mobile communications to provide data rate speeds significantly higher than current wireless systems. In this sense, next-generation broadband mobile communication systems are envisaged to provide ubiquitous wireless access to high-speed mobile terminals. Besides the progress in the radio access, transport infrastructure within the cellular radio access network (RAN) has also experienced significant changes. In particular, the inclusion of internet protocol (IP) as a transport technology in the RAN profoundly changed the cellular network architecture and its related transport network protocols. It also enables the conception of flat network architectures where radio related functionalities could be distributed among different network elements in the RAN.

Along with the improvements of mobile access networks in these fronts (i.e., radio access and transport network), resource management techniques should be adapted accordingly. Resource management techniques are one of the key tools in mobile networks to efficiently manage available network resources. In this regard, resource management solutions for the transport network and air interface have been traditionally developed separately. Particularly, a set of functions are defined to adjust different radio/transport parameters to preserve the quality of service (QoS) of connections based on the status (e.g., availability of resources) of the network in the radio interface/transport part. In this context, this thesis proposes a novel coordinated access resource management (CARM) framework where resource management decisions in the radio access and/or transport network can be influenced by the status of both radio and transport resources (see Figure 1.1). The motivation behind such an approach is the fact that latest changes in the transport network, and particularly the continuous improvements of the radio access technology (RAT) are gradually shifting the resource bottleneck from the radio interface towards the transport network (also known as mobile *backhaul*). The proposed approach aims to capture the actual network conditions (in both radio and transport parts) within the decision-making process of resources so that available network resources could be fully exploited.

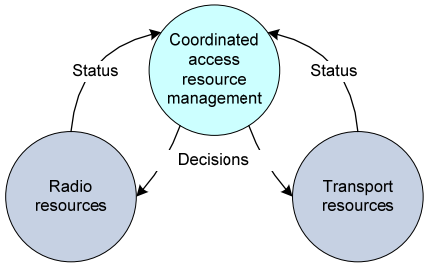


Figure 1.1: Proposed resource management framework.

The main goal of this thesis is the inclusion of transport resources within an advanced resource management framework. The idea behind the proposed approach is that in situations of shortage of backhaul capacity, more efficient decisions in a given resource allocation problem can be taken so that potential congestion events in the transport network could be prevented.

1.2. Thesis Outline

The structure of the thesis is illustrated in Figure 1.2. The remaining of this introductory chapter presents an overview of the thesis and a list of publications derived from the research work. Chapter 2 provides a brief description of the evolution of mobile communications systems, followed by an analysis of relevant advances of the RATs and transport network. Details of different transmission technologies and topologies, along with the motivation of IP within the transport network of the RAN are also discussed in the second chapter. Next, in Chapter 3, it is evaluated how the deployment of IP as transport technology in the RAN impacts on bandwidth capacity requirements in the transport network. The analysis of capacity requirements in an IP-based RAN is performed for the universal mobile telecommunications system (UMTS), and particularly in the context of the UMTS terrestrial access network (UTRAN). In Chapter 4, specific resource management functions to the transport network and air interface are identified. Afterwards, the proposed CARM approach is detailed along with the definition of several resource management functions that can work under the scope of the CARM framework. Among the defined functionalities, the cell selection, or base station (BS) assignment, problem is analytically formulated and evaluated. To this end, an analytical model based on multi-dimensional Markov chains is developed to assess the performance of three different cell selection approaches without focusing on a specific radio access technology (RAT). In Chapter 5 we formulate the BS assignment problem in the context of a system based on wideband code division multiple access (WCDMA). Here, a simulated-annealing algorithm that besides air interface resources, it also

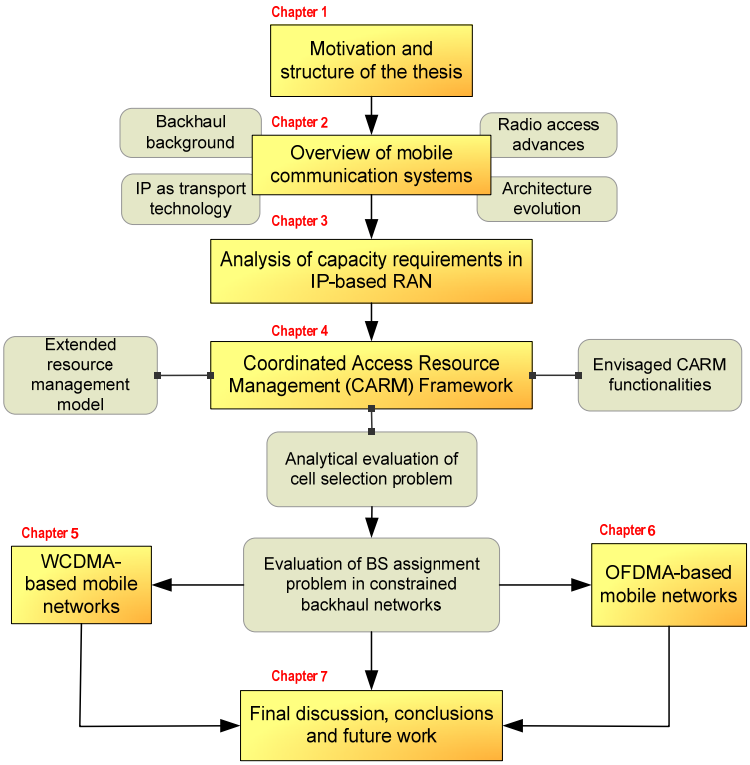


Figure 1.2: Structure of the thesis.

accounts for transport related constraints is developed, evaluated and compared to classical schemes exclusively based on radio aspects. Then, in Chapter 6 a more sophisticated BS assignment algorithm is developed for broadband OFDMA-based systems. Chapter 7 concludes the thesis with a summary of its main contributions and a brief outlook on future related research possibilities.

1.3. Main Outcomes

1.3.1. Publications

Part of the content of this thesis has either been published or submitted for publication during the period of research of the author in the Mobile Communication Research Group of the Department of Signal Theory and Communications, at Universitat Politècnica de Catalunya. A list of the papers is given as follows:

- "*Design and Evaluation of a Backhaul-Aware Base Station Assignment Algorithm for OFDMA-based Cellular Networks*", in IEEE Transactions on Wireless Communications (vol. 9, no. 10, pp. 3226-3237, ISSN: 1536-1276, DOI: 10.1109/TWC.2010.082110.091735), October 2010, by H. Galeana and R. Ferrús.
- *"Backhaul-aware BS Assignment: A Feasible Approach Towards More Efficient Backhaul Usage in Mobile Broadband Networks"*, submitted to IEEE Communication Letters, by H. Galeana and R. Ferrús.
- *"Enhanced Base Station Assignment Approach for Coping with Backhaul Constraints in OFDMA-based Cellular Systems"*, in proceedings of the European Wireless Conference (EW), April 2010, by H. Galeana and R. Ferrús.
- *"User Allocation Algorithm with Rate Guarantees for Multi-rate Mobile Networks with Backhaul Constraints"*, in proceedings of the IEEE Vehicular Technology Conference (VTC), April 2009, by H. Galeana, A. Lainz and R. Ferrús.
- *"A Cost-based Approach for Base Station Assignment in Mobile Networks with Limited Backhaul Capacity"*, in proceedings of the IEEE Global Communications (GLOBECOM), December 2008, by H. Galeana, F. Novillo and R. Ferrús.
- *"Performance Analysis of Transport and Radio Load Balancing Strategies for BS Assignment in Mobile Access Networks"*, in proceedings of the IEEE International Symposium on Personal Indoor and Mobile Radio Communications (PIMRC), September 2008, by F. Novillo, H. Galeana, A. Lainz, R. Ferrús and J. Olmos.
- *"A Base Station Assignment Strategy for Radio Access Networks with Backhaul Constraints"*, in proceedings of the ICT Mobile and Wireless Communications Summit, June 2007, by H. Galeana, F. Novillo, R. Ferrús and J. Olmos.
- *"Evaluation of a Cell Selection Framework for Radio Access Networks considering Backhaul Resource Limitations"*, in proceedings of the IEEE International Symposium on Personal Indoor and Mobile Radio Communications (PIMRC), Athens, Greece, 3-7 September 2007, by J. Olmos, R. Ferrús, and H. Galeana.
- *"Comparison of Transport Capacity Requirements of 3GPP R99 and HSDPA IP-based Radio Access Networks"*, in proceedings of the IEEE International Symposium on Personal Indoor and Mobile Radio Communications (PIMRC), September 2007, by H. Galeana, R. Ferrús and J. Olmos.

- *"Transport Capacity Estimations in Over-provisioned UTRAN IP-based Radio Access Networks"*, in proceedings of the IEEE Wireless Communications and Networking Conference (WCNC), March 2007, by H. Galeana, R. Ferrús and J. Olmos.

1.3.2. Research Projects

Part of the work done in the scope of the thesis has served as basis for the realization of some research activities that have been disseminated in technical reports of the following research projects.

- Sistema de Comunicaciones Móviles Profesionales de Banda Ancha TelMAX. National research project founded by the Center for the Development of Industrial Technology (CDTI) of Spain.
- Advanced Resource Management Solutions for Future All IP Heterogeneous Mobile Radio Environments (AROMA). European research project founded by the European Union within the context of The Sixth Framework Programme (FP6), Information Society Technologies (IST), Specific Targeted Research Projects (STREP). Ref. IST-4-027567.

2 *Mobile Communication Networks*

2.1. Introduction

In the last decades mobile communications has experienced an intensive development and have become one of the hottest areas in the field of telecommunications. The demand of new services and applications to mobile devices continues to grow at a remarkable rate. Nowadays, mobile communication devices play an important role of our daily life and they are seen as indispensable parts of the current society.

The history of mobile communication networks dates back to 1979, with the deployment of the first generation (1G) mobile communication system. This first operational mobile communication system, however, was based on analog signaling techniques. Developments in mobile communications progressed slowly in the early stages, but advances in digital signal processing made possible the enhancement of 1G systems. The development of the second generation (2G) systems began to be deployed around the world (e.g., GSM in Europe) in the early 1990s. As the demand of more efficient support of packet data services and higher data rates increases, 2G systems were evolved into what it is commonly referred to as 2.5G cellular systems (i.e., GPRS and EDGE systems).

Although 2.5G cellular systems aimed to meet the expectations of higher data transmission speeds to support the growing popularity of mobile data services, they did not achieve all the capabilities promised by third generation (3G) systems. The step towards 3G systems was fundamentally different from earlier transitions since it concentrates on providing significantly increased radio system capacities and per-user data rates by introducing an enhanced air interface technology based on wideband code division multiple access (WCDMA), and a new radio access network (RAN) architecture. In this context, the universal mobile telecommunication system (UMTS) standardized within the third generation partnership project (3GPP) constitutes the more representative standard of current 3G systems.

At present, research and development is being undertaken for the definition of the beyond 3G (B3G) and fourth generation (4G) standards in order to provide truly mobile broadband access. In this regard, the definition of future mobile communication systems is in general focusing on two major fronts in the RAN. First, air interface technologies are evolving to provide increased data rates and enhanced quality of service (QoS) capabilities. For instance, the first step taken by the 3GPP towards enhancing the air interface in the UMTS system was the introduction of high-speed downlink packet access (HSDPA) and enhanced uplink, referred to as high-speed uplink packet access (HSUPA). These technologies provide 3GPP standards with a radio access technology (RAT) that is highly competitive in the mid-term future. Nevertheless, as the demand in high data rate services and requirements of operators are dramatically increasing, the 3GPP started in recent years the specification of an evolved access technology referred to as long term evolution (LTE) [1], to ensure continued competitiveness with other systems, such as Mobile WiMAX [2], in a 10-

year perspective and beyond. The key enabling technology implementing the physical layer of both LTE and Mobile WiMAX systems is the orthogonal frequency division multiple access (OFDMA).

The second major front of evolution is related with the mobile network infrastructure. Specifically, the inclusion of internet protocol (IP) networking technologies in mobile networks is profoundly changing the overall network architecture and protocols. The architecture evolution mainly focuses on the adoption of flat IP network architectures where the radio functionality is pushed to the edge of the access network and the connectivity within the RAN is achieved through the use of IP as a network layer protocol.

2.2. Radio Access Networks

The cellular network architecture has two main components: the RAN and the core network (CN). The former one handles all radio-related functionality, while CN is responsible of routing calls and data connections to external networks. The RAN is the part of the mobile network that comprises a number of geographically dispersed base stations (BSs). The BS is the network element that communicates with mobile devices over the radio link (also known as the "air interface") by means a RAT. The BSs are connected to the rest of the elements in the RAN using different (wireless/wired) transmission media. This infrastructure is normally referred to as the *mobile backhaul network* or *transport network*. This section is devoted to provide a brief description of the access technology and the backhaul network.

2.2.1. Radio Access Technology

Multiple access techniques are used to allow many simultaneous users to access to the bandwidth radio spectrum of the radio communication system. There are three fundamental multiple access techniques to share the available spectrum: frequency division multiple access (FDMA); time division multiple access (TDMA); and, code division multiple access (CDMA). Additionally, OFDMA is seen as a hybrid access technique of FDMA and TDMA. We briefly detail these well know radio access methods in the following.

In FDMA-based radio system, the total available bandwidth is subdivided into a number of narrower band channels. Each user transmits and receives at different frequencies as each user gets a unique frequency slot. On the other hand, in TDMA the available spectrum is divided into multiple time slots, and each user is given a time slot to transmit and receive. Lastly, CDMA is a spread spectrum technique where the narrow band is multiplied by a large bandwidth signal that is a pseudo-random noise code. In contrast to FDMA and TDMA, in CDMA all users share the same frequency bandwidth and transmit simultaneously, but use different spreading code sequences to separate each user. In a CDMA-based system the transmitted signal is recovered at the received by correlating the received signal with the pseudo-random noise code used by the transmitter.

The above commented access technologies have been widely used in mobile communication systems. For instance, the GSM/GPRS mobile network standards make use of TDMA access technology, whereas the basis of 3G mobile systems is the CDMA access technology. Nowadays, the most promising multiple access technique already adopted by next generation mobile communication standards is OFDMA [3]. As the name implies, OFDMA is based on OFDM [4], which is a well known multicarrier transmission technique where the available (spectrum) bandwidth is divided into many subcarriers (each one being modulated by a low rate data stream). In order to achieve multiple user access, a set of subcarriers are allocated to users in the system.

In OFDMA, the time varying conditions of the radio channel allows for the exploitation of multiuser diversity, which allows a better usage of the available bandwidth. In particular, due that channel conditions of users are independent to each other, it is possible to select, for a given subcarrier and time instant, the user having good channel conditions (e.g., high signal to noise ratio). Furthermore, the fluctuations of channel conditions are also exploited by means of adaptive

modulation and coding (AMC) techniques [5], where the main idea is to use different modulation and coding schemes depending on the conditions of the radio channel. In this sense, users having good channel conditions would be allowed to transmit at higher data rates (e.g., using QPSK), while users with bad channel conditions transmit at lower data rates (e.g., using a 16QAM scheme).

Another important technology used along OFDMA to improve the performance of the radio access is the so-called multiple-input and multiple-output (MIMO) [6], that consist in the use of multiple antennas at both the transmitter and receiver in order to minimize errors and optimize data speed. MIMO is a higher spectral efficiency technology that offers significant increases in data throughput.

2.2.2. Mobile Backhaul Network

The term *mobile backhaul network* or simply *backhaul* is commonly used to refer to the part of the mobile access network that is responsible of the interconnection between the network elements in the RAN by means different topological configurations and transmission technologies. In the UMTS terrestrial access network (UTRAN), the backhaul links each Node B to the radio network controller (RNC). As well, in recent cellular network systems, the backhaul interconnects the BSs to the access gateway (aGW). For instance, the aGW would correspond to the ASN_GW network entity in Mobile WiMAX, or to the Serving Gateway in LTE network architecture.

Figure 2.1 illustrates a generic architectural structure of a mobile backhaul network. As seen in this figure, the hierarchical architecture of the transport network is divided into three stages: last mile, second mile, and backbone. These stages are related to the amount of traffic aggregation they support. Focusing on the backbone part of the mobile backhaul network, the transmission infrastructure used in this stage is normally optical fiber since high volumes of traffic are consolidated at each of the connection points. On the other hand, the transmission infrastructure in the backhaul network (i.e., last mile and second mile stages) is more difficult to deploy. This is because there are different factors that influence the selection of the physical medium to be deployed in these two stages to interconnect the different elements in the backhaul. This may include the BS density, terrain characteristics and distances, the target coverage area (urban or rural), and the availability and costs of transmission technologies. These factors influence the choice of technologies and design of the backhaul network. Furthermore, as mobile standards usually provide some flexibility in how the backhaul should be implemented, mobile operators can adopt different strategies and rely on available technologies in the market to build its own backhaul infrastructure, or also lease part of the transmission network to a carrier.

The backhaul network is one of the major contributors to the high costs of building out and running a cellular network. Some reports estimate that the backhaul constitute around 25% of the overall operational expenses incurred by cellular operators [7]. In fact, a key challenge to mobile operators is to reduce backhaul costs. In this sense, mobile operators are continuously seeking for cost-effective solutions for the backhaul network in order to squeeze more out from the available network resources. The following subsections introduce some basics of the backhaul infrastructure, particularly, the possible configuration topologies and the types of transmission technologies available for providing the connectivity in the backhaul network.

2.2.2.1. Backhaul Topologies

There are many factors that should be taken into account when selecting the topology for the backhaul network. One of the main factors to be considered is the amount of traffic that is expected to be carried on a given segment of the backhaul network. In fact, the two different stages involved in the backhaul (see Figure 2.1) are clearly differentiated by the level of supported traffic.

Different backhaul topologies can be used to deploy the backhaul infrastructure so that operators could collect or aggregate traffic from a large number of low speed bandwidth links (e.g. E1/T1 normally located in the last mile) into high-speed links (e.g., STM-1, STM-4, etc) of the

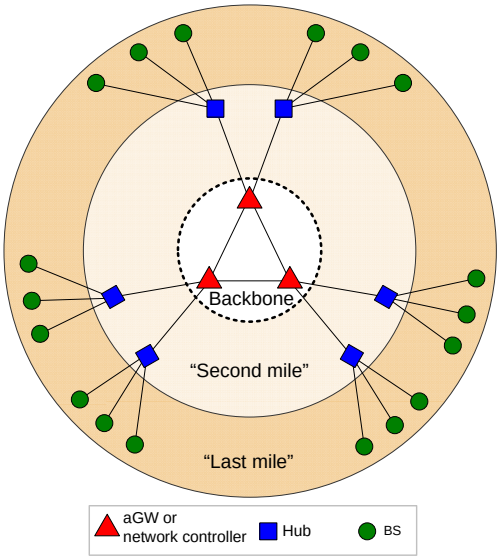


Figure 2.1: A schematic representation of mobile backhaul network [7].

backbone. The aggregation of traffic represents a good strategy for operators since backhaul infrastructure is very expensive, and thus it is required to achieve a more efficient use of transport resources in the backhaul and provide important cost reductions. The traffic aggregation functions can be deployed in some nodes at the different stages in the network architecture. The lowest level of traffic aggregation is located in the last mile stage of the access network architecture. Here, the aggregated traffic from all mobile users being served by each BS is transported to the next stage via narrowband links. The second mile provides the first level of traffic aggregation in the access network since it consolidates traffic from a number of BSs. The consolidation of traffic is performed at hub sites, which additionally could also integrate a BS in the same site. Finally, second-mile links are connected to high-capacity aggregation nodes of the backbone stage.

The effectiveness level of traffic aggregation depends on the network topology used to interconnect network elements in the RAN. Figure 2.2 illustrates the commonly used topologies for the backhaul network. As seen in this figure, different topological configurations like star, ring, tree, chain or a mixed combination can co-exist in the backhaul network [8]. It is worth noting that since there is no standardized solution to the backhaul network, the exact topology would depend on specific needs and requirements of operators, being those configurations that allow a high degree of traffic concentration the most attractive from an economical point of view. For instance, a set of BSs connected in a tree topology may require more bandwidth resources in the transport network than an interconnection scheme where a set of BSs are connected to a single concentration point in a star topology form, as illustrated in Figure 2.2. This is because traffic aggregation is more efficiently in scenarios where greater concentration of traffic is achieved. Finally, traffic aggregation is possible at different levels of the network, either by installing independent or integrated equipments in the sites of the BSs.

2.2.2.2. Backhaul Transmission Technologies

The transmission technologies that can be used by mobile operators to implement the physical layer in the backhaul can be categorized as follows.

- Leased copper lines. Leased lines are today extensively used for mobile network backhaul because they can save a mobile operator from having to manage its own transmission infrastructure. Most of the leased copper lines are based on the legacy time division multiplexing (TDM) technology. The main disadvantages of leased lines are their high cost and the lack of granularity of the bandwidth provided, that is, the vast majority of last mile leased infrastructure is TDM-based copper (e.g., E1/T1 leased lines over copper wires). Although the

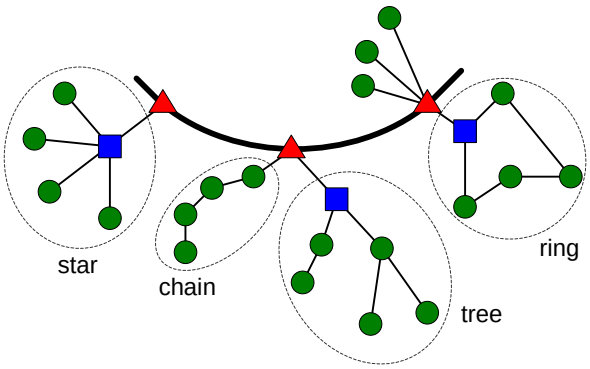


Figure 2.2: Transmission topologies in the mobile backhaul network.

use of leased T1/E1 lines can make economic sense when only a few are required, the cost of this approach scales linearly with capacity, making it poorly suited for backhaul in a 3G/4G environment.

- Microwave radios. Microwave radios interconnect cell sites in the backhaul via point-to-point or point-to-multi-point links. Microwave links are the most common self-owned infrastructure technology in the mobile backhaul network (some reports indicate that around 60% of worldwide cell-sites are connected via this technology [9]). This transmission technology is an attractive way for operators to reduce backhaul-related operating expenditure (i.e., they are generally less expensive to operate).

- Optical fiber. Optical fiber is a high bandwidth transmission technology. Due to its associated cost, this transmission technology is commonly used in the backbone network where there exists a high level of traffic aggregation.

- DSL backhaul. Due to their modest cost structure, digital subscriber line (DSL)-based technologies have become a prominent candidate for cellular traffic backhaul that is not delay sensitive. DSL is widely used for residential broadband and its current data-only requirement makes the DSL suitable for fem-to-cell as well as for best effort cellular applications.

- Satellite backhaul. For very remote locations, satellite links are the only viable means of backhaul from the cost performance perspective. The absolute cost is nevertheless still relatively high.

2.2.3. IP in Mobile Communications Networks

Mobile backhaul networks have traditionally been realized using TDM solutions and asynchronous transfer mode (ATM). These technologies are normally transported over the plesiochronous digital hierarchical (PDH) solutions in the backhaul and SDH/SONET solutions in the backbone network. TDM solutions were initially successfully adopted for circuit-switched voice traffic but, however, they are no longer appropriate to neither meet the capacity requirements of broadband mobile data networks nor cope with the dynamic and highly fluctuant traffic pattern of future broadband cellular systems. In order to meet these requirements, traditional TDM-based backhaul networks are currently migrating towards “All-IP” packet-based networks. The concept of All-IP in mobile networks is used to refer to a mobile system that provides IP-based services and that makes use of IP as the main transport technology in the access network.

The deployment of IP in mobile networks is motivated by the fact that IP has become the de-facto networking protocol. However, its deployment in mobile networks has taken more time mainly because of the inherent complexity of the cellular network architectures. The Mobile Wireless Internet Forum (MWIF) consortium, currently merged with the Open Mobile Alliance

[10], was among the first international industrial forums that investigated an alternative transport technology for mobile communication systems. In this regard, the growing popularity and deployment of the internet protocol (IP) in telecommunication networks influenced the mobile communication sector to investigate the introduction of IP in 3G mobile communication systems. The performance analysis carried out by different companies of the MWIF consortium concluded that IP was a viable option for implementing the transport network of 3G mobile RANs [11]. One of the most important and substantial drivers behind the inclusion of IP in the RAN is the potential cost savings. In brief, an IP-based RAN has the following remarkable advantages over ATM: cost effectiveness, wide deployment, easier network maintenance and management, as well as scalability and flexibility features for a smoother evolution towards next generation networks. Furthermore, the statistically multiplexing feature of IP can effectively reduce bandwidth occupation and improve the utilization of transmission resources in the RAN.

In an IP-based RAN the different available IP protocols are used to support one or more key aspects of network operations. These may include network layer routing and transport of user packets throughout the RAN and support of QoS [12]. Nevertheless, the use of IP within the mobile access networks also involves different challenges such as strict time requirements (i.e., voice and other delay sensitive traffic should be transported in a timely manner). This is particularly important because IP by itself does not offer QoS guarantees. Hence, the deployment of IP in the RAN requires the use of additional protocols to meet the different requirements of the RAN: provide appropriate level of end-to-end QoS to traffic flows by means of a QoS architecture; and efficient network utilization by means of resource management.

It is worth noting that the use of IP inside the access network is a different topic from the provision of IP-based services. The latter refers to the provision of Internet connectivity to mobile users by means the mobile access network, whereas the former one is related to the deployment of IP as a network layer protocol to interconnect the different network elements in the RAN. In the following, we summarize the two primary modes in which IP can be deployed within the mobile access network, namely the *transport mode* and the *native mode* [13], [14].

2.2.3.1. Transport mode.

Under the transport mode, IP is merely used as transport technology, and hence the destination IP address of an end-user is not required to perform packet forwarding decisions in intermediate nodes within the access network. Instead, packets are transported within the network in an encapsulated manner by an intermediate layer. This enable to keep many of the legacy components of the 3G access network unchanged while upgrading just the transport network. In order to illustrate the use of IP as transport technology, herein we consider the case of the UTRAN.

The transport technology defined by 3GPP for initial UTRAN deployment (R99 specifications) was based on ATM, using the ATM Adaptation Layer 2 (AAL2) [15]. From release 5 (R5) IP was standardized as an alternative transport solution [16]. The suitability of IP as a transport technology in UTRAN was firstly evaluated in [11], and also assessed in other subsequent works published in the literature [17], [18], [19]. These works evaluated the attractiveness of IP in the UTRAN through comparative studies versus the performance of the ATM/AAL2, concluding that the IP technology is an efficient transport network solution in terms of link utilization and delay performance in the UTRAN. Furthermore, these contributions also pointed out that IP constitutes a more attractive solution than ATM/AAL2 from the cost point of view due that the success deployment of IP in data networks have lowered the price of IP networking equipments.

The standardization of IP as a transport option is intended to be layer 2 independent, which gives more flexibility to operators in choosing the link layer technologies for the backhaul. According to R5 specifications, the introduction of IP requires that the radio network layer (RNL) functional split shall not be changed depending on the transport network layer (TNL) technology. Note that if the RNL is different for different TNLs, backward compatibility is lost or complicated and an implementation is potentially complicated when changing transport. The general structure

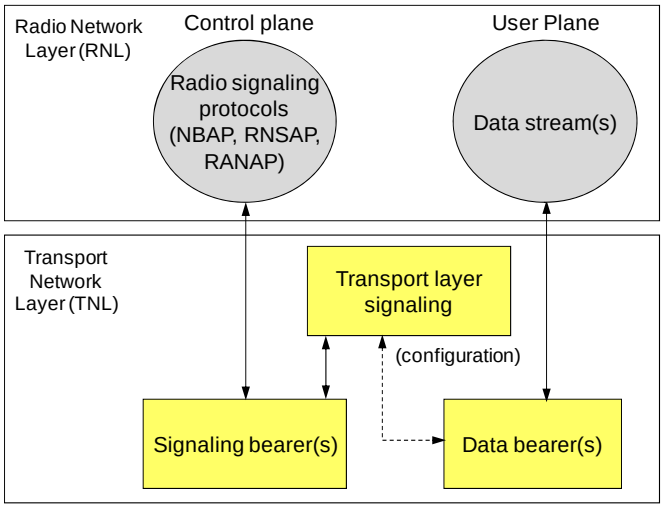


Figure 2.3: Radio and network layers structure in UTRAN.

of UTRAN interfaces is based on the principle that the layers and planes are logically independent of each other. This architectural principle of separation of RNL and TNL is illustrated in Figure 2.3 [20]. All UTRAN related issues are visible only in the RNL, while the TNL represents the transport technology used in UTRAN interfaces (e.g., Iu, Iur, Iub). The control plane includes the UTRAN application protocols, RANAP (over Iu), RNSAP (over Iur) or NBAP (over Iub), and the signaling bearer for transporting these application protocol messages. The user plane includes the data stream(s) and the data bearer(s) for the data stream(s). The data stream(s) is/are characterized by one or more frame protocol (FP) specified for that interface.

Figure 2.4 illustrates an IP-based UTRAN reference architecture. The underlying IP transport technology in the backhaul is responsible for transporting user and control planes, as well as data and O&M information between Nodes B and RNC in the UTRAN. In this architecture, we can distinguish between end nodes (hosts) and intermediate nodes or routers, responsible for forwarding IP packets. In this sense, Nodes B will be usually equipped with an IP host but, in case a given Node B serves as an intermediate node within the backhaul network, it will be integrated with an IP router. In an IP-based transport, the Iub interface between RNC and Node B is supported over the IP transport technology. This implies that FP frames transported in the Iub interface are encapsulated into IP packets, whose destination address in the uplink direction is fixed and refers to the RNC. On the other hand, the destination address of IP packets in the downlink direction belongs to the Node B currently serving a given mobile user. The determination of the serving Node B(s) is made by the RNC using maintained link-layer state for all the currently served mobile users.

The TNL in UTRAN is responsible of providing the appropriate QoS requested by the RNL. For instance, WCDMA radio control functions impose stringent delay requirements on the TNL. Notice that for outer-loop power control to function properly, the round trip delay is preferred to be less than 50 ms, corresponding to a one way delay of 25 ms [21]. In this sense, one of the main challenges of the IP-based transport is that it should meet the QoS requirements in a cost-effective

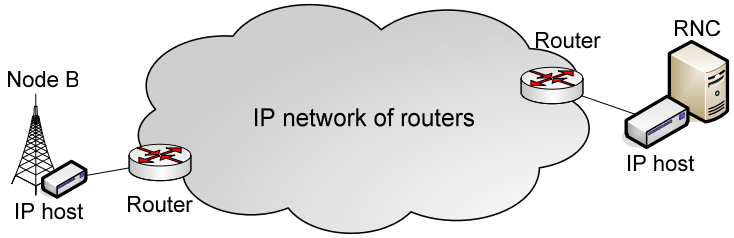


Figure 2.4: Illustration of IP hosts and routers in the IP-based UTRAN.

way in terms of efficiency and maximal resource utilization, particularly in narrowband backhaul links. However, the basic protocol stacks defined by 3GPP for both user and control plane introduce serious overheads [22], and therefore it is necessary the use of mechanisms such as robust header compression (RoHC) and multiplexing to overcome this disadvantage [16].

Another requirement for the TNL is related to the differentiation of multiple real-time traffic classes in the access network. In this sense, to assure optimal operation for real-time traffic with the most stringent delay requirements, small packets (e.g., voice, signalling and synchronization) should be differentiated from real-time data traffic in routers [23], [24], making possible to give different preferences to each traffic type. However, even if voice traffic has higher priority than real-time data traffic, the delay of voice packets can significantly increase due the transport of large data packets. For instance, a 1000 byte long data packet may increase the voice delay by 4 ms at an E1 link. This effect can be minimized by means of segmentation methods to split large packets (i.e., FP packet data units) in smaller segments so that the transmission delay is kept low [22].

2.2.3.2. Native mode.

In the native mode IP packets are transported within the RAN using regular IP forwarding (i.e., based on the destination address). The main characteristic of this mode is that no additional intermediate transport-oriented layers are needed. The absence of intermediate protocol layers inherently implies a higher efficiency since transport overheads are reduced. Besides, a mobile network employing this mode do not require network elements that are specific to any RAT, and hence can be used by the operator for building a converged network with multiple access technologies.

Nevertheless, this mode constitutes a drastic upgrade to the mobile network due that it would imply the replacement of most of the current specific nodes in the RAN by widely used standard IP equipments. As well different functionalities used for packet forwarding, such as the GPRS Tunneling Protocol (GTP), would require to be replaced by other protocols or functions.

2.2.4. Trends in Network Architecture

The cellular systems were initially designed using a hierarchical network architecture, were specialized network elements collectively form a hierarchical cellular system. However, with the advent of newer air interface technologies, the hierarchical approach (originally conceived for voice service and low-speed data) does not constitute a sustainable approach for the next generation of mobile communications systems, as explained in the following.

When 2G and later 3G mobile communication systems were designed, there were two reasons to make them hierarchical. First, when cellular systems were first devised, sharing the expensive vocoders over a large number of users led to considerable savings when deploying such cellular systems. The savings results from not having to deploy such expensive vocoders in all the cell sites. Secondly, since wireless voice transmissions are compressed, fewer bits needed to be transmitted over the backhaul network, so that more voice calls could be handled on a single T1 or E1 connection. Later on, with the introduction of CDMA systems, the hierarchy had an additional benefit for performing macro-diversity transmission and reception. Here, downlink data is prepared by a central anchor and then distributed to a number of BSs for simultaneous transmission over the air interface. A mobile can thus combine the information from multiple “legs” before decoding the information. This type of transmission is particularly helpful in combating fast-fading radio channels. Similarly, in the uplink, a central controller, such as the RNC in the UMTS system, can select the best voice uplink packet before transmitting the received packets to the vocoders. In this case, all protocol processing is performed centrally in the controller entity.

It is currently argued, however, that the main reasons to build cellular systems in a hierarchical manner have disappeared [25]. First, advances in electronics have made the cost argument disappear, and thus there is no reason the cost of electronics needs to dictate the cellular system

architecture. As a result, every BS today can be equipped with cost effective processing environments to perform all access specific functions (including protocol processing). Second, with the transition from circuit-switched voice to voice over IP (VoIP), voice streams are already compressed over the backhaul between the end-points and, with the increase of data usage, voice streams are expected only comprise a small part of the overall bandwidth. Lastly and more importantly: instead of using spatial diversity, that uses multiple receiving antennas, to combat fast fading, time diversity (i.e., fast retransmission from one BS) can be used instead. This latter argument is backed up with the progressive adoption of AMC techniques, fast scheduling and Hybrid ARQ (H-ARQ) in the LTE and Mobile WiMAX solutions.

In order to efficiently deliver mobile broadband services, operators require a network infrastructure that simultaneously provides lower costs, lower latency, and greater flexibility. The key to achieving this goal is the adoption of flat, all-IP network architectures. Particularly, with the shift to flat network architectures, mobile operators can:

- Reduce the number of network elements in the data path to lower operations costs and capital expenditure.
- Partially decouple the cost of delivering service from the volume of data transmitted to align infrastructure capabilities with emerging application requirements.
- Minimize system latency and enable applications with a lower tolerance for delay; upcoming latency enhancements on the radio link can also be fully realized.
- Evolve radio access and packet core networks independently of each other to a greater extent than in the past, creating greater flexibility in network planning and deployment.
- Develop a flexible core network that can serve as the basis for service innovation across both mobile and generic IP access networks.
- Create a platform that will enable mobile broadband operators to be competitive, from a price/performance perspective, with wired networks.

2.3. Backhaul as New Network Bottleneck

The idea that the backhaul network can constitute a capacity bottleneck in the mobile RAN is in general difficult to understand and accept. This is because the air interface has been traditionally assumed the only limiting factor of resources in the access network. Nevertheless, there are strong arguments that support the idea that, in some deployment scenarios, the backhaul network could become the network bottleneck.

At the initial rollout of 3G systems, mobile operators reuse as much as possible the backhaul infrastructure from legacy 2G systems. As commented before, the infrastructure in legacy systems is mostly based on TDM solutions, which are not appropriate to meet capacity requirements of emerging broadband communication systems [26], [27]. Therefore, considering the huge impact that backhaul has on operating and capital expenditures, mobile operators are carefully reviewing their cellular backhaul strategies before making further network infrastructure upgrades. In fact, migration to more cost-effective technologies in the backhaul is expected to be carried out gradually. It is worth noting that while mobile operators would prefer to deploy optical fiber to more cell towers and thus have enough backhaul capacity at each cell site, the fact is that there are far too many cells for this to be a near-term strategy to solve the bandwidth problem in the backhaul. It is clear that is not cost-effective to do so due that the deployment of fiber optic cables would require substantial investments.

Nowadays, the rollout of more spectral efficient air interface technologies is imposing stringent capacity requirements to the backhaul network. One of the main challenges of this trend for mobile

operators is how to properly scale the backhaul network to support the ever increasing air interface capabilities [28]. As the radio access has been improved over the years, and demand for higher throughput and higher-data-rate services increased, the bottleneck of the mobile access networks is progressively being shifted from the radio interface towards the backhaul network. In this context, a backhaul network able to cope with peak data rates of the air interface (as it was dimensioned in legacy GSM networks) no longer constitutes an efficient option. This is because the peak rate supported in the cells of next generation broadband systems could be quite high compared to the served mean aggregate rate due to the use of AMC and soft-reuse techniques. AMC techniques allow to assigns the most appropriate modulation and coding scheme (MCS) to users depending on their channel conditions. By means of AMC, users enjoying good radio conditions can use a high order modulation scheme with low coding redundancy in order to transmit at high data rates. As a result of the use of this type of techniques, the aggregate traffic rate supported in a given cell could eventually be close to the peak rate of the cell depending on users' spatial distribution (i.e., most served users close to the BS and enjoying good radio conditions).

All reasons stated above make us to believe that, even though mobile operators are going to progressively introduce more cost-effective transmission technologies to cope with fast growing mobile traffic along with newer radio equipment, backhaul limitations may not be discarded in some network deployments [29], [30], [31]. In this context, best practices for efficient backhaul design have been recently issued by NGMN Alliance and there is an increasing number of solutions pushing for the adoption of more cost-effective transmission technologies than those used in most current deployments [32], [33] along with new resource management functionality specifically tailored to tackle backhaul congestion (e.g., [34]). As a matter of fact, flow control mechanisms have been already introduced in current mobile networks to partially mitigate traffic peaks in the backhaul at the expenses of an increased delay in some services [35], [36]. Attending to previous considerations, cellular network capacity limitations due to a shortage of backhaul capacity may not be underestimated in some network deployments.

2.3.1. Resource Management

Considering that the backhaul network is fast becoming a new network bottleneck, resource management solutions play an important role in order to make more efficient use of the available backhaul capacity and avoid eventual congestion situations. Resource management in IP-based backhaul networks aim to maximize the network capacity while maintaining the levels of QoS requested by the connections in the access network. Although there is not a general categorization for describing resource management approaches in the IP-based transport, in this section we enumerate different approaches considered within a resource management framework aimed to manage transport resources within the backhaul.

2.3.1.1. Admission Control Mechanisms

Admission control (AC) mechanisms to the transport network of a mobile RAN has been mainly addressed focusing on the case of ATM transport and less attention has been paid to the case of IP-based transport. There are two basic approaches to the admission control problem in the IP-based transport network [37]:

- Parameter-based admission control (PBAC). This approach computes the amount of network resources required to support a set of flows given a priori known flow characteristics, that is, traffic descriptors. PBAC algorithms can be generally analyzed by formal methods due to the deterministic nature of the traffic descriptors. This approach can be realized by using a bandwidth broker (BB) agent [38].
- Measurement-based Admission Control (MBAC). This approach relies on measurement of actual traffic load in making admission decisions. MBAC algorithms can only be analyzed

through experiments on either real networks or a simulator. Given the reliance of MBAC algorithms on source behavior that is not static in general, service commitments made by such algorithms can never be absolute. Measurement-based approaches to admission control can only be used in the context of service models that do not make guaranteed commitments (e.g., controlled load service in Integrated Servives –IntServ– QoS architectures).

There exist other AC approaches for IP-based transport networks [34], where the traffic load of backhaul bottleneck links is taken into account in the admission control decisions so that congestion events in the backhaul are prevented. The proposed solutions are used along with active queue management (AQM) that are intended to achieve high link utilization with low queuing delays.

2.3.1.2. Resource Management in DiffServ

Due that IP by itself lacks of QoS mechanisms, its implementation as a transport technology in mobile networks requires the use of a QoS solution. The Differentiated Services (DiffServ) approach [39] is one of the most recent IP QoS architectures proposed by the Internet Engineering Task Force (IETF) that can be used to provide QoS support in the IP-based transport network. The DiffServ approach in the UTRAN can classify and prioritize the traffic (e.g., at each Node B in the border of the network domain) into different classes, so that each traffic can be managed differently. For instance, expedited forwarding (EF) [40] can be used for the real-time traffic, while assured forwarding (AF) [41] or best effort (BE) can be used for the non-real real traffic. In this way, it is possible to give preferential treatment to real-time traffic over non-real time traffic. Nevertheless, the DiffServ architecture does not define resource reservation schemes, so that traffic flows of each class in the IP-based transport would compete for the available resources.

Over such a basis, the resource management in DiffServ (RMD) framework proposed in [42], [43], extends the DiffServ principles to provide resource management and admission control in the IP-based transport. In RMD, two types of resource reservation protocols are used: the Per Domain Reservation (PDR) protocol and the Per Hop Reservation (PHR) protocol. The PDR protocol is used for resource management in the ingress and egress nodes of the DiffServ domain, while the PHR protocol is used for resource reservation per DiffServ traffic class in the interior nodes in the communication path of the DiffServ domain.

The RMD proposal works as follows. Once a QoS request arrives at the ingress node (i.e., Node B), the RMD framework determines whether sufficient bandwidth resources are available at all interior nodes in the communication path between the ingress node (Node B) and egress node (RNC) to support the flow. In the case of PHR protocol, it is defined such that the availability of resources is checked by means of measurements before any “QoS requests” are admitted, without maintaining any PHR reservation state in the nodes in the communication path. The measurements are done on the average real traffic data load. The main advantage of this PHR group is that MBAC mechanisms have the potential of more efficient resource utilization [44]. The only state information maintained for the measurement based PHR relates to the measured available bandwidth in each interior node of the DiffServ domain. This is referred to as “per hop behavior” (PHB).

2.3.1.3. Congestion Control Mechanisms

The congestion control approaches in mobile access networks are restricted to monitoring the usage of air interface resources due that it is assumed that there exist a backhaul with sufficient capacity so that there is no packet losses in the transport network. However, as argued before, the enhancements of the air interface to achieve higher data rates, are imposing stringent capacity requirements to the backhaul network. Hence, if the aggregate traffic rate supported in a given cell exceeds the engineered backhaul capacity of the BS, a congestion situation in the backhaul network

will be produced. A congestion situation in the backhaul network may result in high packet losses and increased delays, which may potentially cause the violation of service commitments.

In this context, [45] proposes different resource management schemes where the main idea is to regulate the traffic to be supported in the backhaul network by adjusting the admission control criterion at the air interface so that congestion in the transport network is avoided. Particularly, the congestion control algorithms aim to maximize resources in the IP-based RAN while maintaining service commitments. The proposed schemes in [45] schemes use a measured packet loss rate to calculate a power scaling factor, which is then used to scale the admission control thresholds for the air interface, thereby indirectly adjusting the traffic load entering the IP-based RAN and reducing congestion.

Finally, the AROMA project [46] presented different proposals to this matter. In particular, the different proposals rely on radio resource management (RRM) and common RRM (CRRM) strategies to cope with overload/congested situations in the radio access of the RAN. On the other hand, congestion in the transport network can be addressed by either specific transport resource management mechanisms or relying on coordinated resource management strategies, to incorporate the status of transport network resources into their decision-making process. This latter idea is the one developed and evaluated in this thesis.

2.3.1.4. Other Methods for Transport Congestion Avoidance

There are other methods to appropriate detect congestion in the IP-based transport network, such as window-based, rate-based or combination of both

- A method to reactively reduce the overload in the transport network is proposed in [47]. This method is achieved by limiting the MAC-d transport format set (TFS) for nonguaranteed packet switched bearers, providing higher utilization of transport resources and a low level of losses when the offered load exceeds the engineered transport network capacity.
- In [48] and [49] present a method where the load offered to the transport network is proactively controlled by limiting the MAC-d transport formats. This overload avoidance approach relies on knowledge of the momentary available bottleneck capacity.
- In [50] it is proposed a rate-based flow control algorithm for HSDPA in order to provide high transport network utilization and maintain the delay and loss in the transport network low. The proposed method uses the transport network congestion detection functionality standardized by the 3GPP.

2.3.1.5. Capacity Over-provisioning

Sometimes it is argued that over-provisioning is the most straight forward way to solve the problem of capacity shortage in the backhaul. The capacity over-provisioning involves the provision of an amount of bandwidth in such a way that overload in the backhaul do not occur. This possibility is only feasible, from the economical point of view, to the backbone network where there exist high levels of traffic aggregation. In this context, Chapter 3 of the thesis is devoted to estimate the amount of capacity required in the backhaul network of the UTRAN.

2.4. Summary

This chapter presents an overview of the evolution of mobile communication systems, and also provides a description of different components on the mobile radio access network. Special emphasis has been placed on addressing the different advances on both the radio access technology and the so called backhaul network. This latter covers details such as typical network architectures, most commonly used transmission technologies, and also the role of the IP within the transport of

backhaul networks. This chapter also discusses the fact that backhaul network is fast becoming a new resource bottleneck in the RAN. Over such a basis, the chapter finally concludes presenting different approaches found in the literature aimed to manage transmission resources in order to avoid or alleviate potential congestion situations in the backhaul part of the mobile network.

3 *Transport Capacity Requirements of IP-based Radio Access Networks*

3.1. Introduction

The increasing role of IP technology in modern telecoms has been already discussed in previous chapter of this thesis, as well as the main reasons behind its introduction as a transport technology in Beyond 3G (B3G) mobile access networks. The IP technology became an alternative solution to asynchronous transfer mode (ATM) or circuit-switched networks. Among several benefits of an IP-based transport in mobile access networks, the most important one is the potential cost-savings it brings to mobile operators. The use of IP within the access networks facilitates the integration of different radio access technologies (RAT) operating over a unique transport, thus enabling the development of heterogeneous access networks. Nevertheless, an IP-based radio access network (RAN) also presents different challenges and particularly those related with the strict timing requirements that the transport network should fulfill. As well, advanced radio control functions require that the transport of user traffic over the transport network should fulfill stringent delay bounds, regardless if the traffic is real-time or non real-time [21], [51]. In this sense, an accurate capacity provisioning of transport resources in access networks is crucial. Furthermore, due to the costs associated to the transport network, it is mandatory that the IP-based transport network in RAN, dubbed as IP-RAN, should meet different timing requirements in a cost-effective manner in terms of efficiency and maximal utilization of available resources in the transport network.

In this chapter we aim to analyze the impact of the introduction of IP in the RAN on the transport capacity required to meet quality of service (QoS) requirements. The transport capacity requirements are evaluated in the context of an UMTS Terrestrial radio access network (UTRAN). We first develop a simulation model to characterize the IP-based transport in the UTRAN. Particularly, the Iub interface [52] that connects a base station (BS) with a radio network controller (RNC) in the UTRAN is modeled. Over such a basis, transport capacity requirements are estimated considering different mean traffic loads supported over the Iub interface so that delay bounds imposed to the transport network can be met.

Capacity requirements in the transport network are evaluated considering the case of "best-effort" traffic, which would correspond to the worst case scenario where the transport network and its associated protocol stack do not include QoS mechanisms. This method used to dimension the transport network capacity is referred to as "over-provisioning". At this regard, it is normally argued that over-provisioning in mobile access networks is not an economically viable solution [53]. However, there are not references devoted to quantify an over-provisioning solution for the dimensioning of an IP-based radio access network. This chapter tries to provide a useful insight into this issue. It is important to remark here that most of the studies dealing with over-provisioning planning in IP-based networks have been mainly addressed so far to backbone networks [54]. Thus, it is deemed mandatory to have a specific analysis in the context of an IP-RAN where, unlike IP-

based backbone networks, there exist particular conditions (e.g., different levels of traffic aggregation, characteristics of the applications using the transport, delay restrictions imposed by the radio applications) that should be taken into account in the capacity modeling and assessment process.

The analysis of transport capacity requirements in UTRAN is carried out for two different scenarios. In the first scenario, traffic is mainly supported over dedicated channels (DCHs) in the radio interface, whereas in the second scenario the traffic is supported over high speed downlink packet access (HSDPA) channels. Each scenario imposes quite different conditions and restrictions to the transport network. The rest of the chapter is structured as follows. Section 3.2 introduces the defined approach to estimate capacity requirements in the IP-RAN. Then, in Section 3.3 the evaluated scenarios are described, followed by the simulation setup in Section 3.4. Section 3.5 provides some numerical results and finally a summary of the chapter is given in Section 3.6.

3.2. Transport Network Characterization

The network dimensioning procedure followed in this chapter starts by defining a network model to characterize the IP-RAN. The transport network characterization is realized assuming different simplifications for the transport network. Then, the method followed to estimate the transport capacity requirements in the IP-RAN is detailed.

3.2.1. Simplified IP-RAN Network Model

Considering the common topologies and transmission technologies of current access networks, we now proceed to identify relevant parameters for characterizing the IP-based UTRAN. Figure 3.1 illustrates an example of the transmission components in the RAN. The interconnection of network nodes can be achieved using star, ring, tree, and mixed topological configurations (as discussed in Chapter 2). The reason behind the use of a particular topology configuration depends on the expected level of traffic aggregation in the access network. It can be seen in this figure that traffic aggregation functions can be deployed in some nodes from lower to upper levels in the access network. In fact, traffic aggregation is essential because it allows a more efficient use of bandwidth and also simplifies network management.

From the generic topology given in Figure 3.1, a simple IP-RAN model can be derived to represent any single path between a given Node B and its associated RNC. An example of how the topology can be mapped into a single-path IP-RAN model is shown as well in Figure 3.1. Although this is a simple model, it captures the main characteristics of an IP-RAN:

- Network links of different capacity can be configured.
- It is possible to capture different topologies by configuring the amount of traffic aggregation at each level. For instance, a star topology connecting several Nodes B can be modeled assuming a higher concentration of traffic than the case of a tree topology with few Nodes B.
- The size of the network can also be taken into account by means of adjusting the number of hops in the path between the Node B and RNC.
- The model can also capture the existence of low capacity last mile links (between a single Node B and the first level of traffic aggregation), as well as high capacity backbone links for traffic consolidation.
- The model can be consistently justified in terms of current topologies and available transmission technologies.

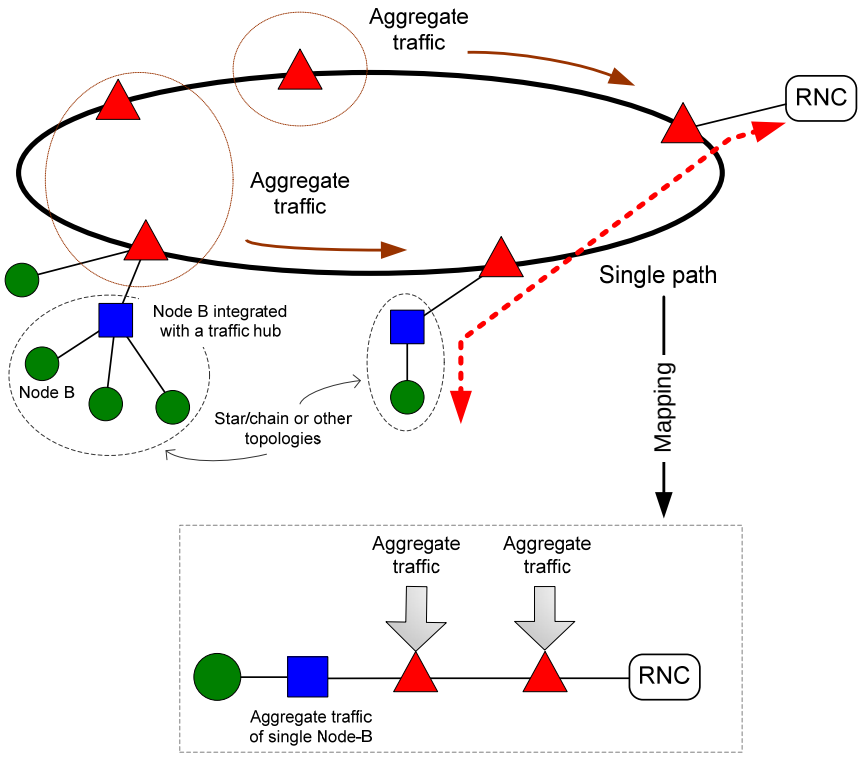


Figure 3.1: Mapping between a RAN topology and the IP-RAN network model for a single path within the topology.

3.2.2. Capacity Estimation Approach

The estimation of transport capacity requirements in the UTRAN is done assuming that the transport network in the UTRAN is entirely based on IP technology. This implies that nodes in the UTRAN are connected through an IP-based transport network responsible for transporting user plane and control plane, as well as operation and maintenance information in the UTRAN. Hence, the Iub interface between RNC and Node B is supported over the IP transport. It is worth noting that as standardization of IP as transport technology in mobile networks is intended to be independent of the layer 2, IP transport architecture is limited to nodes implementing an IP layer, as argued in section 2.2.3.1. Details of the transport network characterization and dimensioning approach are given in the following.

The dimensioning of an IP transport network leads to different approaches depending on the timeframe under consideration. For instance, a long-term analysis composed by days or even weeks requires the identification and characterization of the periods (i.e., hours, minutes) with highest traffic peaks. This is similar to the definition of the busy hour concept in the telephone networks. Over the considered long-term periods, a short-term approach in the order of minutes can be followed to analyze the system dynamics, such as the presence of traffic bursts, making possible to estimate the capacity required to prevent queue build-up or excessive delays.

In our study, the long-term characterization relies on the knowledge of the number of concurrent connections supported, and their traffic characteristics, in a given instant, in each Node B of a given UTRAN deployment. Using the traffic information and also knowing the network routing information, the amount of aggregated traffic traversing each link in the transport network can be estimated. Specifically, if we consider any small period of 5 minutes, we can state that the mean traffic rate supported in a given network path can be obtained by the sum of the mean values of the traffic generated by the concurrent connections traversing the single path.

Over such a basis, for a given number of concurrent connections supported in a given network path, or equivalently, for a given mean aggregated bit rate, the minimum capacity required in the path in order fulfilling a given delay bound can be estimated. The capacity requirement in the transport network is expressed in terms of a parameter referred to as "over-provisioning factor", denoted as β, which relates the excess of capacity required in a single transport network path to the mean aggregated traffic in the path, denoted as R_b. The transport capacity in a given network path can be expressed as follows:

$$C_{\text{path}} = R_b \cdot (1+\beta) = \left(\sum_i R_b^i \right) \cdot (1+\beta) \qquad (3.1)$$

where R_b^i represents the mean bit rate of user connection i traversing over the path. There are several factors affecting the required capacity: the aggregated mean bit rate, the protocol stack overhead, the traffic pattern characteristics (i.e., statistical properties) of the individual sources and particularly the considered QoS requirements that should be satisfied. As the delay is one of the main QoS restrictions that the transport network should meet, in our analysis we assume an upper bound on the delay experienced by IP packets transverse the path under observation. Furthermore, as the delay is random variable, it is assumed that the considered delay bound is satisfied whenever it is met for 99.9% of the total packets traversing the path in a given reference time period.

With respect to the dependency of the proposed capacity estimation approach to the traffic pattern characteristics of the sources, two different traffic models showing quite different dynamics are analyzed: voice traffic and web browsing. For each traffic type, a detailed characterization of the complete Iub protocol stack is addressed so that the mechanisms used at each layer are taken into account for reproducing the IP traffic supported in the transport network. The analysis of the two types of services is done separately, without mixing services. Therefore, the obtained results provide the link capacity needed to support a given amount of traffic of a particular type of service.

The analysis of capacity requirements in mixed services scenarios is out of the scope of this work. It is worth noting that in those situations the capacity requirements would be highly dependent on the QoS model used to share resources for each service type. Thus, transport capacity requirements can be different depending on how the QoS is handled on the IP-RAN. In any case, the results obtained in this chapter could be applied in the quantification of the amount of bandwidth needed for a given type of traffic in a best-effort network, as well as in a given multi-protocol label switching (MPLS) path, or in a given per-hop behavior (PHB) in the case of differentiated services (DiffServ) networks. It is worth noting here that the benefits arising from the statistical multiplexing of sharing different MPLS paths or PHBs in the same transport resources are not captured in our model but it still can be used as an upper bound.

Figure 3.2 illustrates the defined single path IP-RAN network model in the context of UTRAN. Here, the reference network path carries the mean aggregated traffic coming from the RNCs to the cloud of Node Bs (downlink direction). Notice that this path corresponds to the Iub interface in the UTRAN. In order to estimate the required capacity in the transport network, we first review the background of the Iub interface, and then define a capacity dimensioning approach.

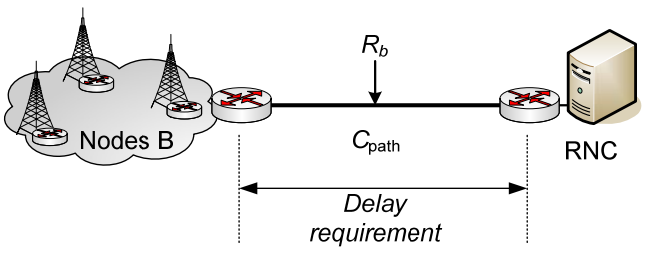


Figure 3.2: Single-path IP-RAN network model.

3.3. Evaluated Scenarios

In UTRAN, the data generated at higher layers is carried over the air interface using transport channels mapped onto different physical channels. In this sense, our focus in this chapter is on the performance analysis of the Iub interface when different transport channels are assumed. The Iub user plane includes various frame protocols (FPs), options for the support of random access channels (RACH/FACH), dedicated channels (DCH) and downlink shared channels (HS-DSCH). These latter channels are the ones used in HSDPA.

In this chapter, we evaluate transport capacity requirements in IP-based access networks assuming the use of DCHs and HS-DSCH channels. In the following details of these two scenarios are given, along with their corresponding protocol stacks and delay requirements.

3.3.1. Dedicated Channels Scenario

The objective in this scenario is to analyze the impact on transport network resources when DCHs are used in the air interface. These channels were introduced in the release 99 of the UMTS standard, and are commonly used in radio access bearers (RABs) for voice as well as for non real-time data services. In the release 99 of UMTS, all the medium access control (MAC) functionalities reside in the RNC, which performs packet scheduling based on the load measurements provided by the Node B and the user terminal. Figure 3.3 shows the user plane protocol stack of this scenario.

The radio link control (RLC) protocol handles segmentation/reassembly and retransmission of user data between the user terminal and the RNC. The MAC layer is responsible for mapping logical channels onto appropriate transport channels, as well as the selection of the data rates being used. At the output of the MAC layer bursts of transport blocks (TBs) are generated every transmission time interval (TTI) of the corresponding transport channel. Then, for each dedicated channel, the DCH framing protocol (DCH-FP) layer assembles the bursts transmitted in one TTI into one FP frame that is subsequently delivered to the IP transport network layer [55].

As described in [56], delay in the UTRAN depends on many factors and components such as the processing at each network node, transport network, and radio interface. However, there is no 3GPP specification defining specific delay requirements to be fulfilled over the Iub interface. The tolerable delay bounds in the transport network for dedicated channels are dependent whether the user traffic is real-time or not [21]:

- For real-time traffic, the tight end-to-end delay of the applications imposes a rather stringent requirement for the delay budget of UTRAN transport.
- For non-real time traffic, the UTRAN transport delay is governed by radio functions of outer-loop power control and soft-handover control.

The latter requirements result in particularly tight delay budgets to be satisfied. In [11] the considered delay requirement, between the Node B and RNC, for voice services and web-browsing

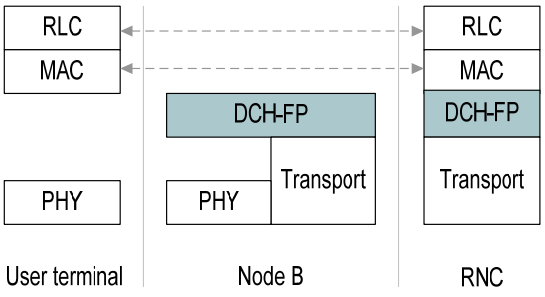


Figure 3.3: Protocol stack for dedicated channels.

services is around 5 ms around 50 ms, for 99.9% of transmissions, respectively. It is worth noting that in the case of voice, the service itself is the main factor that influences the decision of which delay requirement should be used. On the other hand, in the case of data services the radio functions are the limiting factor to take into account.

The above mentioned delay bounds are taken as a reference for this scenario as indicated in Table 3.1. Along with the previous values, and in order to assess the sensitivity of simulation results to the delay requirements being considered, we also consider softer delay restrictions: 20 ms is also considered for voice traffic, and tighter delay restriction of 5 ms for web-browsing traffic.

3.3.2. High-Speed Channels Scenario

The second scenario takes into account the use of high-speed channels over the air interface. We focus on HSDPA radio channels that lead to higher data rates to be supported in the Iub interface. Unlike the release 99 where MAC layer was completely located at the RNC, in this scenario a fast packet scheduling functionality is now introduced at the Node B (MAC-hs). In HSDPA, the Node B directly handles retransmissions using automatic request (ARQ) functionality, leading to faster retransmissions than the UMTS release 99. Under this scenario, different traffic patterns characteristics need to be transported in the access network due to the fact that radio packet scheduling is moved to the Node B. Furthermore, as HSDPA introduces new elements in the protocol architecture, this scenario aims to study the impact on transport network requirements on the Iub interface due to the use of high-speed channels. Figure 3.4 depicts the user plane protocol stack of this scenario.

It can be observed that the RNC retains only part of the dedicated MAC (MAC-d) mainly to handle logic channel multiplexing. The RLC layer stays unchanged, although some optimizations for real-time services such as VoIP are introduced. The use of buffering in the Node B permits a peak rate for the connection as high as the terminal and Node B capabilities allow. Having the transmission buffer at the Node B also requires flow control mechanisms to be applied in order to prevent an overload situation in the Node B buffer if radio conditions in the downlink make data to be retained at the Node B. Also in the downlink direction, the Node B buffer should not get empty as long as there are still user data pending for transmission at the RNC. The FP protocol specified to carry HSDPA data in the Iub interface is referred to as the high-speed downlink shared channel FP (HS-DSCH FP).

The delay requirements for HS-DSCH FP frames are mainly due to the service itself since neither outer-loop power control nor soft-handover are supported on these channels. According to this, softer delay restrictions than in the previous scenario can be considered for voice (e.g., 50 ms) and data traffic (e.g., 150 ms). However, attending to potential delay values given in [57] for long term evolution (LTE), values ranging between 1 ms and 15 ms are accounted for packet transmissions in the transport network. Thus, in accordance with these arguments, delay upper bounds for voice traffic are 5 ms and 50 ms, whereas for data traffic delays of 5 ms and 150 ms are considered in this scenario (see Table 3.1).

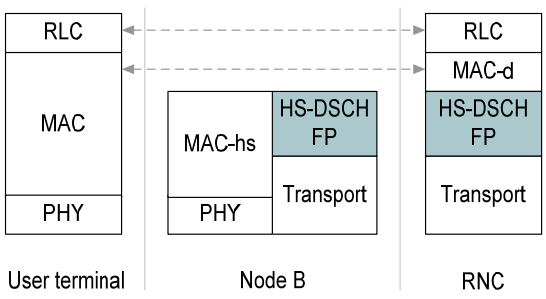


Figure 3.4: Protocol stack for high-speed channels.

Table 3.1: Delay requirements for voice and web-browsing traffic.

Transport scheme	Delay requirement	
	Voice	Web
Iub with DCHs	5-20 ms	5-50 ms
Iub with HS channels	5-50 ms	5-150 ms

3.4. Simulation Setup

In order to evaluate bandwidth requirements in the transport network of UTRAN, simulation models were implemented in OPNET simulator [58], a commercial network simulator. Relying upon the IP-related networking simulation components provided by OPNET, the new implemented components are the voice traffic model and web-browsing traffic model; and the layered structure of the Iub interface with IP transport.

3.4.1. Traffic Models

The simulation models used for the voice traffic and web-browsing traffic are detailed in this section. The traffic models considered in our study have been defined following the recommendations considered in previous studies [11], and also in 3GPP technical reports [16]. The voice model consists of a series of ON and OFF periods with a service rate of 12.2 Kbps, which corresponds to one of the bit rates achieved by the adaptive multi-rate (AMR) codec specified by 3GPP. The ON and OFF states are exponentially distributed with a mean duration of 3 sec [11]. We assume that all users' sessions are kept active during the entire simulation elapsed time. The background noise description packets sent by the codec during the silence periods are not considered in the voice traffic model.

In the case of data traffic we consider a web-browsing traffic model. A web session is modeled as a sequence of packets corresponding to the download of pages. The generation of packets is modeled using a truncated Pareto distribution, where the mean packet size, denoted as μ_m, generated by the model can be estimated as [19]:

$$\mu_m = \frac{1}{1-\alpha}\left(\frac{k^\alpha}{m^{\alpha-1}} - \alpha k\right) \tag{3.2}$$

where α=1.1 is the shape parameter of the Pareto distribution, k=81.5 is the minimum packet size (in bytes), and m=6000 is the maximum packet size (in bytes). This results in a mean packet size of 366 bytes. The number of pages in a session is a geometrically distributed random variable with a mean of 5 pages. The number of packets per downloaded page is modeled by a geometrically distributed random variable with a mean of 25 packets. Packet-calls are separated by an interval (reading time) which is a geometrically distributed random variable with a mean of 10 seconds. Table 3.2 summarizes the considered voice and web-browsing traffic model parameters.

3.4.2. Iub Simulation Model

In accordance to guidelines provided in [16], we consider a modular structure for the Iub interface modeling. The structure is separated into the following different modules: link, IP transport, Radio Protocols/FP, and traffic sources. Figure 3.5 shows a diagram of the IP-based Iub reference model used for DCHs and HS channels. The link and IP transport modules are common for both scenarios, and the main differences between the two depicted structures are related to the Radio Protocols/FP functions and traffic model assumptions. A description of the different components is given in the following paragraphs.

Table 3.2: Parameters of voice and web-browsing traffic models.

Traffic model	Parameter	Value
Voice	Source type	ON/OFF
	Call duration (s)	30
	Inter-call time (s)	10
	Packet size (bytes)	32
	Packet inter-arrival time (ms)	20
	ON period duration (s)	3
	OFF period duration (s)	3
Web-browsing	Packet inter-arrival time (ms)	11.9
	Number of packets per page	25
	Number of pages per session	5
	Reading time (s)	30
	Shape parameter, α	1.1
	Minimum packet size (bytes), k	81.5
	Maximum packet size (bytes), m	6000

In the modeling of the link, we consider a single queue onto which all concurrent connections are multiplexed, and where the service time is a linear function of the IP packet size. Furthermore, the buffers in the link model are assumed to be large enough to accommodate potential overloads, and hence there is no packet loss.

The IP transport module includes the following components: segmentation, multiplexing queues and packetizer. The segmentation module guarantees that large FP frames are fragmented in order to fit into the maximum container payload. The multiplexing queue retains FP frames from various streams (i.e., user connections), so that the packetizer can arrange several of them into the same IP packet. This process introduces an additional delay to the streams (e.g., FP frames wait in the multiplexing buffer until either there is enough data to build a complete transport packet or when the maximum allowed waiting time in the packetizer has been reached). Moreover, the following overheads are considered in the IP transport module:

- Overhead/Stream, added to each FP-PDU so that several of them can be packet into the same container.
- Overhead/Container, added to the set of FP frames multiplexed in a single IP packet.
- The UDP/IP overhead of the packet to be delivered to the transport.

We have adopted a generic Iub interface modeling, where the FP frames of different user flows are concatenated into a single IP packet in multiplexing queues of the IP transport module. This model can be used to capture the behavior of different multiplexing methods by simply considering the corresponding overheads added to each stream and also the overhead/container [11].

The modeling of the Radio Protocols/FP block is addressed taking into account two main aspects: overheads and queuing. On one hand, overheads values are derived from headers added in the packet data convergence protocol (PDCP), RLC, MAC and FP layers, which depend on the service type and on the considered scenario. In particular, in the scenario of DCHs, voice traffic is assumed to be supported under the transparent RLC mode whereas web traffic uses RLC acknowledge mode. The use of the PDCP layer is considered for header compression in the case of data services. In the scenario of HS channels, voice traffic is seen as voice over IP (VoIP) so that PDCP is also used in this case for header compression. We assume that after compression the resulting header size is approximately 4 bytes [59]. On the other hand, RLC/MAC queuing is introduced to account for the effect of having a maximum bit rate for the DCHs scenario. Otherwise stated, we have considered DCH channel rates of 256 Kbps for web traffic, and channel rates of 12.8 Kbps for voice traffic (i.e., a packet of 32 bytes is sent every 20 ms).

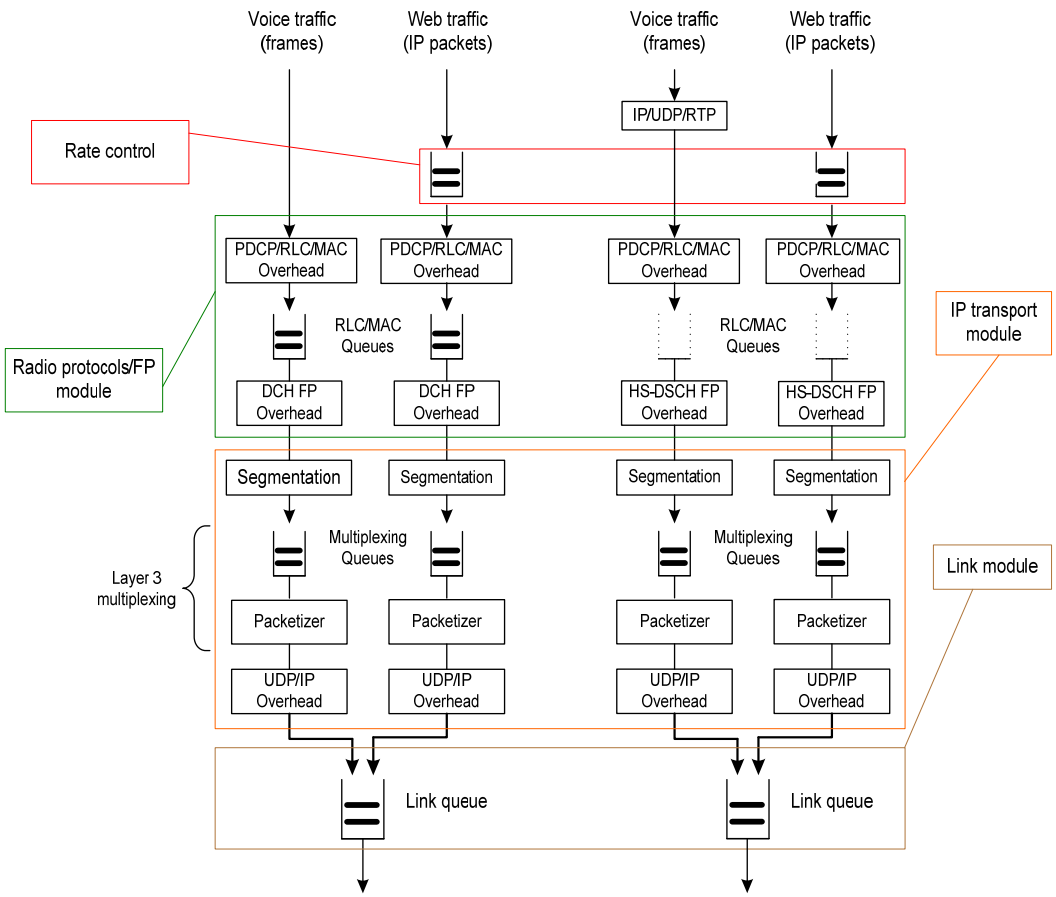


Figure 3.5: Iub modeling for DCH channels (left side) and HS channels (right side).

In the DCH scenario, the maximum bit rate should be enforced by the MAC scheduler at the RNC. Notice that in the scenario of HS channels, we do not model this effect since the assumption here is that data arriving at the RNC can be directly forwarded to the Node B scheduler, that is, no RLC/MAC buffer waiting time is considered in the RNC for HS channels scenario.

Finally, we have also captured in the model the effect of a rate limiter for web traffic. The purpose of a rate control is to limit the maximum data rate. This avoids large traffic bursts reaching the RLC/MAC buffers at the RNC. In both scenarios, the maximum data rate allowed by the rate control is limited to 512 Kbps. Although this rate control is depicted in Figure 3.5 as an extra queue, it is simulated in our model by adjusting the web-browsing traffic model in order to generate traffic at a given data rate. In this sense, in order to achieve a maximum data rate of 512 Kbps, the inter-arrival packet time of web-browsing traffic has been adjusted to 5.95 ms. Notice that in a real system, this rate limiter could be located at the gateway of the packet-switched core network. Table 3.3 and Table 3.4 list the overheads considered for DCHs and HS channels, respectively [11], [19].

Table 3.3: Voice and web-browsing traffic overheads for DCHs scenario.

Module	**Component**	**Overheads**	
		Voice	**Web**
Radio protocols/FP	PDCP/RLC/MAC	0 bytes	2 bytes
	DCH-FP overhead	8 bytes	5 bytes
IP transport	Overhead/Stream	3 bytes	3 bytes
	Overhead/Container	8 bytes	8 bytes
	UDP/IP Overhead	28 bytes	28 bytes

Table 3.4: Voice and web-browsing traffic overheads for HS channels scenario.

Module	Component	Overheads	
		Voice	Web
Traffic source	IP/UDP/RTP	4 bytes	—
Radio protocols/FP	PDCP/RLC/MAC	2 bytes	2 bytes
	HS-DSCH FP overhead	10 bytes	10 bytes
IP transport	Overhead/ Stream	3 bytes	3 bytes
	Overhead/Container	8 bytes	8 bytes
	UDP/IP Overhead	28 bytes	28 bytes

3.5. Numerical Results

This section presents simulation results of transport capacity requirements in an IP-based UTRAN. Simulation results are presented in terms of the extra capacity required, given by means of the β factor introduced in equation (3.1), that is required to meet delay requirements in the transport network. This is performed under different levels of mean traffic load (aggregated from a number of voice or web-browsing traffic sources). Specifically, for a given simulation run we consider a specific mean service traffic load of either voice or web-browsing service type. The number of concurrent connections is adjusted before each simulation case so that the required mean amount of traffic entering the network could be achieved. Then, we collect the packet delay statistics of a period of 5 minutes, excluding a warming time that is required to assure a stabilized mean traffic load.

In this section, the capacity requirements of the evaluated scenarios are firstly presented. This is done for different mean traffic loads and considering the delay requirements detailed in section 3.3. In the second part of this section, a sensitivity analysis is performed in order to study how different aspects considered in the transport network modeling and traffic models impact on the capacity requirements study.

3.5.1. Capacity Requirements

The simulation results for DCHs under voice and web-browsing traffic are presented in Figure 3.6 and Figure 3.7, respectively. These figures present the extra capacity (in terms of the over-provisioning factor, β) respect to the mean traffic that is required to meet the considered delay requirement. The mean rate values of aggregated traffic considered in the analysis are: 2 Mbps, 4 Mbps, 8 Mbps, and 16 Mbps. The legends in the figures relate the colors of the bar graphs to the considered delay bounds expressed in milliseconds (ms).

It can be seen that β values ranging from 50% to 70% suffice to support the different mean loads of voice traffic in order to fulfill the considered delay requirements. For both services the degree of over-provisioning decreases as the mean aggregated traffic in the network is increased, which means that the traffic peak rates are less pronounced. Notice that in the case of voice traffic (see Figure 3.6) small differences are observed when comparing the lowest and highest mean traffic loads. For instance, for the mean load of 2 Mbps and 16 Mbps, it is required an over-provisioning factor of around 69% and 52%, respectively, in order to meet a delay requirement of 5 ms. This translates to a total transport capacity of around 3.4 Mbps and 24 Mbps, respectively.

Unlike voice traffic, simulation results for the case of web-browsing traffic shows that the over-provisioning factor exhibits more drastic changes as the mean traffic load in the transport network is increased. Notice that while over-provisioning factors of around 70% are required to support mean voice traffic of 2 Mbps, and meet a delay requirement of 5 ms, in the case of web-traffic the extra capacity needed is around 160% in order to support the same level of mean traffic. These

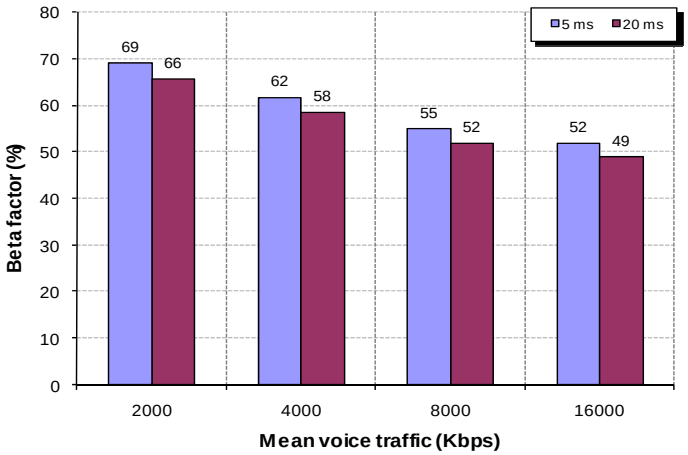


Figure 3.6: Over-provisioning factor (β) for DCHs with voice traffic.

differences are due to the nature of each service (i.e., the web-browsing service generates a more bursty traffic pattern than voice service). In addition, the web-browsing traffic model considered in our study also accounts for a given level of *self-similarity* [60], due that the traffic model generates packets according to a heavy-tailed Pareto model. Hence, the web-browsing traffic is generated with higher temporal correlations than the classic Poisson, or exponential, traffic models.

Figure 3.9 and Figure 3.8 show the capacity requirements of the HS channels scenario when voice and web-browsing traffic, respectively, is assumed. Notice that a significant increase in the over-provisioning factor, with respect to DCHs scenario, for both voice and data services is appreciated. If we compare the DCHs and HS channels scenarios for the case of voice traffic, Figure 3.6 and Figure 3.9, respectively, it is observed that the over-provisioning factor increases around 20% when introducing HS channels. This increase is mainly due to the larger overhead values incurred in the protocol stack of this scenario in order to support voice over IP (VoIP). It is worth noting that the mean bit rate values indicated in the abscissas axis of the graphs only account for voice frames.

On the other hand, in the case of web-browsing traffic, quite different situations can be envisaged. Focusing on Figure 3.7 and Figure 3.8, we can observe that, when comparing the same delay constraint, transport capacity requirements in HS channels is higher. This is due to the high

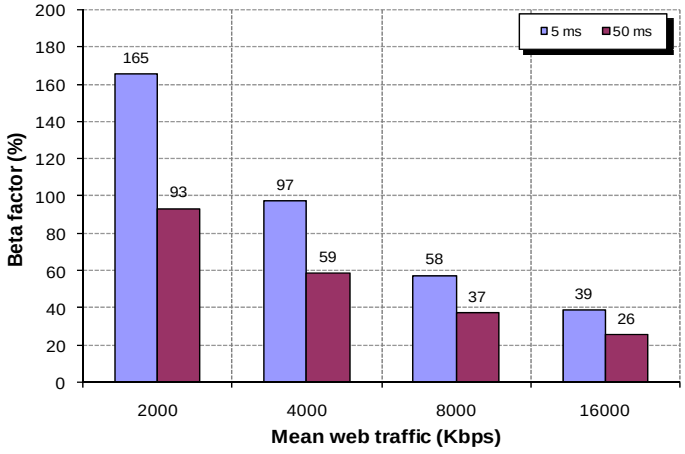


Figure 3.7. Over-provisioning factor (β) for DCHs with web traffic.

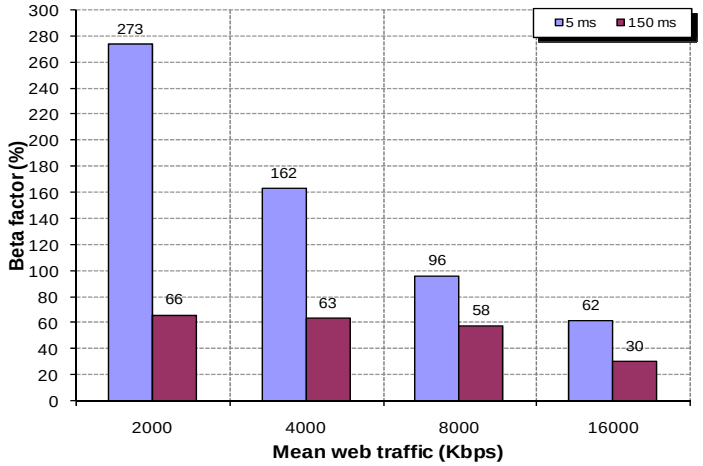


Figure 3.8: Over-provisioning factor (β) for HS channels with web traffic.

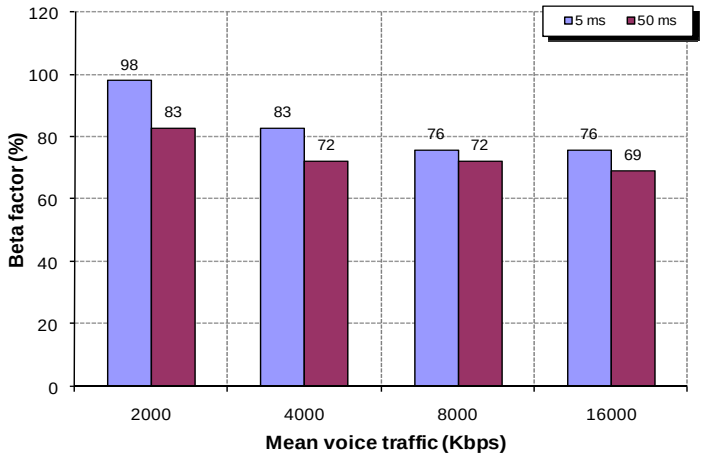


Figure 3.9: Over-provisioning factor (β) for HS channels with voice traffic.

variability of the traffic injected by the RNC into the transport network because, unlike the scenario with DCHs, there is no smoothing effect introduced by the RLC/MAC queuing. Comparing these figures it is also observed that transport capacity requirements in the HS channels scenario are lower than in the case of DCHs scenario, when the considered delay bounds are higher. Although this could seem an unfair comparison, notice that the delay bound in DCHs is not due to the service itself but to specific radio functions within the FP-DCH protocol. Contrarily, this limitation does not exist in the scenario of HS channels so that softer delay requirements can be applied whenever the final service is not deteriorated.

3.5.2. Sensitivity Analysis

In this section we extend the previous simulation results. The aim is to assess the sensitivity of the estimated capacity to a number of selected factors such as the network size, traffic model parameters and DCH rates. Prior to the sensitivity analysis we study the contribution to the over-provisioning factor due to the protocol overhead in the Iub interface so that the part of the over-provisioning factor that exclusively dependent on traffic statistics is clearly identified.

3.5.2.1. Protocol Stack Overhead Analysis

So far, it has been analyzed the transport capacity requirements in the IP-based transport network under different cases (i.e., service type and transport channel), considering specific delay requirements for the transport network. To this end, a detailed characterization of a generic Iub protocol stack has been addressed in section 3.4.2. Now in this section, we analyze the amount of overheads introduced by the considered protocol stack for each of the studied cases. The overhead values are derived from the corresponding headers that are added in each of the different modules of the Iub interface model depicted in Figure 3.5, as detailed in the following paragraphs.

In the Iub interface model, the data traffic in the traffic source module, i.e., payload plus IP/UDP/RTP overheads, is received at the RLC layer and fragmented into RLC Packet Data Unit (PDU) of 40 bytes, which in turns pass them to the MAC layer. Then, the MAC layer adds 2 bytes to each of the fragmented packets and sends data to the FP layer, which in turn adds the FP headers to conform the FP PDU streams. Next, the overhead/stream is added to each FP PDU stream so that several of them can be packet into the same container.

The assembled FP PDUs are then received in the multiplexing queues and then send to the packetizer where the size of the introduced container/overhead is 28 bytes to each IP packet. Finally, an UDP/IP overhead of 8 bytes is added to the IP packet. This leads to an overhead of 36 bytes for each packet to be transmitted over the IP transport network. In addition, the multiplexing queues also incorporate the following parameters: maximum waiting time, minimum container size, and maximum container size. Table 3.5 contains the values of the multiplexing queues that we have considered in simulations. Therefore, the multiplexer conform a ready-to-transmit packet by multiplexing the received FP PDU streams, and adding the corresponding overheads.

The payload within each FP PDU stream depends on the type of services it carries. In the case of web-browsing traffic, the payload within each FP PDU is 40 bytes (i.e., due that the packets generated by the web-browsing traffic model are segmented in the RLC layer). On the other hand, voice traffic model generates voice frames with a constant size of 32 bytes. It is worth noting that in the scenario of HS channels the voice service is seen as voice over IP (VoIP) an overhead for the IP/UDP/RTP headers has to be added. This amount of overhead is limited to 4 bytes since it is assumed that header compression is applied (i.e., PDCP layer).

In this context, the generated IP packet to be transferred over the transport network is depicted in Figure 3.10. The size of the generated IP packet is denoted as IP_{packet}. This IP packet contains n FP_PDU streams plus the overhead added to each stream, and also the overheads added to the set set of FP_PDUs multiplexed in a single IP packet. In particular, the size of the overheads of the IP packet, denoted as $IP_{overheads}$, is equal to the value of the overhead/container added in the packetizer plus the IP/UDP overhead. The generated IP packet size could be expressed as follows:

$$IP_{packet} = IP_{overheads} + (n)(FP\_PDU + overhead/stream) \quad (3.3)$$

In order to analyze the amount of overhead introduced by each protocol stack in the Iub interface for each of the two considered scenarios, it is assumed that when the maximum waiting time in the multiplexer has been reached there are enough FP PDU streams stored in the queues so that the total size of the packet to be generated can be as greater as the maximum container size we have considered in the multiplexing queue (see Table 3.5).

Table 3.5: Parameters of the multiplexing queues.

Parameter	Value
Maximum waiting time	5 ms
Minimum container size	1200 bytes
Maximum container size	1500 bytes

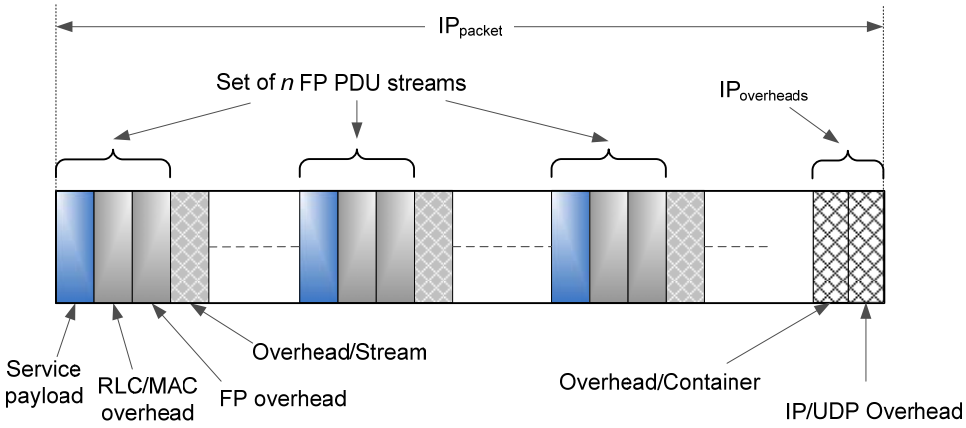


Figure 3.10: Generated IP packet format.

Therefore, the exact number of FP PDU streams can be estimated assuming that the size of the assembled packet in the multiplexer is IP_{packet}=1500 bytes. This can be expressed as:

$$n = \left\lfloor \frac{IP_{packet} - IP_{overheads}}{FP\_PDU + overhead/stream} \right\rfloor \tag{3.4}$$

With the number n of FP PDU streams, the exact total packet size generated can be calculated using equation (3.3). Over such a basis, it is possible to determine the portion of the traffic that correspond solely to the amount of overheads introduced in the Iub interface, and can be expressed as:

$$\text{Overhead}(\%) = \frac{IP_{packet} - (n)(S_{payload})}{n \cdot S_{payload}} \times 100 \tag{3.5}$$

where $S_{payload}$ is the service payload with a value of 32 bytes and 40 bytes, for voice traffic and web-browsing traffic, respectively. Using the detailed procedure, the overhead percentage for each scenario can be computed. Resulting values for the two considered scenarios are summarized in Table 3.6.

In the following we now consider simulation results presented in section 3.5.1 in order to evaluate the contribution to the over-provisioning factor due to the protocol stack overheads. The bars in Figure 3.11 and Figure 3.12 show the over-provisioning factor for the two considered scenarios under voice and web-browsing traffic, respectively. The third series in these figures show the overhead introduced by each Iub protocol stack for each case. The difference between the over-provisioning factor and the overhead value correspond to the extra capacity needed to overcome traffic fluctuations. Focusing in the case of voice traffic (see Figure 3.11), the overhead percentage is close to 40 % for DCH channels and rises up to around 65% for HS channels. This overhead increase is due to the additional overheads resulting from the support of the VoIP protocol stack in the case of HS channels. On the other hand, the overhead, for web traffic, as shown in Figure 3.12, overhead due to protocol stack is found to be around 19.6% for DCHs and 7.3% for HS channels (where the maximum allowed MAC-d PDU is 625 bytes). This implies that under web-browsing traffic the usage of DCH channels or HS channels have less impact on transport capacity requirements than in the case of voice traffic. Therefore, transport capacity requirements in these cases are mainly influenced by the dynamics of the traffic itself.

Table 3.6: Overhead percentages.

Scenario	Voice	Web
DCH channels	37.6 %	19.6 %
HS channels	63.3 %	7.3 %

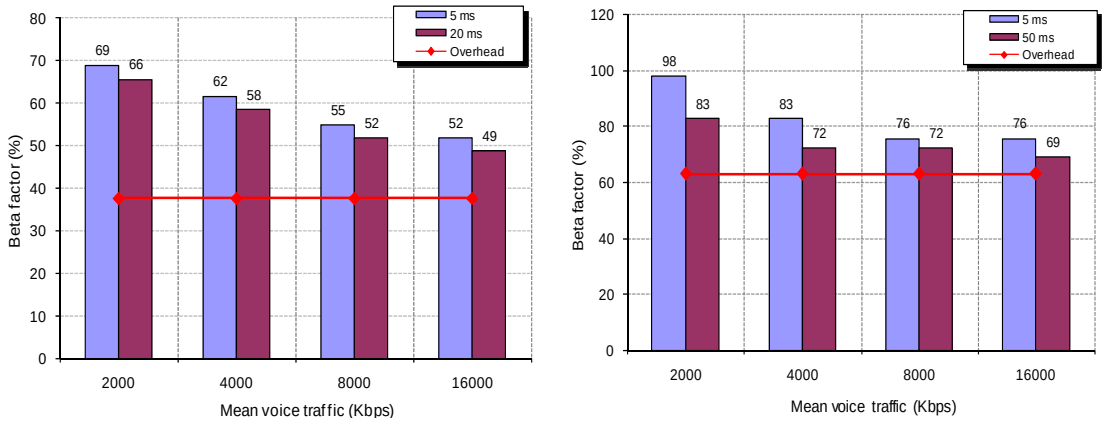


Figure 3.11: Impact of protocol overheads on capacity requirements for voice traffic with: (a) DCHs, (b) HS channels.

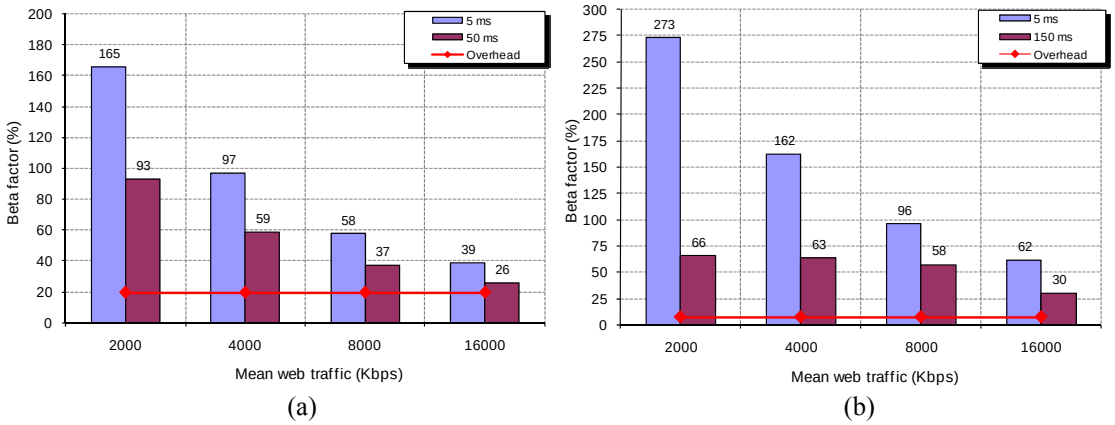


Figure 3.12: Impact of protocol overheads on capacity requirements for web traffic with: (a) DCHs, (b) HS channels.

It is worth noting that the overhead percentages incurred by the protocol stacks are dependent on the considered multiplexing method used to concatenate multiple streams into a single IP packet. In any case, the generic approach followed to model the Iub interface can be used to capture the behavior of different multiplexing approaches, such as the ones listed in the following.

- CIP. The composite IP is a layer 3 multiplexing scheme, where FP PDU streams are assembled to fit the CIP packet payload and form a CIP container [16]. A segmentation/re-assembly mechanism allows to split large FP PDU streams into small segments. Several CIP containers are multiplexed into one IP packet. The proposed protocol stack is shown in Figure 3.13(a).

- Lightweight IP Encapsulation (LIPE). This is a layer 3 scheme, see Figure 3.13(b), to multiplex low bit rate audio (or multimedia) packets into a single UDP/IP session [61], [62].

- AAL2/UDP/IP is other layer 3 multiplexing scheme where multiplexing of FP frames into IP packets is carried out above the IP layer by AAL2 layer [51]. The protocol stack of this alternative is illustrated in see Figure 3.13(c).

- PPP multiplexing (PPPmux) is a layer 2 scheme (see Figure 3.13) that uses the point-to-point-protocol (PPP). As illustrated in Figure 3.13(c) one FP frame from the RNL is encapsulated into one IP packet with possible UDP/IP header compression (cUDP/IP). The key idea is to

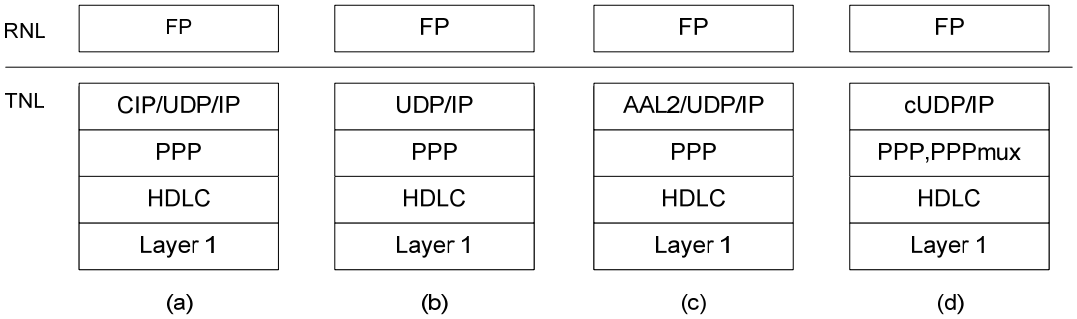


Figure 3.13: Protocol stacks for IP transport optimization.

concatenate multiple PPP encapsulated frames containing IP packets into a single PPP multiplexing scheme [63].

3.5.2.2. Sensitivity to DCH rates

After observing different cases of capacity requirements under different mean traffic loads for each scenario, in this section we investigate the impact on capacity requirements due to the DCH rate focusing on the web-browsing traffic. We consider a simulation setup where the aggregated mean traffic load is 2 Mbps, and considering three different DCH rates: 128 Kbps, 256 Kbps and 384 Kbps. Furthermore, three different delay constraints are also considered for each case. The results of this study are shown in Figure 3.14.

From a resource consumption point of view, the worst combination (in terms of transport capacity requirements) is when the speed of DCHs is set to 384 Kbps and hard delay requirements, such as 5 ms, should be satisfied in the IP-based transport network. The use of high channel rates increase the levels of burstiness of the traffic being supported over the transport network, thus leading to higher capacity requirements for the transport network in order to meet the considered delay requirements. On the other hand, DCHs with lower data rate serve as a kind of traffic shaping of the traffic entering the RLC/MAC queues, and thus allowing a more relaxed IP transport. Notice that about a channel rate of 384 Kbps requires around 25% of more capacity than in the case of a channel rate of 256 Kbps in order to meet the stringent delay requirement of 5 ms.

3.5.2.3. Sensitivity to traffic models

It has been inferred that the inherent characteristics of traffic substantially determines transport capacity requirements in the IP-based access network. The dependence of the results to the burstiness level of the traffic model is evaluated in this section. To this end, other types of services that exhibit different traffic pattern are implemented. The selected services for this study are streaming and background. These services are simulated using the web-browsing traffic Pareto model by adjusting the values of the following parameters:

- For streaming data, a still image service is considered where data packets of mean size 60 Kbytes are send each 2 seconds. The minimum data packet size is 22 Kbytes, while the maximum is set to 147 Kbytes. The shape parameter of the Pareto distribution is equal to 1.1.
- For background data, a fax service is considered with the following parameters: inter-arrival time of 10 seconds, the mean data packet size is 200 Kbytes, the minimum packet size is 56 Kbytes, and maximum packet size is 1.1 Mbytes. The value for shape parameter of the Pareto distribution is equal to 1.1.

Simulation results for this case are presented in Figure 3.15. This figure depicts the mean traffic that can be supported over the Iub interface in order to meet a given delay requirement. We use as a reference case the transport capacity requirements for the web-browsing service when considering

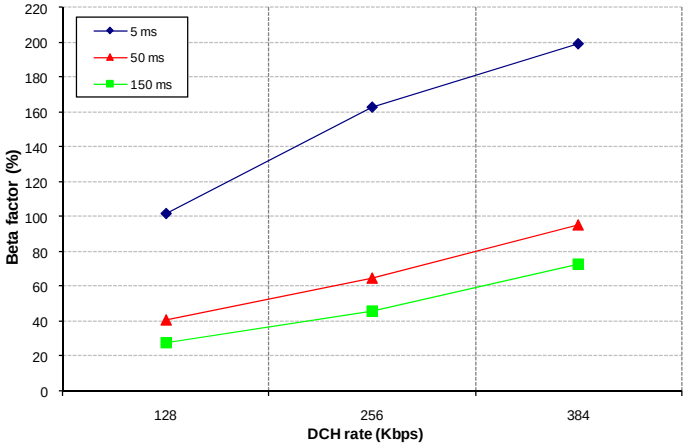


Figure 3.14: Over-provisioning factor (β) under different channel rates and delay requirements.

DCHs or HS channels, with a delay requirement of 5 ms. More specifically, the transport capacities considered in simulations are set to the one required, in the case of web-browsing service, to support a mean web-browsing traffic of 2 Mbps. In this sense, a transport capacity of 5.3 Mbps is used for the case of DCH channels, and a transport capacity in the Iub of 7.45 Mbps for the case of HS channels.

As seen in Figure 3.15, lower mean traffic values of web-browsing traffic can be supported for a given transport capacity in the Iub interface than in the case of background and streaming traffic. For instance, in the case of DCHs, it is seen that while the mean web-browsing traffic supported in the transport network (with 5.3 Mbps of bandwidth) is 2 Mbps, mean traffic values of about 3 Mbps and 4 Mbps can be supported in the case of background and streaming traffic, respectively. This translates to an over-provisioning factors of around 165%, 76%, and 32% for web-browsing, background and streaming traffic, respectively. This again shows that web-browsing traffic exhibit more pronounced levels of burstiness than the two other services, and thus more transport capacity in needed to meet the delay requirement of packets in the transport network.

3.5.2.4. Sensitivity to the diameter of the network

The impact of the number of hops in the IP-RAN on delay requirements is analyzed in this section. The traffic supported in the IP-based transport network passes through a specific number

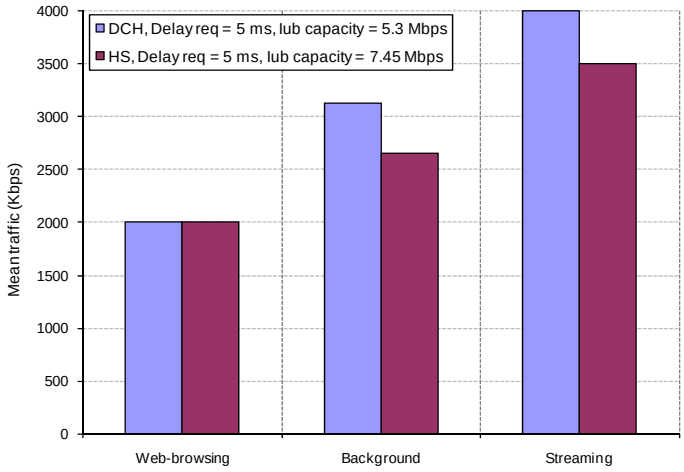


Figure 3.15: Mean traffic supported in the transport network for different services assuming the transport capacity required to meet the 99.9% of the delay requirement.

of IP routers in the path between RNC and Node B. This network path must accommodate enough capacity at each hop in order to fulfill end-to-end (ETE) delay requirements. The number of hops in a network path is normally referred to as network diameter.

The total delay requirement to be satisfied by the transport network has to be distributed along the path components that incur in some type of delay. The delay components can be decomposed into "per-hop" delays, and these in turn into "per-link" and "per-node" delay components. The per-link delay components imply the propagation delay, whereas per-node delay components involve three sub-components: serialization delay, processing delay, and queuing delay.

Over such a basis, the bandwidth that must be provisioned in each forwarding node in order to statistically guarantee, with a given probability, a maximum tolerable ETE delay in the transport network can be approximated if the mean delay per-node is known. In order to determine the mean value of the delay experienced by traffic in a given node, it is necessary first to assemble the delay budget of an IP network path taking into account the aforementioned delay components. To this end, we assume that a given path in the IP-based transport network is composed by N identical routers. In this context, the delay budget, denoted to as D_{budget}, of the transport network can be expressed as follows:

$$D_{\text{budget}} = D_{\text{prop}} + \sum_{i=0}^{N} D_{n,i} \tag{3.6}$$

where D_{prop} is the propagation delay in the entire network path, and $D_{n,i}$ is the delay incurred at each router i which in turn can be computed as follows:

$$D_{n,i} = D_{\text{ser},i} + D_{\text{proc},i} + D_{\text{queue},i} \tag{3.7}$$

where $D_{\text{ser},i}$ is the serialization delay, $D_{\text{prop},i}$ is the processing delay, and $D_{\text{queue},i}$ is the queuing delay. The first two components of equation (3.7) as well as propagation delay in equation (3.6) are likely to be the deterministic part of the delay budget and they are relatively easy to determine. On the other hand, queuing delay represents the stochastic part of the delay and it is more difficult to predict because it depends on the congestion level experienced in each network node. In this sense, we consider a number of simplified conditions to analytically approximate the delay experienced in a network path.

We assume that the total delay observed in a given network path exclusively depend on the waiting times in the queues, which is assumed to be exponentially distributed and independent. Then, the queuing delay through N identical nodes can be estimated using the closed-form formula presented in [64]. This formula express the (1-ε)-quantile of the total end-to-end delay requirement for a Poisson traffic traversing N identical nodes as the sum of the average total queuing time and the number of times the standard deviation of the total queuing time. This can be written as [64]:

$$D_{\text{ETE}} = \mu_N + \alpha_N(\varepsilon)\sigma_N \tag{3.8}$$

The values of α_N solely depends on the diameter of the network (i.e., the number of N hops) and the value of ε, that is defined as the portion of the traffic that does not meet the delay requirement. The standard deviation, referred to as σ_N, is determined by:

$$\sigma_N = \sqrt{N} \times D_{\text{mean}} \tag{3.9}$$

where D_{mean} is the mean delay in a network node. With respect to the average waiting time, denoted as μ_N, in equation (3.8), this can be expressed as:

$$\mu_N = N \times D_{\text{mean}} \tag{3.10}$$

Substituting equations (3.9) and (3.10) into equation (3.8), and assuming a maximum tolerable value of ETE delay, the mean delay, denoted as D_{mean}, in a hop can be obtained. Then, from the mean delay value and attending to the assumption that the delay is exponentially distributed, the (1-ε)-quantile of the delay incurred in a single hop can be calculated.

Figure 3.16 presents the delay incurred at each hop when considering four different end-to-end requirements, and different network sizes. Particularly, varying the number of nodes between 2 and 16 we compute the part of the ETE delay that should be satisfied in one hop of a certain path.

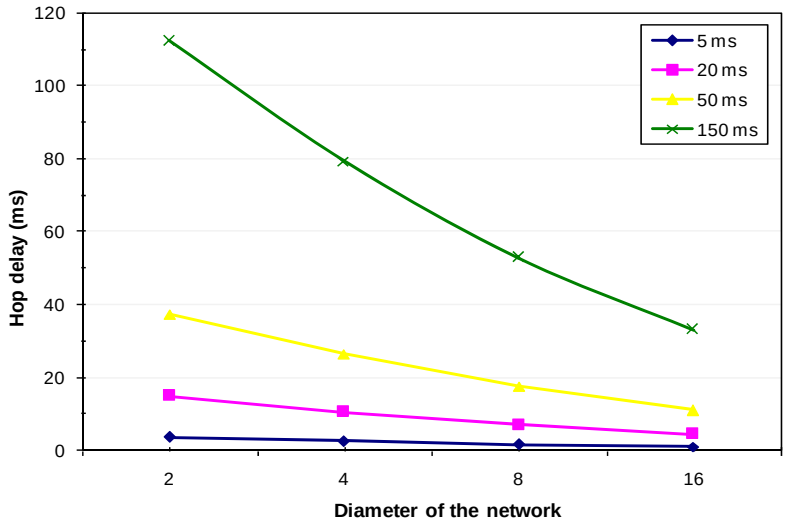


Figure 3.16: Hop-delay for different end-to-end delay requirements and network sizes.

Moreover, different delay values have been considered as ETE requirement. It can be seen that while the number of hops increase, the per-hope timing requirement become more stringent.

3.6. Summary

This chapter has provided an analysis of transport capacity requirements in the access network when introducing IP as a transport technology. It also provides detailed characterization of the Iub interface. Capacity requirements are estimated assuming an over-provisioning approach and considering two different scenarios. In the first scenario, traffic is mainly supported by means of DCH channels in the radio interface, while in the second one the traffic is supported over HSDPA channels. For each scenario, the protocol stack and delay requirements imposed by the radio layer have been specified. A simulation model for bandwidth estimation has been defined to assess the required capacity, expressed in terms of a parameter known as the over-provisioning factor that relates the extra-capacity required in a link respect to the aggregated mean bit rate supported on that link.

The capacity required in the IP-based transport network has been shown to depend on the aggregated mean bit rate, but particularly on the characteristics of the individual traffic sources and on the considered delay bounds. It has been shown that while over-provisioning factors around 40-60% can suffice for voice services in the DCH scenario, web-browsing services can demand values higher than 100% just to satisfy the delay requirements imposed by the Iub mechanisms and not by the traffic itself. In the scenario with HS channels, it is shown that higher over-provisioning factors are required for voice (mainly due to the higher overhead of VoIP solutions). Contrarily, lower over-provisioning factors can be applied in the HSDPA scenario (e.g., around 75% for 2 Mbps aggregates) to support web-browsing traffic because of the lower delay restrictions that in this case can be assumed in the Iub interface since neither outer loop power control or soft handover must be supported.

4 Coordinated Access Resource Management Framework

4.1. Introduction

As discussed in Chapter 2, the current trend to all-IP architectures in mobile communications creates the challenge to support quality of service (QoS) over IP networks. At the same time, the introduction of new high-speed and data-intensive mobile services claims for a highly efficient use of both radio and backhaul network resources. Taking this into account, in this thesis we face a situation where limiting resources (bottlenecks) do not have to be necessarily on the air interface of the access network, but can be in the backhaul (transport) capacity of the mobile access network. Therefore, QoS and efficient network resource management guarantees would depend on both wireless (radio interface) and wired (backhaul) segments of mobile access networks.

The possibility that transport network resources could put important limits to QoS has been little explored in the literature so far. In this chapter we propose a novel resource management framework that considers both radio and transport resource occupancy in its decision-making process. In this context, in Section 4.2 we firstly present a QoS reference architecture where QoS management functions specific to the transport network (hereafter referred to as transport resource management, TRM) as well as radio resource management (RRM) QoS functions are identified. The proposed framework, referred to as Coordinated Access Resource Management (CARM), is also introduced in this section, along with the identification of different functions that can be developed within the context of the CARM approach. Over such a basis, Section 4.3 presents a generic framework to analyze the benefits of a coordinated mobility control (cell selection) strategy in mobile networks with transport capacity limitations, regardless of the studied radio access technology (RAT). Specifically, in this section an analytical model based on multi-dimensional Markov chains is developed to assess the performance of different cell selection strategies that may consider both radio and transport constraints. Finally, the main conclusions of this chapter are summarized in Section 4.4.

4.2. Resource Management Functional Model

The proposed CARM framework has been developed in the context of the IST AROMA Project [46] in order to fully exploit access network resources and handle potential bottleneck situations in the mobile backhaul network. This approach leads to a new paradigm where backhaul resources are considered not only at the network dimensioning stage but are included in an integrated resource management scheme. In this section, the CARM functional framework is introduced as a feasible approach to take into account backhaul resource usage information within different resource management functions along with radio resource considerations.

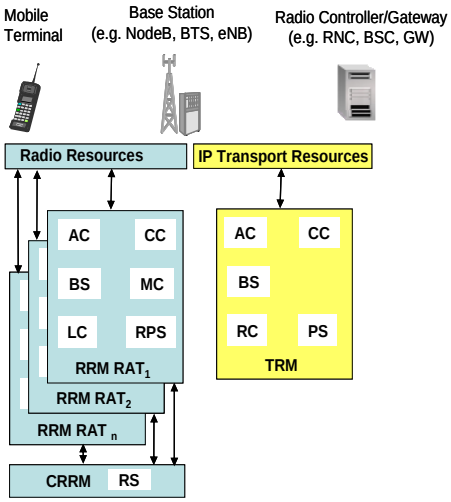


Figure 4.1: Separate QoS functionalities for RRM and TRM [46].

4.2.1. Reference QoS Framework

The UMTS end-to-end QoS architecture is specified in [65] and [66]. The QoS architecture is based on the concept of *bearer service* that specifies a transport service between two points within the network along with its expected QoS behavior. The bearer model follows a layered approach where a bearer service between two points in the network is made up of the concatenation of underlying bearer services between intermediate points (if any). The proposed CARM model is built upon this reference bearer service architecture and focus on the resource management functions needed to manage the configuration and operation of the bearer services between terminals and base stations (radio segment) and base stations and network nodes in the access (e.g., radio network controllers, base station controllers for UTRAN and GERAN respectively) or core network (e.g., serving gateway for LTE).

Figure 4.1 highlights the different pools of resources to be managed by QoS supporting functionalities and show the key QoS functionalities needed for both RRM and TRM. In principle these functions could be conceived separately for both QoS management domains:

- With respect to RRM, the proposed functions take into account, as a guideline, the functions proposed in [65], [66], and were already studied and validated in the context of the European IST EVEREST Project [67]. Hence, considering a multi-RAT scenario, RRM functions comprise: admission control (AC), congestion control (CC), bearer selection (BS), mobility control (MC), radio link control (LC), radio packet scheduling (RPS) and RAT selection (RS), this latter being considered as a common RRM (CRRM) when managing multiple RATs.

- With respect to TRM, we have identified the proposed functions by taking into account recent research efforts towards the introduction of QoS management in the context of DiffServ IP networks, [68], [69], [70] and by realizing that the desirable QoS resource management objectives in the transport network layer (TNL) claim for a set of functions that mirror, to certain extent, those already familiar at the radio network layer (RNL), as explained in Section 2.2.3.1. Hence, key TRM functions are admission control (AC), congestion control (CC), bearer selection (BS), packet scheduling (PS) and route control (RC).

4.2.2. Envisaged CARM functionalities

When trying to jointly optimize the use of both radio and transport resources by means of RRM and TRM functions identified in

Figure 4.1, a strong interaction can be predicted among several functions of both domains. This idea motivates the development of a coordinated resource management framework encompassing both radio and transport segments to implement some specific resource management functionalities. These functions are referred to as CARM functions. In particular, it is proposed CARM functions to cover: admission control, congestion control, bearer selection, mobility control and RAT selection. Notice that decisions taken by these functions have an impact in the utilization of radio and transport resources, and thus an interaction between radio and transport network could help to take more accurate decisions in order to use resources of both segments more efficiently. For instance, with respect to the mobility control function, it is foreseen that the transport network will also have to be informed and checked for available resources before making the decision of a radio layer handover, and so this function is also included among the CARM functions. Furthermore, the RAT selection, which originated as a CRRM function, should also be included as a CARM function due to the possible dependence of RAT selection decisions on the TNL link loads. On the other hand, the link control on the RRM side, as well as route control on the TRM side are considered specialized functions that only make sense within their own scope (or pool of resources). The same can be said about packet scheduling in both RRM and TRM. It is in the packet scheduling function where some important QoS decisions (like priority) are separately enforced in the relevant nodes from RNL or TNL.

Figure 4.2 summarizes the name and scope of all the proposed QoS management functions, while Figure 4.3 shows the role of the coordinated QoS functions in the medium-term and long-term architectures. By medium-term architecture we refer to scenarios where the transport technology within the access network is entirely based on IP (i.e., 3GPP release 6), and where the Iub interface must still be fully supported over IP. On the other hand, by long-term architecture we mean a scenario aligned to 3GPP long term evolution (LTE) efforts [1], where the RNL consists only of two types of nodes, the access gateway (aGW) and the Evolved UTRAN Node B (eNB), and where the Iub interface with its stringent delay constraints, is no longer needed.

4.2.2.1. CARM Functions

The objectives of each of the envisaged functions in the CARM functional framework are given in the following.

- RAT selection. The RAT selection function is in charge of selecting the most appropriate RAT, either at call establishment or during the session life-time through the so-called vertical handover procedure, given the requested service and QoS profile. The RAT selection decision could be influenced by many factors, including non-technical issues like the operator's policies or business model. It is considered a CARM function due to the need to take into account the link load at the transport layer before making a RAT selection.

- Bearer selection. The bearer selection is in charge of selecting the required resources to support the requested QoS profile at the radio and transport bearer services. This implies the configuration of new radio and transport bearers given the requested QoS profile and selected RAT. It also includes dynamic mapping of requested QoS parameters to the transport QoS parameters.

- Admission control. It maintains information of available/allocated resources in both the radio and the IP transport network and performs resource reservation/allocation in response to new service requests, at call establishment or during vertical/horizontal handover, with a given QoS profile. From the radio point of view it takes into account, for example, the interference level

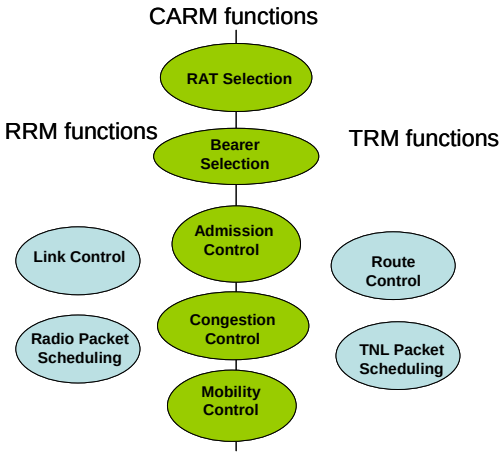


Figure 4.2: Proposed QoS management functions [46].

and the availability of codes (in a WCDMA radio interface), and from the transport network it can take into account, for example, the current occupation of the bottleneck link.

- Congestion control. It is in charge of taking the actions required to handle overload events in the radio or transport network side. This function will implement the operator's policy for congestion situations, for example, give priority to real-time/premium/business users over non-real-time/consumer users, etc, and take the necessary actions to reduce the duration of the congestion event. The methods used to handle congestion include a range of options, for the radio and for the transport part, which are operator/implementation specific. Congestion control needs coordinated actions from the radio and transport resource management. As an example, the possible actions range from changing the transport format combination Set (TFCS, UTRAN specific) to some users in the RRM part, to setup alternative routes or enforce link-sharing strategies for packet scheduling in the transport part.
- Mobility control (cell selection). This function is basically in charge of deciding the best cell to connect each terminal either at session set-up or in a handover process. Mobility control decisions can take into account measurements from the UE and the Node B and may take other

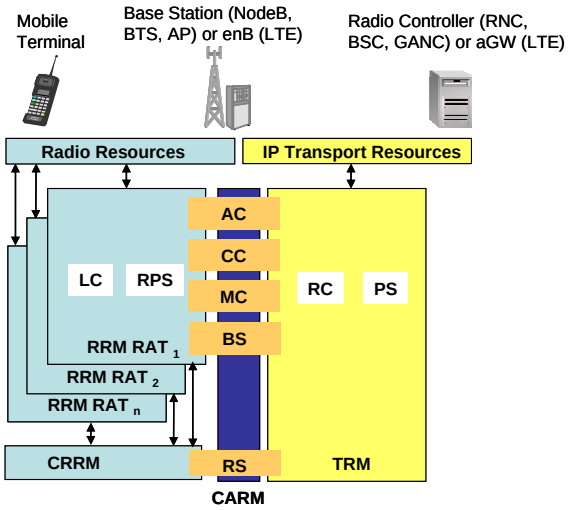


Figure 4.3: Role of CARM functions in a 3GPP architecture [46].

inputs, such as neighbor cell load, traffic distribution, transport and hardware resources and operator defined policies into account. We envisage that the transport resource availability could be checked before making a handover, since it is possible that a given cell with free radio resources cannot accept a handover call due to a congested link in the transport network. It is also possible that transport resource availability can influence the decision about which is the optimum cell to direct the handover to.

4.2.2.2. RRM Specific Functions

Objectives of the RRM functions identified in Figure 4.2 are detailed in the following:

- Radio link control. It is in charge of dynamically adjusting the radio link parameters of the mobile terminals in order to preserve the QoS for established sessions. This function will typically include power control and link adaptation mechanisms. Power control aims at dynamically adjusting the power transmitted by all the terminals in a given cell. The required power for each user depends on several factors like radio-link propagation losses, amount of interference in the cell and type of service and mobility of the user. Link adaptation functions dynamically adjust modulation and coding to maximize the throughput given the radio channel conditions.
- Radio packet scheduling. This function is in charge of maximizing resource occupation by scheduling packets for established sessions taking into account several factors, like the QoS of the session, the interference level of the cell and the channel quality for the particular user. Radio packet scheduling is a short term strategy that tries to use free resources that could otherwise remain underutilized.

4.2.2.3. TRM Specific Functions

Finally, on the TRM side the description of the two identified functions is provided below:

- TNL route control. This QoS management function is in charge of selecting appropriate routes in the transport network to guarantee the efficient use of TNL resources and the QoS requested by the TNL IP bearers. This function will be applied when setting up new QoS IP bearers and it is also envisaged that this function should continuously monitor the link utilization and buffer occupancy of the transport network nodes in order to prevent congestion and maintain efficient use of network resources. The implementation of this function will relay on appropriate load balancing techniques, path establishment with QoS constraints (like constraint routing-label distribution protocol, CR-LDP), as well as network resiliency mechanisms in case of link/node failures.
- TNL packet scheduling. This function is in charge of implementing, at the IP transport network nodes, the appropriate QoS queuing decisions so the different flows (or aggregates of flows) receive the right QoS treatment at every node. For that purpose the packets are marked at the ingress node with a mark that identifies them as belonging to a given QoS "behavior aggregate" (assuming a DiffServ scheme, for example). The implementation of TNL packet scheduling could range from simple priority queuing to the more sophisticated link-sharing techniques.

4.3. Analytical Evaluation of the Cell Selection

After the identification of different CARM functions, this section introduces a generic framework to analyze the benefits of a coordinated mobility control (cell selection) strategy in mobile networks with transport capacity limitations, regardless of the studied RAT. To this end, an

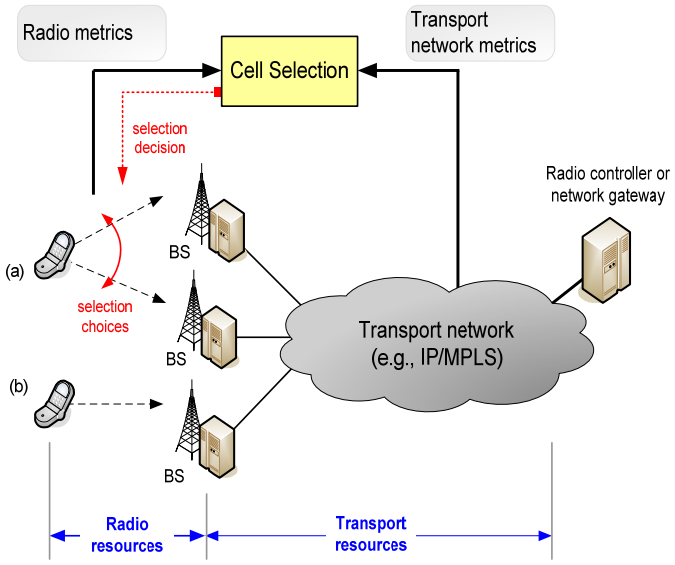


Figure 4.4: Cell selection framework

analytical model based on multi-dimensional Markov chains is developed assess the performance of different cell selection strategies that may consider both radio and transport constraints. The proposed model is used to compare different cell selection algorithms under scenarios with limited transport capacity. In particular, three different algorithms are modeled by means of the Markov chain model. Two algorithms are baseline cell selection algorithms that rely exclusively on radio criteria. The third algorithm is an enhancement approach that combines the classical minimum path loss (MPL) criterion and transport resource utilization.

4.3.1. Scope of Cell Selection Framework

Figure 4.4 illustrates the scope of the proposed cell selection framework. As shown in this figure, the decisions taken by the cell selection function are expected to be taken considering radio and transport resource status. In this scenario it is assumed there exists overlapped cell coverage in some locations of the service area. Thus, some terminals (e.g., terminal (a)) will have more than one candidate cell to handle their connections. A single RAT (homogeneous scenario) with a frequency reuse factor of one is considered. This scenario is claimed to be the most critical in terms of using information other than radio metrics to control the cell selection process. Notice that in the case of scenarios with multiple frequency layers and/or heterogeneous RATs, terminals can be distributed among frequency layers or RATs considering transport limitations as well, but the total or partial decoupling of the radio resource pools used in each frequency/RAT could make this decision less critical in terms of incurred radio degradation (i.e., there is no interference between frequency layers or different RATs).

It is important to remark that this framework is very generic due that the analysis of the cell selection problem is not particularized to a given RAT and transport network solution. Therefore, the cell selection framework depicted in Figure 4.4 could be applied from 2G/3G networks with a TDM/ATM backhaul (e.g., GSM/UTRAN) up to evolved architectures e.g., LTE [71], where the backhaul network can be deployed over IP/MPLS networking technologies.

Three cell selection strategies are analyzed under the considered cell selection framework:

- **Best Server Cell Selection (BS_CS).** Under this algorithm, terminals are always connected to their radio best-server cell [55]. Although this algorithm leads to use radio resources efficiently, when considering potential bottlenecks in the transport network, some new sessions can be blocked due to transport saturation of the best-server and it does not take advantage of spare transport capacity in some neighboring cells. Hence, this algorithm does not exploit any

transport capacity gain. The BS_CS algorithm is mainly used as the reference for the next two algorithms.

- **Radio Prioritized Cell Selection (RP_CS).** Unlike the previous algorithm, in the RP_CS algorithm all the cells having a difference in path loss, with respect to the best-server cell, below a certain path loss margin (PLM) are considered as candidate cells. Then, among the candidate cells whose transport is not saturated, the one showing minimum path-loss is selected. This mechanism clearly results in certain capacity gain in the use of the transport resources, but it comes at the expense of some radio degradation due to the potential selection of non-optimal cells. Notice that the RP_CS algorithm is the one commonly used in legacy networks with cell redirection support (e.g., cell redirection mechanism in UMTS [72]). Notice also that this algorithm is unaware of transport occupancy unless a transport blocking condition arises in the target cell.

- **Transport Prioritized Cell Selection (TP_CS).** The TP_CS algorithm behaves similar like the RP_CS algorithm whenever the transport occupancy on the candidate cells is below a certain threshold. However, when the transport occupancy exceeds such a threshold, the algorithm prioritizes the candidate cells according to their transport occupancy and not according to their radio path loss. The main idea behind this approach is to postpone as much as possible the transport saturation by means of a transport-aware distribution (transport aware load balancing) of the terminals with more than one candidate cell in the radio domain.

4.3.2. System Model and Problem Formulation

We consider a network with a set of N access points (APs) or BSs, denoted as $B=\{AP_1,\ldots,AP_N\}$, that covers a geographical area where a total of λ "calls" per second are generated (traffic generation is not necessarily uniformly distributed in the service area). The term "call" is used here in a wide sense since the problem formulated can be applied to different time scales, as illustrated in Figure 4.5. Particularly, a "call" can be thought as a session request when focusing on time scales above tenths of seconds (e.g., voice call or data session). As well, by "call" we can also consider a short data transfer within a session in a time scale of seconds or lower (e.g., packet calls within browsing data sessions). In any case, we consider that a call from a terminal i is required to be served by a single AP within B so that a given bit rate, denoted as R_i, is guaranteed. This implies that sufficient resources at both the air interface and the transport network should be allocated to terminal i to meet its bit rate requirement. The mean duration of the call service time is denoted as $1/\mu$ (s).

The amount of radio resources required by terminal i to support the requested bit rate R_i depend on the characteristics of the RAT under consideration. For instance, in code division multiple access (CDMA) systems a given transmission power has to be allocated to guarantee the requested bit rate, while in time division multiple access (TDMA) systems the bandwidth is assigned in the form of orthogonal time-slots. In order to avoid dealing with a specific RAT, in this chapter we follow a well-know approach to allocate connections to APs based on trying to serve each terminal

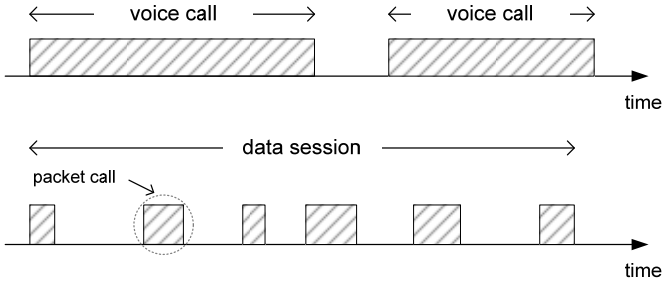


Figure 4.5: Scope of a call in two different time scales.

from the APs with the highest channel gain (i.e., minimum path loss criterion). Hence, in this work, the characterization of the air interface is mainly addressed by accounting for the channel gain where path loss and long-term radio channel conditions are captured. Over such a basis, we assume that the number of APs that can serve terminal *i* is limited to a candidate set of APs, denoted as CS_i. The candidate set is composed of all APs having a difference in channel gain with respect to the best-server cell (i.e., the AP with the best channel gain) not higher than a certain path loss margin, denoted as Δ (dB). For instance, the candidate set CS_i for terminal *i* is defined as:

$$CS_i = \left\{ \mathrm{AP}_n : \mathrm{AP}_n \in \mathrm{B},\ 10\log\left(\frac{\max\limits_{n=1\ldots N}(h_{in})}{h_{in}} \right) \le \Delta \right\} \tag{4.1}$$

where h_{in} is the channel chain between terminal *i* and AP_n. Hence, from the radio perspective, is it assumed that an efficient usage of the air capacity is directly coupled to the allocation of connections to APs with highest channel gains. Moreover, as our focus in this chapter in on studying capacity limitations due to transport overload, we do not consider potential radio capacity limitations that, otherwise, would be dependent on specific radio technologies (e.g., maximum downlink transmit power in CDMA or number of available time-slots in TDMA).

As for the transport network model, the transport resources needed to serve a given call can be directly related to the bit rate R_i required by each call. In this case, capacity limitations are accounted by assuming a given provisioned transport capacity in each AP_n denoted as C_n. Hence, the aggregated bit rate served by AP_n cannot exceed the provisioned capacity. This capacity limitation can be expressed by:

$$\sum_{i \in T_n} R_i < C_n \tag{4.2}$$

where T_n denotes the set of terminals being served by AP_n. Attending to this generic system model, the cell selection algorithms detailed in Section 4.3.1 are formulated as follows. The solution to the cell selection problem using the BS_CS algorithm is straightforward, as illustrated in Figure 4.6. For each terminal the cell with the highest cannel gain is selected. The new call from terminal *i* is accepted whenever there is sufficient available transport capacity in the best-server. The available capacity at the selected AP*n*, denoted in the flowchart as AC_n, is computed as:

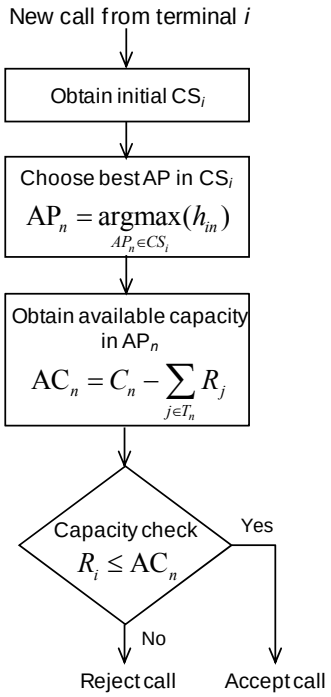


Figure 4.6: Flowchart of BS_CS algorithm.

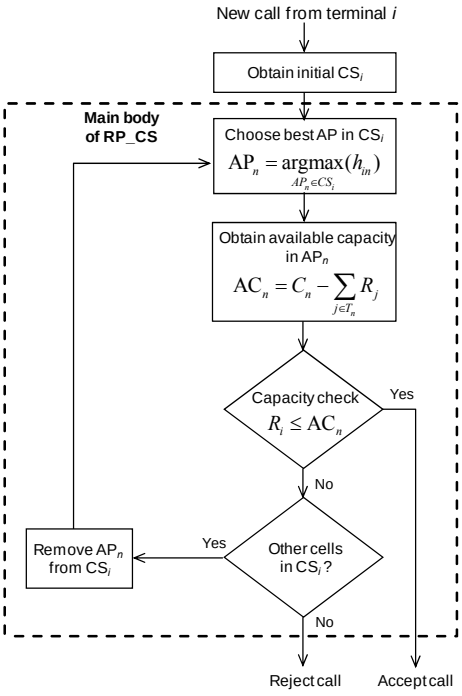


Figure 4.7: Flowchart of RP_CS algorithm.

$$\mathrm{AC}_n = C_n - \sum_{j \in T_n} R_j \tag{4.3}$$

where the second term in equation (4.3) denotes the aggregated data rate obtained summing up the bit rate R_j of each connection j already being served by AP_n. Hence, if the available capacity is sufficient to accommodate terminal i and meet its rate requirement R_i, the connection is accepted, otherwise it is rejected.

Unlike the BS_CS, in the RP_CS algorithm (see Figure 4.7) if a call from terminal i cannot be accommodated in the best server AP due to insufficient transport resources, it is verified whether it can meet its rate requirement in another AP present in its candidate set. Thus, with RP_CS a call from terminal i will only be rejected if there is no spare transport capacity in any of the APs belonging to its candidate set.

Finally, the TP_CS algorithm is shown in Figure 4.8. Notice that the TP_CS algorithm behaves similar to the RP_CS algorithm in situations when the transport load of APs is low. The distinction between low and high load transport conditions is determined based a threshold value of the transport capacity occupation, denoted as C_{th}, in the APs. To this end, the TP_CS algorithm first determines the available capacity of the APs in the candidate set of terminal i. Among the available capacity values of the APs it picks the minimum one and verifies if it is below the defined threshold value. If at least one of the APs is under high load conditions (i.e., at least one of the APs exceeding the threshold) the TP_CS algorithm determines the serving AP for terminal i based on a Bernoulli trial with a set of probabilities $\{\mathrm{P}_n \; ; \forall_n : \mathrm{AP}_n \in \mathrm{CS}_i \}$, where $\mathrm{P}_n = f(\mathrm{AC}_n)$ is a function of the available capacity of the APs in CS_i.

4.3.3. Analytical Model

We adopt an analytical approach to find the capacity gain in the utilization of transport resources that could be attained by using a coordinated cell selection strategy. Assuming some simplifying hypothesis, like having an infinite population of users with a single common service, Poisson distribution of call arrivals and exponential call service time, the capacity gain (in the sense of minimum transport blocking probability) can be obtained by solving the flow equations of a

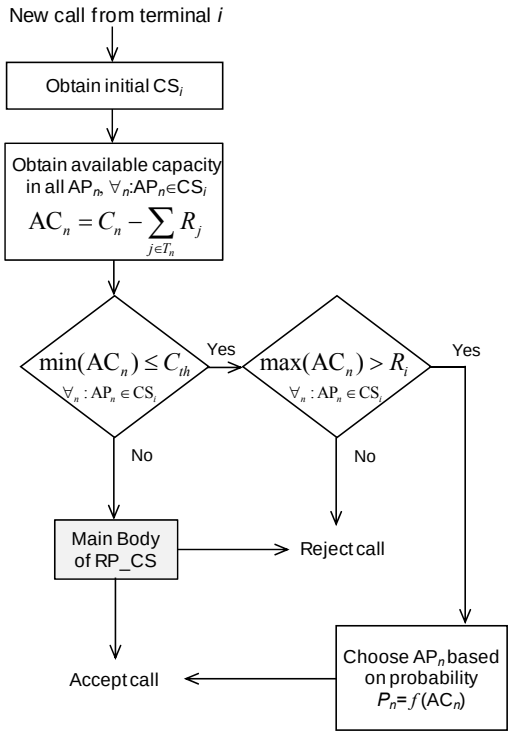


Figure 4.8: Flowchart of the TP_CS algorithm.

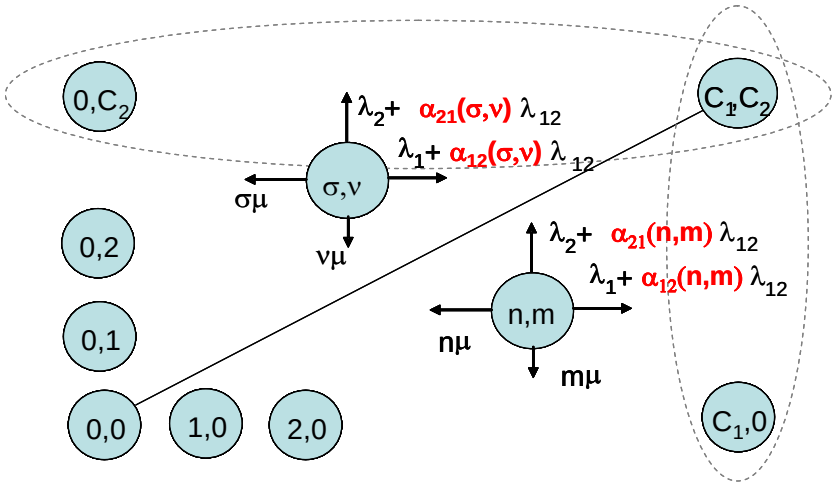


Figure 4.9: State diagram for a bi-dimensional Markov chain.

multi-dimensional Markov model. The analytical model is mainly intended to capture session blocking due to transport saturation.

Figure 4.9 shows the state diagram for a bi-dimensional case, where C_1 and C_2 are, respectively, the transport capacities of the links serving two candidate cells for cell selection. The transport capacity is measured in terms of the required throughput of a single connection of the kind of service requested by the users. So cell number 1 could accept, from the transport point of view, up to C_1 simultaneous connections of the target service. Using queueing system naming conventions we may say that cell 1 has C_1 servers. In state (n, m) we have n calls being served at cell 1 and m calls at cell 2. The upper left corner state is the blocking state. λ_1 and λ_2 are, respectively, the rates at which the connection requests arrive at cell 1 and at cell 2 from terminals that, due to coverage, technology or pricing constraints, have a single candidate cell, while λ_{12} is the rate of incoming connections for the terminals that can choose among cell 1 and cell 2.

The overall system rate of arrivals would be given by $\lambda=\lambda_1+\lambda_2+\lambda_{12}$ [calls/s]. μ [calls/s] is the service rate of accepted calls, while $\alpha_{12}(n, m)$ and $\alpha_{21}(n, m)=1-\alpha_{12}(n, m)$ are state dependent coefficients that must be chosen to optimize the system capacity. If transport resources occupation information is available (as in the case of the TP_CS algorithm), these coefficients should send a bigger fraction of the traffic from the terminals that can choose towards the cell with more free transport resources at any given moment. The proposed method to calculate $\alpha_{12}(n, m)$ with TP_CS is:

$$\alpha_{12}(n,m) = \frac{E(C_1 - n)}{E(C_1 - n) + E(C_2 - m)} \qquad (4.4)$$
$$\alpha_{12}(C_1, C_2) = 0$$

where $E(x)$ means the free Erlang capacity (given the desired system blocking probability) of a cell with x free servers. Notice that for the limiting states (those where one of the cells is blocked) we have $\alpha_{12}(C_1,m)=0$ and $\alpha_{21}(n,C_2)=0$, meaning that all the λ_{12} calls are directed to the not saturated cell. If detailed transport occupation information is not available (RP_CS) we still can use the model by taking $\alpha=1/2$ for all the states but the limiting ones. By taking $\alpha=1/2$ to model the RP_CS strategy we are assuming that the overlapped coverage regions show path-loss symmetry and so half of the λ_{12} calls are best-served by C_1 and half by C_2. We also assume that in the limiting states $\alpha_{12}(C_1,m)=0$ and $\alpha_{21}(n,C_2)=0$. Notice that current cell selection strategies already apply this criterion since the only information required is if the transport for any given cell is saturated.

This description can be generalized to a multi-dimensional model with any number of candidate cells. In the general case of an scenario where some terminals could have up to a maximum of N candidate cells, the state diagram of the corresponding N-dimensional Markov chain looks as

shown in Figure 4.10. The coordinates of each state are the number of transport resources currently occupied at each of the cells. We call C_i, (i=1,..,N), to the transport capacity of the links serving each candidate cell, where the transport capacity is measured in terms of the required throughput of a single connection of the kind of service requested by the users. Using queueing system naming conventions we may say that cell i has C_i servers.

In Figure 4.10 γ_i (i=1,..,N) are the probability flows departing from the current state (S_1, S_2,…,S_N) and reaching the neighboring state whose coordinate i is S_{i+1}, (i=1,…,N), while μ [calls/s] is the service rate of accepted calls. For the sake of clarity, the probability flows entering state (S_1, S_2,…,S_N) are not shown in Figure 4.10. The general expression for γ_i (i=1,…,N) can be written, for any state, as:

$$\gamma_i = \lambda_i + \sum_{\substack{n=1 \\ (n\neq i)}}^{N} \alpha_{in}\lambda_{in} + \sum_{\substack{n=1 \\ (n\neq i)}}^{N}\sum_{\substack{n=1 \\ (n\neq i)}}^{N} \alpha_{inm}\lambda_{inm} + \sum_{\substack{n=1 \\ (n\neq i)}}^{N}\sum_{\substack{m=n+1 \\ (m\neq i)}}^{N}\sum_{\substack{p=m+1 \\ (p\neq i)}}^{N} \alpha_{inmp}\lambda_{inmp} + \cdots + \sum_{\substack{n=1 \\ (n\neq i)}}^{N}\sum_{\substack{m=n+1 \\ (m\neq i)}}^{N}\cdots\sum_{\substack{q=N \\ (q\neq i)}}^{N} \alpha_{inm\ldots q}\lambda_{inm\ldots q} \qquad (4.5)$$

where $\lambda_{inm\ldots q}$ [calls/s] is the call arrival rate for terminals whose set of candidate cells is $\{i,n,m,...,q\}$ and the constants $\alpha_{inm\ldots q}$ are the state dependent load steering coefficients that must be chosen to optimize the system capacity. The terms in equation (4.5) stem from the fact that the full set of terminals in our scenario can be classified into N different disjoint subsets, $\{T_1, T_2,...,T_N\}$, where a terminal belongs to the subset T_j ($1 \le j \le N$) if it has exactly j candidate cells. Hence, two terminals from a given subset T_j would belong to the same traffic class if both share the same set of candidate sets. Any given traffic class is uniquely identified by the sequence of j different subscripts in $\lambda_{inm\ldots q}$. By convention, the first subscript in $\alpha_{inm\ldots q}$ means the cell where the fraction of traffic $\alpha_{inm\ldots q} \cdot \lambda_{inm\ldots q}$ is directed to. That is, $\alpha_{inm\ldots q}$ is the probability that any new session of class $\lambda_{inm\ldots q}$ is directed to cell i. The order of subscripts in $\lambda_{inm\ldots q}$ is not relevant. For instance, $\lambda_{123}=\lambda_{213}=\lambda_{312}$. Equation (4.5) is written in such a way that any traffic class which has cell i as a candidate cell appears only once. We call λ to the global system rate of arrivals, which is the addition of the call rates of all the 2N-1 different traffic classes. Within any subset T_j, and for any cell i, the traffic classes including cell i in the candidate set of cells is:

$$\binom{N-1}{j-1} \qquad (4.6)$$

So, the total number of additive terms in equation (4.5) is equal to:

$$\sum_{j=1}^{N}\binom{N-1}{j-1} = 2^{N-1} \qquad (4.7)$$

For the state (S_1, S_2,..,S_N) and for the traffic from terminals belonging to subset T_j and class

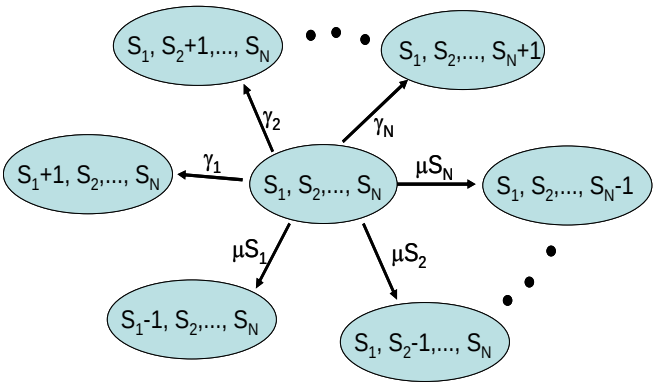


Figure 4.10: Generalized state diagram for an N-dimensional Markov chain.

$\lambda_{inm...q}$, we calculate $\alpha_{inm...q}$ in one of two ways:

$$\alpha_{inm...q}(S_1,..,S_N) = \frac{1-\delta(C_i - S_i)}{\sum_{k=1}^{j}\left[1-\delta(C_{g(k)} - S_{g(k)})\right]} \qquad (4.8)$$
$$\alpha_{inm...q}(C_1,C_2,..,C_N) = 0$$

$$\alpha_{inm...q}(S_1,..,S_N) = \frac{E(C_i - S_i)}{\sum_{k=1}^{j} E(C_{g(k)} - S_{g(k)})} \qquad (4.9)$$
$$\alpha_{inm...q}(C_1,C_2,..,C_N) = 0$$

In equation (4.8), $\delta(n)$ means Kronecker's delta. In equation (4.9) $g(1)=i$, $g(2)=n$, $g(3)=m,\ldots,g(j)=q$ and $E(x)$ means the free Erlang capacity (given the desired system blocking probability) of a cell with x free servers. Equation (4.8) models a scenario where all terminals connect to their best server (unless it is saturated) and assumes symmetric coverage, so that traffic originated in the overlapped coverage regions is equally shared by all the overlapping cells (this would be the case of RP_CS strategy with uniformly distributed spatial traffic generation). Equation (4.9), in turn, makes use of transport occupation information to send a bigger fraction of the traffic from the terminals that can choose towards the cell with more free transport resources at any given moment (this would be the case of the TP_CS strategy). Notice, in (4.8) and (4.9), that in the limiting states no traffic is sent towards the saturated cells, but the traffic is still served by the cells with free resources, since the sum of all the load steering coefficients for a given traffic class equals unity.

Our design proposal for the TP_CS strategy is to use equation (4.9) only at the states where C_i-$S_i<L$ for at least one value of i ($i=1,...,N$). In the rest of the states we will apply equation (4.8). So L is a parameter of the TP_CS strategy. For the RP_CS strategy equation (4.8) is always used. Since the total probability rate departing from any of the states must be zero, it is possible to write a linear system of equations to find the probability of being at any of the states, and so the system blocking probability.

4.3.4. Performance Metrics

Two key performance metrics to compare cell selection strategies can be derived from the proposed Markov Model. One metric is the trunking gain and the other is the degree of radio degradation.

4.3.4.1. Trunking Gain

We aim to determine the capacity gain in the utilization of transport resources that cell selection strategies can achieve in scenarios where the transport network constitute the resource bottleneck. Such capacity gain is here referred to as trunking gain. Given a traffic distribution and cell selection strategy, by solving the Markov Model, the probability of being in each state is determined and so the system blocking probability, that is, the fraction of sessions which are rejected due to saturation states. By reversing the procedure (using a numerical root-finding algorithm) it is possible to fix the desired system blocking probability and find the value of λ that leads to that blocking probability. By repeating this method for the cases of RP_CS and TP_CS cell selection and for the reference case, the trunking gains, denoted as t_g, can be found as:

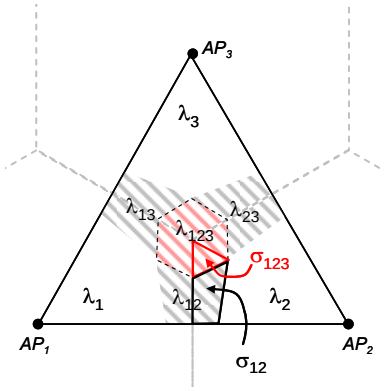


Figure 4.11: Regular cell deployment and regions where the path-loss difference to any of two or any of three APs is less than a given margin.

$$t_g\big|_{RP\_CS} = 100 \cdot \left(\frac{\lambda_{RP\_CS}}{\lambda_{BS\_CS}} - 1 \right) (\%)$$
$$t_g\big|_{TP\_CS} = 100 \cdot \left(\frac{\lambda_{TP\_CS}}{\lambda_{BS\_CS}} - 1 \right) (\%) \qquad (4.10)$$

where λ_{RP_CS} and λ_{TP_CS} are, respectively, the values of λ that lead to the desired blocking probability when applying either the RP_CS or TP_CS cell selection strategies. λ_{BS_CS} is the traffic that leads to the desired blocking probability when applying the reference strategy BS_CS.

The reference strategy for calculating the trunking gain is the case where all new sessions are connected always to their best-server. If the best-server is saturated then the session is dropped. This means that only $\lambda_1, \lambda_2,..,\lambda_N$ are different from zero and $\gamma_i=\lambda_i$ (i=1,...,N). In the reference case, and assuming equal transport capacity per cell, the blocking probability, for a given λ and a given number of servers per cell, can be obtained with the classical Erlang-B expression.

4.3.4.2. Path Loss Increase

The method used to compute the path-loss increase statistics is detailed in this section. We assume a single-RAT scenario using a regular hexagonal cell deployment with uniform user distribution per square meter. The analysis is focused on the coverage region of three sector cells (see Figure 4.11) served by the access points AP_1 (C_1 servers), AP_2 (C_2 servers) and AP_3 (C_3 servers), respectively. In Figure 4.11 the best-server area of an AP is the region where that AP is the closest one. In this case, the global rate of arriving calls, denoted as λ, can be distributed into the following traffic classes: $\lambda=\lambda_1+\lambda_2+\lambda_3+\lambda_{12}+\lambda_{13}+\lambda_{23}+\lambda_{123}$, where the exact distribution of rates is found by numerical integration of the areas of overlapped coverage, that is, the grey and the pink dashed areas in Figure 4.11. Obviously, the width of the overlapped coverage regions increases for increasing values of the considered path loss margin. In order to find the boundaries of the overlapped coverage regions a simple exponential power decay law has been assumed:

$$P_R = P_T \cdot k \cdot d^{-\beta} \qquad (4.11)$$

where P_T and P_R are, respectively, the power transmitted by the serving AP and the power received at the terminal, d is the distance from the terminal to the serving AP, k^{-1} is the attenuation for d=1 and β=3.5. For our purposes the value of k is irrelevant, since we are only interested in the difference of attenuations (in dBs) to each of the candidate cells.

Using numerical integration, the mean path-loss increase for a single call of class λ_{12}, denoted as ρ_{12}, originated inside the best-server area of AP_2 but connected to AP_1 is computed as:

$$\rho_{12} = \frac{\iint_{\sigma_{12}} \left(\frac{d_1}{d_2} \right)^{\beta} d\sigma_{12}}{\sigma_{12}} \tag{4.12}$$

where the integral is extended over the area σ_{12} (see Figure 4.11). In equation (4.12), d_1 and d_2 are, respectively, the distances from $d\sigma_{12}$ to AP_1 and AP_2. In the same way the mean path-loss increase for a single call of class λ_{123}, denoted as ρ_{123}, originated inside the best-server area of AP_2 but connected to AP_3 is computed as:

$$\rho_{123} = \frac{\iint_{\sigma_{123}} \left(\frac{d_3}{d_2} \right)^{\beta} d\sigma_{123}}{\sigma_{123}} \tag{4.13}$$

where σ_{123} is the area shown in Figure 4.11 and d_2 and d_3 are, respectively, the distances from $d\sigma_{123}$ to AP_2 and AP_3. As a simplifying worst-case assumption we take a mean path-loss increase equal to ρ_{123} also for the traffic of class λ_{123} generated inside the best-server area of AP_2 but connected to AP_1.

In order to calculate the global mean path-loss increase we must average ρ_{12} and ρ_{123} over all the possible system states. For a given state (S_1, S_2, S_3) (where $0 \le S_1 \le C_1$, $0 \le S_2 \le C_2$, $0 \le S_3 \le C_3$), the probability that a call of class λ_{12}, generated inside the best-server area of AP_2, is not connected to AP_2 is:

$$\begin{cases} 0; & (\alpha_{21}(S_1, S_2, S_3) \ge 0.5) \\ \dfrac{\frac{\lambda_{12}}{2} - \alpha_{21}(S_1, S_2, S_3)\lambda_{12}}{\frac{\lambda_{12}}{2}} = 1 - 2\alpha_{21}(S_1, S_2, S_3); & (\alpha_{21}(S_1, S_2, S_3) < 0.5) \end{cases} \tag{4.14}$$

Notice that in the case that the traffic is considered fully symmetric (i.e., the total traffic generated inside the best-server area of AP_2 is $0.5\lambda_{12}$, of which, if $\alpha_{21}(S_1, S_2, S_3)<0.5$, only $\alpha_{21}(S_1, S_2, S_3)\cdot\lambda_{12}$ calls are connected to AP_2). If $\alpha_{21}(S_1, S_2, S_3) \ge 0.5$ then AP_2 is already absorbing all its traffic quota. The mean path-loss increase due to traffic of class λ_{12} is then:

$$\overline{\Delta}_{12} = 1 + (\rho_{12} - 1) \cdot \sum_{S_1=0}^{C_1} \sum_{S_2=0}^{C_2} \sum_{S_3=0}^{C_3} u\left(\frac{1}{2} - \alpha_{21}(S_1, S_2, S_3) \right) \cdot \left[1 - 2\alpha_{21}(S_1, S_2, S_3) \right] \cdot \Pr\left(S_1, S_2, S_3 \right) \tag{4.15}$$

where $u(x)$ is the unitary step function (with $u(0)=0$) and $\Pr(S_1, S_2, S_3)$ is the probability of state (S_1, S_2, S_3).

In the same way, the probability that a call of class λ_{123}, generated inside the best-server area of AP_2, is not connected to AP_2 is:

$$\begin{cases} 0; & (\alpha_{231}(S_1, S_2, S_3) \ge \frac{1}{3}) \\ \dfrac{\frac{\lambda_{123}}{3} - \alpha_{231}(S_1, S_2, S_3)\lambda_{123}}{\frac{\lambda_{123}}{3}} = 1 - 3\alpha_{213}(S_1, S_2, S_3); & (\alpha_{213}(S_1, S_2, S_3) < \frac{1}{3}) \end{cases} \tag{4.16}$$

If $\alpha_{213}(S_1, S_2, S_3) \ge 1/3$ then AP_2 is already absorbing all its traffic quota. The mean path-loss increase due to traffic of class λ_{123} is then:

$$\overline{\Delta}_{123} = 1 + (\rho_{123} - 1) \cdot \sum_{S_1=0}^{C_1} \sum_{S_2=0}^{C_2} \sum_{S_3=0}^{C_3} u\left(\frac{1}{3} - \alpha_{213}(S_1, S_2, S_3) \right) \cdot \left[1 - 3\alpha_{213}(S_1, S_2, S_3) \right] \cdot \Pr\left(S_1, S_2, S_3 \right) \tag{4.17}$$

where ρ_{123} is the mean path-loss increase for a single call of class λ_{123} originated inside the best-server area of AP_2 but not connected to AP_2. Finally the global mean path-loss increase, for calls originated inside the overlapped coverage areas is:

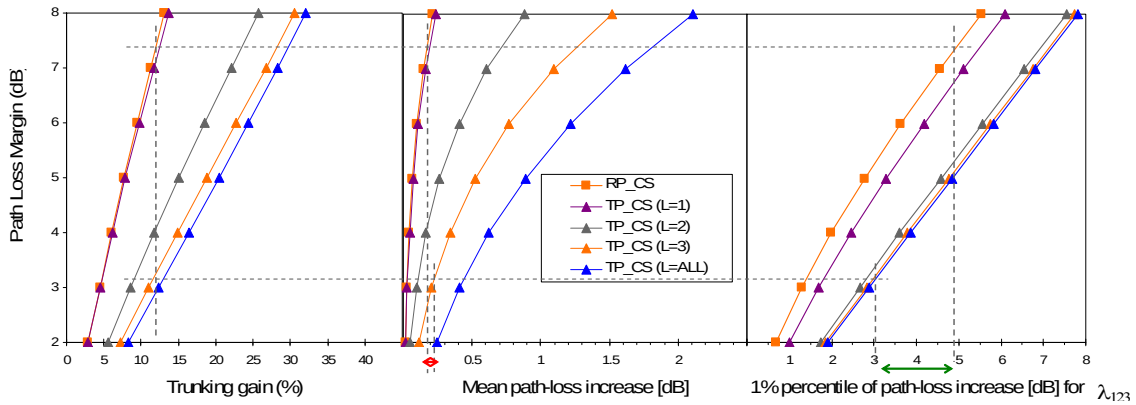


Figure 4.12: Comparison of TP_CS and RP_CS strategies (symmetric AP transport capacities).

$$\overline{\Delta} = 10 \cdot \log\left(\overline{\Delta}_{12} \frac{\lambda_{12}}{\lambda_{12} + \lambda_{123}} + \overline{\Delta}_{123} \frac{\lambda_{123}}{\lambda_{12} + \lambda_{123}} \right) [\mathrm{dB}] \qquad (4.18)$$

In order to get more insight about the statistics of the path-loss increase for this type of calls, the 1% percentile of the path-loss increase (defined as the path loss increase that is only exceeded with probability 1%) for calls of class λ_{12} and for calls of class λ_{123} is also calculated. The applied methodology consists of a numerical calculation of the cumulative distribution function of $(d_1/d_2)\beta$ and $(d_3/d_2)\beta$ inside the areas σ_{12} and σ_{123} (respectively) followed by averaging over all possible system states as in (4.15) and (4.17).

4.3.5. Results and Discussion

In this paragraph the results obtained for the scenario described in the previous paragraph are presented. The desired system blocking probability is 1% in all cases. Figure 4.12 is a comparison of RP_CS and TP_CS cell selection strategies for symmetric AP transport capacities. In particular, a number of servers equal to 8 is considered for the three cells (this can be a high number when focusing on high data rate services over cellular cells, e.g., 384 Kbps in a UMTS cell). The curves are parameterized by the value of L, which was explained in section 4.3.3 (L=ALL means that equation (4.9) is used at all the states).

In the leftmost graph we realize that the higher the PLM, the higher is the trunking gain (more terminals can choose among several cells), but the TP_CS strategy clearly outperforms the RP_CS strategy. It is also evident that the TP_CS strategy with L=3 achieves almost the same trunking gain

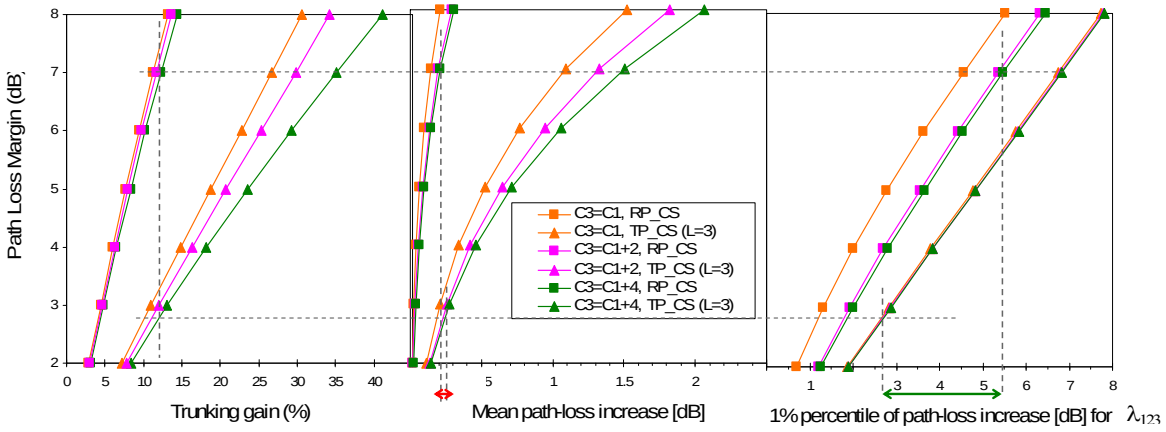


Figure 4.13: Comparison of TP_CS and RP_CS strategies (asymmetric AP transport capacities).

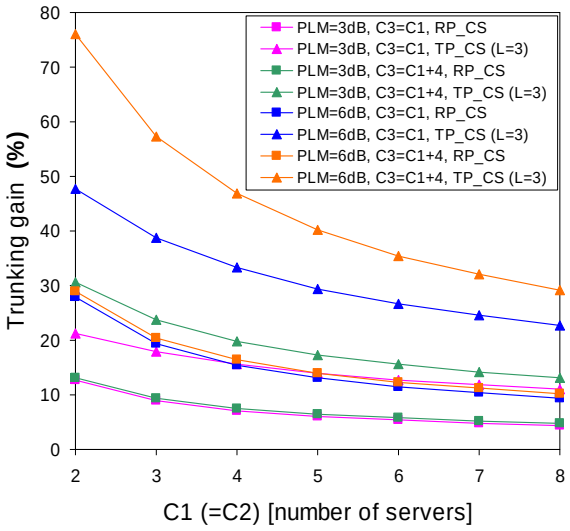


Figure 4.14: Trunking gain as a function of the number of servers per AP (symmetric and asymmetric transport capacities).

as the TP_CS with *L*=ALL but leads to less path-loss increase.

In the central graph we see that the trunking gain is achieved at the cost of certain mean path-loss increase. But if we compare the RP_CS strategy and the TP_CS strategy for the same trunking gain (see dashed lines), the path-loss increase of both strategies is almost the same. So, even taking into account our estimation of the radio degradation, the TP_CS strategy still outperforms the RP_CS strategy. Finally, the rightmost graph of Figure 4.12 allows comparing the 1% percentile of the path-loss increase (defined as the path loss increase that is only exceeded with probability 1%) for the λ_{123} traffic. At the same trunking gain we can expect almost 2 dB less path-loss increase, for the worst case calls, with the TP_CS strategy.

Figure 4.13 shows the trunking gains that can be obtained assuming an asymmetric distribution of transport resources, that is $C_2=C_1=8$ and $C_3=C_1+n$ (n=0,2,4). These results address the case where one of the candidate cells (cell 3) has an upgraded backhaul link while the others do not. The same comments already given for Figure 4.12 apply also here, but notice how the trunking gain of the RP_CS strategy is independent of the asymmetry in the number of servers, while an increased asymmetry leads to an increased trunking gain of the TP_CS strategy. On the other hand, Figure 4.14 shows the dependency of the trunking gains with the number of servers for symmetric and asymmetric transport capacities. For a low number of servers (low capacity links and high speed services) the trunking gains of the TP_CS strategy can be quite high.

Attending to the obtained results, it has been demonstrated that cell selection strategies that account for both radio and transport network status are able to provide a higher capacity gain than strategies that are only based on radio criteria. The analytical modeling of the cell selection problem has been carried out without focusing on a specific RAT. Hence, we consider that the proposed cell selection strategy that is based on the CARM concept deserves to be further evaluated in the context of specific RAT implementations.

4.4. Summary

This chapter has introduced the CARM framework, where the status of the transport network is incorporated in the decision making process along the radio status information. This approach is justified by the fact that backhaul bottleneck situation are likely to appear in some network scenarios, as argued in Chapter 2. Over such a basis, we have identified a set of functions which could be developed within the scope of the envisaged CARM functional model. The rest of the

chapter has been devoted to evaluate the cell selection problem considering a generic mobile RAN scenario where the transport network could represent the network bottleneck. To this end, an analytical framework developed in the context of the AROMA project [46] is used to assess the benefits of including metrics related to transport resource occupancy in the decision-making process of a cell selection strategy. The proposed framework is used to estimate the capacity gains, and the possible radio degradation (i.e., path loss increase), that can be attained to the proposed cell selection strategy. Performance analysis have shown that, under scenarios where transport network resources can get saturated, it is possible to make use of the transport status occupation to drive cell selection, even in those scenarios where a cell selection other than the non best radio server can be considered a priori as not appropriate. It has been demonstrated that the proposed cell selection approach is able to mitigate transport limitations by conveniently allocating those connections that have less impact on the radio degradation.

5 *Evaluation of BS Assignment Problem in WCDMA Cellular Networks*

5.1. Introduction

The coordinated access resource management (CARM) developed in Chapter 4 defines a set of resource management functions that takes into account transport network metrics within the resource allocation process. Furthermore, in Chapter 4 has also provided an initial assessment of a coordinated cell selection strategy (that combines the classical path loss criterion and transport resource utilization) in the context of a generic framework, that is, without focusing on a specific radio access technology (RAT). So far, it has been demonstrated that, in scenarios where the transport network constitutes the bottleneck, a coordinated (radio and transport) cell selection strategy outperforms cell selection schemes that rely exclusively on radio criteria, in terms of the capacity gain in the utilization of transport resources each strategy can achieve.

Over such a basis, this chapter is devoted to evaluate the proposed the CARM framework, and specifically focusing on the cell selection, also referred to as base station (BS) assignment, problem in a scenario of a single RAT deployment. In this context, considering the cell selection framework introduced in Figure 4.4, in this chapter we assume that the radio access is based on wideband code division multiple access (WCDMA). Therefore, it is assumed that capacity limitations (bottlenecks) could appear at the air interface or transport network.

The structure of the chapter is as follows. In Section 5.2 the related work to the BS assignment problem in WCDMA cellular networks is firstly presented, followed by the system model in Section 5.3. In Section 5.4 three different BS assignment strategies to be evaluated are defined. Afterwards, in Section 5.5 a simulated-annealing BS assignment algorithm is proposed to solve the BS assignment problem in the context of WCDMA cellular networks. Simulation results and performance evaluation of the considered BS assignment strategies are discussed in Section 5.6. A summary of the main conclusions of the chapter is finally given in Section 5.7.

5.2. Related Work

The basic idea of the BS assignment problem is to select, from a set of candidate BSs, the most appropriate BS or group of BSs (in case of macro-diversity support) to handle radio transmission to/from mobile users. The literature contains a wide range of contributions, in which different formulations and algorithmic solutions to this problem have been proposed considering diverse assignment criteria (e.g., minimizing transmitted power) as well as attending to possible constraints (e.g., satisfying a minimum signal to interference and noise ratio, SINR). Following paragraphs detail related work on the topic.

One of the most common BS assignment approaches is the minimum path loss (MPL), where the BS assignment problem can be formulated as an unconstrained optimization problem. The "best" solution is obtained straightforwardly by assigning each user to the BS that provides the highest radio link gain [55]. This approach constitutes the core of many BS assignment algorithms used in current 2G/3G cellular systems where absolute and/or relative received signal level thresholds are used to decide upon the serving BS. Its main disadvantage comes from the fact that interference and radio load conditions are not considered in the assignment process.

Another well-known BS assignment approach consists of incorporating both channel gain and interference level in the allocation criterion, which is particularly important in cellular environments. This is normally done by considering the SINR in the assignment process. In this context, a joint BS and power assignment problem is formulated in [73] as a constrained optimization problem targeted to minimize users' uplink transmitted power while satisfying a given SINR constraint in each connection. Yates [73] develops a distributed iterative algorithm referred to as minimum power assignment (MPA), capable to find an optimal solution for the uplink of any single channel interference based power controlled system. Additional constraints on maximum or minimum transmit power can be added to the MPA algorithm straightforwardly.

It is worth noting here that the MPA algorithm cannot be directly applied to find an optimal solution for the downlink channel where it has been shown [74] that there does not exist a Pareto optimal solution minimizing all individual BSs' transmitted powers. This is because in the uplink, choosing the BS that makes the user to transmit with less power directly turns into a reduction in the interference level observed by other users, thus resulting beneficial for all the users which could even decrease their transmitted power while satisfying the SINR constraint. On the contrary, in the downlink case, moving a user from one BS to another so that the power allocated in the new BS is less than the power required in the previous one, does not necessarily contribute to reduce the level of interference to all the users, since users close to the new BS are now even going to see a higher level of interference. So, in the downlink case, a mutual dependency exists between the SINR values and the BS assignment because of co-channel interference (CCI) that brings more complexity into the BS assignment problem.

In [75], the BS assignment and power allocation problems for the downlink are formulated using the concepts of utility and pricing. A utility function is defined to reflect the satisfaction level of each user with respect to the achieved transmission rate, which in turn depends on the BS assignment and power allocation of the rest of users. The mutual dependency issue is avoided in [75] assuming full load conditions, that is, each BS is assumed to transmit at its maximum power irrespective of the final BS assignments. This assumption leads to authors to: (a) decouple the problem of BS assignment from the power assignment in each BS; and (b) formulate the utility function of a user only dependent on the power allocated to this user (without dependence of the power allocation for the other users). Over such a basis, work in [75] first formulates an optimization problem to obtain the transmission powers of users assigned to a given BS that maximize the sum of all user utilities (i.e., system utility), subject to constraints on the maximum transmission power at the BS and a maximum data rate for each user. The power allocation problem is solved by assuming a fixed BS assignment and using an iterative pricing-based algorithm to share the power of the BSs among its "assigned" users. In particular, each BS follows a bidding process to allocate power to those users, among its assigned ones, willing to pay the more for the resources, whenever the maximum power limit of the BS is satisfied. This approach effectively maximizes the system utility but could also lead to starvation of all the resources by some privileged users. Then, as to the BS assignment process, users left out of the power allocation in their assigned BS (done with a given periodicity) are assigned to a "more cheaper" neighboring BS (resource prices are broadcasted after each bidding period). As a result, radio load among BSs is balanced through the dynamic reassignment of non served users from heavily loaded (more expensive) BSs to lightly loaded (cheaper) BSs.

5.3. System Model

The performance analysis of the BS assignment problem will be addressed considering a single frequency WCDMA network. We focus on the downlink because it is usually considered the more restrictive link due to the asymmetric bandwidth demand between the downlink and the uplink data services. The network consists of N BSs that cover a geographical area in which, at a given instant, there are M users that have to be served. It is assumed that resources in any BS in the system are constrained by two factors: the maximum power limit in the radio interface and the provisioned capacity in the backhaul network. The system state is characterized by an M x N matrix, hereafter referred to as $B = \{b_{ij}\}$, that denotes the BS assignments of all users at a given instant. In particular, $b_{ij} = 1$, if BS j is assigned to user i, and $b_{ij} = 0$, otherwise. In the following we formulate both radio and backhaul related constraints.

5.3.1. Air Interface Constraint Definition

In the downlink of a WCDMA air interface, the signal to interference and noise ratio (SINR) observed by user i have to fulfill the following expression [55]:

$$\mathrm{SINR}_i = \frac{\dfrac{P_{ij}}{L_{ij}}}{\dfrac{(P_j - P_{ij})}{L_{ij}}(1-\alpha_i) + \sum_{k=1,k\neq j}^{N} \dfrac{P_k}{L_{ik}} + P_N} \geq \frac{R_i}{W}\left(\frac{E_b}{N_0}\right)_{\min,i} = \gamma_i \qquad (5.1)$$

where $(E_b/N_0)_{\min,i}$ is the minimum bit energy over noise power spectral density requirement, P_{ij} is the required transmit power devoted to user i being served by BS j, P_N is the noise power at the user terminal, R_i is the bit rate of user i, W is the chip rate, P_k is the total transmit power of BS k, L_{ik} is the path loss between another BS k and user i, and α_i is the orthogonality factor seen by user i ($\alpha_i = 1$ means perfect orthogonality). Thus, attending to equation (5.1), the required transmitted power for user i being served by BS j can be solved as follows:

$$P_{ij} = \frac{\gamma_i}{1+\gamma_i(1-\alpha_i)}\left[(1-\alpha_i)P_j + \sum_{k=1,k\neq j}^{N} \frac{L_{ij}}{L_{ik}} P_k + L_{ij}P_N\right] \qquad (5.2)$$

Thus, the required total transmission power at BS $j \in \{1,\ldots,N\}$ can be obtained by summing up the power of each served user i, which should be lower than or equal to the radio constraint:

$$P_j = \sum_{i=1}^{M} P_{ij} b_{ij} \leq \mathrm{P}_{\mathrm{T,max}} \qquad (5.3)$$

where $P_{\mathrm{T,max}}$ is the maximum transmit power limit of base stations. The resolution of (5.3) when the BS assignment for each user in the system is known is a well studied problem, and feasibility conditions and optimal power allocation can be obtained following the algorithm described in [73]. It is also worth noting that when focusing on the joint power control and BS assignment problem there is not always a Pareto optimal power vector in the downlink as is the case in the uplink [74].

5.3.2. Transport Constraint Definition

In the modeling of the system, it is also assumed that each BS is constrained by the available transport network capacity. The transport capacity provisioned for a given BS is fixed as a function of the amount of traffic that the BS can serve over the air interface. The adopted procedure is as follows.

First, the capacity of the downlink channel in a WCDMA network is computed. In this sense, a common used approach to estimate the air interface downlink capacity is based on the computation of the downlink load factor n_{DL} defined as [76]:

$$\eta_{\mathrm{DL}} = \sum_{i=1}^{M} \frac{R_i}{W} \left(\frac{E_b}{N_0} \right)_{\min,i} \left((1-\alpha_i) + f_{\mathrm{DL},i} \right) \tag{5.4}$$

where M is the number of users served by a given BS, $f_{\mathrm{DL},i}$ is the other-to-own cell received power ratio of user i at the position where it is located that is given by:

$$f_{\mathrm{DL},i} = \sum_{k=1,k\neq j}^{K} \frac{L_{ij}}{L_{ik}} \tag{5.5}$$

Equation (5.4) means that as the load factor move towards one, the downlink capacity approaches to its maximum air interface pole capacity value [76]. Over such a basis, under the assumption that all mobile users have similar service characteristics and quality requirements (i.e., service type, service bit rate R, and E_b/N_0 requirements), it can be shown that the air interface capacity, denoted here as C_{air}, can be estimated as follows:

$$C_{\mathrm{air}} = R \cdot \left\lfloor \left(\frac{W}{R\left(E_b/N_0\right)\left[(1-\alpha)+f_{\mathrm{DL}}\right]} \right) \cdot \left(\frac{\left[(1-\alpha)+f_{DL}\right]}{P_{\mathrm{N}} \cdot L_{\mathrm{avg}} / P_{\mathrm{T}}^{\max} + \left[(1-\alpha)+f_{DL}\right]} \right) \right\rfloor \tag{5.6}$$

where $\lfloor x \rfloor$ is the floor function (i.e., the largest integer value that is equal to or less than x), α is the average orthogonality factor in the cell that is defined as,

$$\alpha = \frac{1}{M} \sum_{i=1}^{M} \alpha_i \tag{5.7}$$

f_{DL} is the average ratio of other-to-own cell BS power received by users, computed as:

$$f_{\mathrm{DL}} = \frac{1}{M} \sum_{i=1}^{M} f_{\mathrm{DL},i} \tag{5.8}$$

and L_{avg} is the average path loss over all connections in the cell:

$$L_{\mathrm{avg}} = \frac{1}{M} \sum_{i=1}^{M} L_{ij} \tag{5.9}$$

The air interface pole capacity, denoted as $C_{\mathrm{air,p}}$, can be derived from equation (5.6) by neglecting the term accounting for the ratio between noise power and received power in front of the terms accounting for the effect of inter-cell and intra-cell, this latter due to non perfect orthogonality. The air interface pole capacity can be expressed as:

$$C_{\mathrm{air,p}} \approx R \cdot \left\lfloor \left(\frac{W}{R\left(E_b/N_0\right)\left[(1-\alpha)+f_{DL}\right]} \right) \right\rfloor \tag{5.10}$$

It is worth noting that equations (5.6) and (5.10) do not consider activity factors so that obtained capacity would correspond to the maximum number of simultaneous transmissions.

In our analysis, the backhaul capacity of BSs, denoted as C_{trans}, in the system is related to the air interface pole capacity by means of a multiplicative dimensioning factor, denoted as β, as it is shown below:

$$C_{\mathrm{trans}} = \beta \cdot C_{\mathrm{air,p}} \tag{5.11}$$

Notice that a dimensioning factor of β=1 in equation (5.11) would mean that the transport capacity has been dimensioned to satisfy the downlink air interface pole capacity resulting from the planning process. It is worth noting that, as pointed out in [77], in order to account for the overheads of the transport protocol stack and signaling margins for control and O&M traffic, a multiplicative factor could also be included in equation (5.11). The inclusion of such new term do not change the analysis presented in this section. Over such a basis, the transport network constraint of a given BS j in the system can be expressed as:

$$\sum_{i=1}^{M} R_i b_{ij} \leq C_{\text{trans}} \tag{5.12}$$

This implies that the aggregated traffic of all users assigned being served by BS *j* should not exceed the total available transport network capacity of the BS.

5.3.3. BS Assignment Problem Formulation

The BS assignment problem that considers radio and transport resources can be formulated as a constrained optimization problem where the aim is to maximize the aggregated utility of all user assignments, referred to as system utility, subject to the availability of resources at the air interface and the transport network. This is can be expressed as:

$$\max_{b_{ij}} \left(\sum_{i=1}^{M} \sum_{j=1}^{N} u_{ij} b_{ij} \right) \tag{5.13}$$

$$s.t. \quad \sum_{i=1}^{M} P_{ij} b_{ij} \leq P_{\text{T,max}} \qquad j = 1,...,N \tag{5.14}$$

$$\sum_{i=1}^{M} R_i b_{ij} \leq C_{\text{trans}} \qquad j = 1,...,N \tag{5.15}$$

$$\sum_{j=1}^{N} b_{ij} \leq 1 \qquad i = 1,...,M \tag{5.16}$$

$$b_{ij} \in \{0,1\} \tag{5.17}$$

where u_{ij} in (5.13) is defined as the utility of the assignment of user *i* being served by BS *j*, (5.14) and (5.15) are the air interface and transport network constraints, respectively, while b_{ij} is the assignment indicator that is equal to one if user *i* is assigned to BS *j*, or zero otherwise.

The above optimization problem is very difficult to tackle due to the inherent characteristic of the downlink, where there exists a mutual dependency between the power required by each user and the assignment solution of the rest of users in the system. The complexity associated with the computation of constraint (5.14) is elevated due to its non-lineal properties.

The approach used in this work to solve the BS assignment problem is as follows. We define a utility function in such a way that it expresses the degree of fulfillment to the resources constraints that each user-BS combination provides. This allows us to develop a BS assignment algorithm that does not require to explicitly consider constraints (5.14) and (5.15). Instead, the availability of resources at each BS are included within the utility function. In the following we present the evaluated BS assignment strategies, and their corresponding utility functions, and afterwards the BS assignment algorithm used to implement the defined assignment strategies is detailed.

5.4. Base Station Assignment Strategies

Three possible BS assignment strategies are defined in this section. Each BS assignment strategy relies on a particular assignment criterion, which is the set of rules that are followed to perform the assignment process of users. The analyzed BS assignment strategies are referred to as MPL, load balancing radio (LBR) and joint radio and transport (JRT).

The MPL strategy assigns each user to the BS with highest channel gain. The LBR strategy aims to balance the radio load in the network. It performs the assignment of users exclusively based on power level consumption at BSs. Likewise, in the JRT approach radio resources are also optimized, but in those cases where transport resources become the network bottleneck it also aims to find an assignment solution that could constitute a tradeoff between optimizing radio resources and preventing backhaul congestion (due to the assignment of users to BSs with insufficient

transport capacity). The latter assignment strategy is the scheme proposed to account also for potential backhaul capacity limitations in the BS assignment process.

The behavior of each BS assignment strategy is modeled using the concept of utility function. A utility function, denoted as u_{ij}, is defined to express the degree of fulfillment to the resource constraint(s) each BS assignment strategy aims to optimize with the assignment of user i to a given BS j. This is achieved by incorporating a penalty factor into the utility function whose value directly depends on the degree of fulfillment to the corresponding constraint(s).

The considered utility functions are monotonically decreasing and concave functions, although different forms of expressing utilities are possible [78]. In this type of functions, the absolute value of its derivative progressively increases as moving towards a condition of minimum utility. Conversely, these functions exhibit softer variations when they are close to the region of maximum utility. Details of the utility functions of the above mentioned BS assignment strategies are given in next sub-sections.

5.4.1. Minimum Path Loss

The basic idea of MPL strategy is simple. The assignment of user i is performed based on selecting the BS whose corresponding radio path loss is the minimum (hereafter referred to as best server [55]). This strategy is used as a reference case for the other two BS assignment strategies. The assignment criterion of the MPL approach is modeled by the following utility function:

$$u_{ij} = 1 - \left(\frac{\min\left(L_{ij} - L_{i,\mathrm{bs}}, \Delta\right)}{\Delta} \right)^2 \tag{5.18}$$

where $L_{i,\mathrm{bs}}$ is the attenuation of user i with respect to the best server choice, L_{ij} is the attenuation that user i would have if it is served by BS j, and Δ is the maximum accepted path loss margin with respect to the best server. Thus, it is easy to see that the maximum utility would be obtained when user i is assigned to its best server, and lower values otherwise. Notice, however, that the utility function of this strategy is unaware of BS power utilization and backhaul capacity constraints.

5.4.2. Load Balancing Radio

The LBR strategy aims to distribute terminals among BSs considering the transmitted power of BSs in the system, whenever propagation losses between terminals and candidate BSs do not exceed a given margin Δ above the minimum path loss (i.e., the one that each user would have respect to its best server). The utility function of LBR strategy is formulated as:

$$u_{ij} = \left(1 - \left(\frac{\min\left(L_{ij} - L_{i,bs}, \Delta\right)}{\Delta} \right)^2 \right) \cdot \left(1 - \left(\frac{\min(P_j, P_{\mathrm{T,max}})}{P_{\mathrm{T,max}}} \right)^2 \right) \tag{5.19}$$

Here $L_{i,bs}$ is the path loss attenuation between user i and its best server, and $L_{i,j}$ is the attenuation of that user i would have if it is served by BS j. Thus, as observed in equation (5.19), as the total power of BS j increase towards its maximum power limit $P_{\mathrm{T,max}}$, the resulting utility u_{ij} associated to the assignment of user i to BS j decreases. Hence, higher utilities are achieved when all BSs tend to use the less transmit power. Similarly, the utility is also decreased when assigning a user to a BS with a path loss approaching Δ above the minimum. The exponents in the path loss and power components are used to adjust the shape of the proposed utility function. Here we assume quadratic exponents, and thus both components have the same weight. Although this strategy is aimed to balance the power consumption at BSs in the system, it does not take into account the availability of backhaul resources. This leads us to the third strategy.

5.4.3. Joint Radio and Transport Balancing

Unlike LBR approach, for the JRT approach we incorporate transport restrictions within the utility function as follows:

$$u_{ij} = \left(1 - \left(\frac{\min\left(L_{ij} - L_{i,bs}, \Delta\right)}{\Delta}\right)^2\right) \cdot \left(1 - \left(\frac{\min(P_j, P_{\mathrm{T,max}})}{P_{\mathrm{T,max}}}\right)^2\right) \cdot \left(1 - \left(\frac{\min(R_j, C_{\mathrm{trans},j})}{C_{\mathrm{trans},j}}\right)^2\right) \quad (5.20)$$

where R_j is the aggregated rate of all users being served by BS j, and $C_{\mathrm{trans},j}$ is the associated transport capacity of the BS j. The rightmost term in equation (5.20) takes into account the transport occupancy of BS j, which implies that if a user is assigned to a BS with high transport/power utilization, the resulting utility will be lower. Although some degree of coupling exists between the BS transmitted power and the served aggregate rate, different situations can arise where one constraint could become more restrictive that the other. For instance, for the same aggregate traffic load, different power levels may be required depending on how far from the BS users are located.

5.5. Simulated Annealing Algorithm

The developed algorithm to solve the BS assignment problem that considers radio and transport network constraints is based on Simulated Annealing (SA), a popular meta-heuristic technique used in combinatorial optimization problems. A brief introduction to the SA technique and details of the implemented SA algorithm are given in this section.

5.5.1. Background

Simulated annealing [79] is an optimization technique inspired by the natural process of annealing solids. The physical process of annealing is the cooling of a metal sufficiently slowly so that it adopts a low-energy, crystalline state. When the temperature of the metal is high, the particles within the metal are able to move around, changing the structure of the metal. As the temperature is lowered, the particles are limited in the movements they can make as many movements have a high energy cost and are increasingly limited to only those configurations with lower energy than the previous state. Simulated annealing draws inspiration from the physical process, in a computational model of the physical system.

The basic simulated annealing algorithm initially defines a high temperature and then reduced to a near-zero value during the execution of the algorithm. Starting from an initial solution to the optimization problem, in each iteration of the algorithm this solution is perturbed in some manner to produce a new solution. The change of both solutions is evaluated in terms of the increase in a given utility function. When the new solution is no worse than the previous one, the new solution is accepted. On the other hand, when the new solution is of lower quality than the existing solution, it may be accepted with a probability dependent upon both the current temperature and the magnitude of the difference in the utility. The probability of accepting a worse solution, denoted as p, is a function of both the temperature and the change in the system utility, and is given by the Metropolis equation [79]:

$$p = \exp\left(\frac{-\delta}{T}\right) \quad (5.21)$$

where δ is the change of the system utility between two different solutions, and T is a control parameter, which by analogy with the original application is known as the "system temperature". Using this function, the SA approach not only accepts solutions that increase the system utility (assuming a maximization problem), but also solutions that decrease it. The behavior of the acceptance probability is shown in Figure 5.1 for two different temperature values.

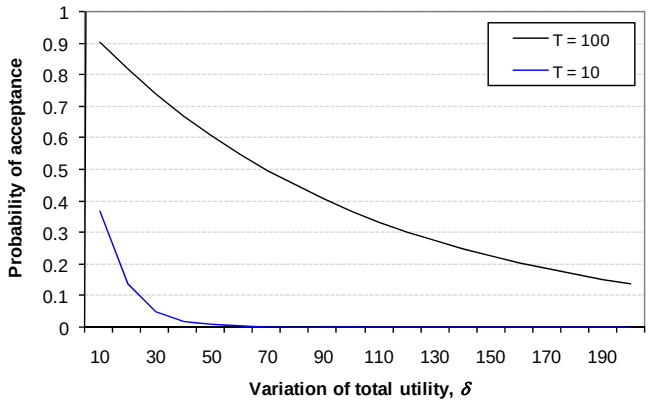


Figure 5.1: Acceptance probability of new solutions in simulated annealing.

5.5.2. Description of the Algorithm

For a given snapshot of the system, that is, number of users scattered in the service area, the algorithm aims to maximize the total system utility U, which is given by:

$$U = \sum_{i=1}^{M}\sum_{j=1}^{N} u_{ij} b_{ij} \tag{5.22}$$

While the algorithm maximizes (5.22), it is verified if a feasible BS assignment solution, where constraints (5.14) and (5.15) can be satisfied, has been found. The pseudo-code of the implemented SA algorithm is shown in Figure 5.2. The algorithm starts by defining an initial assignment solution $B=\{b_{ij}\}_{MxN}$ and an initial temperature T_0. The initial assignment solution B is obtained using the minimum path loss criterion, although different approaches are accepted. The algorithm iteratively decreases the temperature during the search of a feasible solution. For each temperature value T, the inner loop (line 04) could be performed a maximum number of iterations. The algorithm generates a new BS assignment solution B', and the system utility of both solutions is compared by means of $\delta=U(B')-U(B)$. The new assignment solution B' is accepted if it provides an increase ($\delta > 0$) of the system utility respect to the previous solution. However, as suggested by the SA approach, if the new solution leads to a worse system utility (i.e., $\delta < 0$) it is also accepted with a probability given by equation (5.21). After this process of finding a new BS assignment solution, comparing it to the current one, and either accepting or rejecting it is done *max_iter* times, the temperature changes. The rate of change for the temperature depends on the specific problem analyzed and the amount of time for which we want SA to run. The maximum number of iterations is also chosen this way. This process then is repeated until the stopping criterion is reached. Details

```
01  Initialization T_{k=0}, T_end, nbr_iter, max_iter
02  Compute initial assignment solution B, total utility U(B)
03  while T_k ≥ T_end and stopping criterion is not reached do
04      while iter ≤ max_iter and stopping criterion is not reached do
05          Compute a new BS assignment solution B'
06          Compute δ = U(B')-U(B)
07          if δ > 0 then
08              B=B'
09          else
10              if random(0,1) < e^(-δ/T) then
11                  B=B'
12          Determine if stopping criterion is met
13          iter = iter + 1
14      Compute T_{k+1} according with the annealing schedule
```

Figure 5.2: Pseudo-code of the simulated annealing algorithm.

of the parameters involved in the operation of the algorithm are discussed next.

5.5.3. Algorithm Parameters

The performance of the SA algorithm is highly dependent to the adjustment of several parameters. The main component of the SA algorithm is the cooling schedule, which is composed by four different elements: starting temperature, new BS assignment procedure, temperature decrement rule, iterations in each temperature, and a stopping criterion.

5.5.3.1. Starting temperature value.

The definition of the starting temperature T_0 plays a key role in the operation of the algorithm. If it is defined too high, the algorithm can make many movements throughout the solutions space as the probability of accepting worse solutions is high. On the other hand, if the initial temperature is set very low (i.e., acceptance probability of a worse solution is very low), the algorithm will hardly move from the initial solution, and from the beginning it could be trapped on a local optimum.

The starting temperature considered in our implementation is set to T_0=99. In this way (at high temperatures) most of the solutions visited by the algorithm are accepted, and therefore it avoids to be trapped into local minima. In fact, this is one of the main advantages provided by SA as it offers a way of escaping from local minima by means a mechanism of "uphill" move that relies on the decision rule given in equation (5.21).

5.5.3.2. Temperature decrement rule.

Once the starting temperature is defined, a rule to iteratively decrease it as the search progresses should be selected. We consider an exponential annealing schedule [80] where the temperature value for step k+1 is computed according to:

$$T_{k+1} = \varepsilon \cdot T_k \tag{5.23}$$

where ε is a constant used to decrease the temperature, whose value is close to, but smaller than, 1. Different studies have shown that suitable values of ε ranges from 0.8 to 0.99 [81], with better results being found in the higher end of the range. However, the higher the value of ε, the longer it will be the decrementing rate of the temperature in the algorithm. In our implementation we assume a value of ε=0.8.

5.5.3.3. Number of iterations.

The theory of SA states that enough number of iterations at each temperature should be performed so that the system has reached the steady state. In this sense, the steady state can be assumed to be reached when the difference between two consecutive values of the system utility (as computed in line 06 of Figure 5.2) is not higher than a defined margin. Unfortunately, the number of iterations that would be performed at each temperature to achieve this might be exponential to the problem size, thus leading to unacceptable simulation time. Alternatively, a maximum number of iterations to be performed at each temperature can be assumed. This latter approach is the one followed in our algorithm.

5.5.3.4. Stopping criterion.

The algorithm iteratively decreases the temperature so that eventually the stopping criterion can be met. In particular, the search of the algorithm can be halted when a BS assignment solution satisfying radio and transport constraints of each BS in the system is found (referred to as feasible solution). To this end, the algorithm verifies if the new solution found satisfies both radio and backhaul constraints of each BS in the system (line 12 in Figure 5.2). Furthermore, in order to

consider a case where no feasible solution can be found, a final temperature value is also included as a stopping criterion in the algorithm. The final temperature considered in the implementation of the simulated annealing algorithm is 0.374 [81].

5.5.3.5. New solution generation

The new assignment solution *B'* generated (line 05) in the SA algorithm is computed using the following approach. For each user it is estimated the utility increment that it would have respect to each BS contained on its candidate set. The utility increment of a user is defined as the difference between the utility obtained on its current assignment and the utility that this user would achieve if is moved to a new BS. Then, the user and BS which provides the highest utility increment is considered for reassignment, thereby generating a new BS assignment configuration. With the new BS assignment configuration, the power consumption for each user is computed again by means of equation (5.2). Afterwards, the utility of each user and consequently the total system utility are updated accordingly.

The described approach to generate a new solution BS assignment solution constitutes a good tradeoff between avoiding local minima and reducing the overall number of iterations, in comparison to randomly generate a new BS assignment solution.

5.6. Performance Evaluation

This section provides the performance analysis of the BS assignment problem in WCDMA mobile systems. The first step carried out in our analysis is the estimation of the air interface downlink capacity, followed by simulation results of the formulated BS assignment strategies under different network configurations. As argued in the system modeling of section 5.3, the evaluation is done focusing on the downlink performance as it is assumed to be the most restrictive link due to the asymmetric bandwidth demand between uplink and downlink (i.e., the required uplink bit data rate is lower than the downlink bit data rate).

5.6.1. Estimation of Downlink Capacity

As discussed in section 5.3.2, the capacity required in the transport network is modeled in accordance to the amount of traffic that can be supported over the air interface (as captured in equation (5.11)). Assuming a single cell scenario with uniform user distribution the air interface capacity is computed by means of equation (5.6) with parameters as provided in Table 5.1 [55]. The air interface capacity is presented in Table 5.2 for different values of the maximum cell path loss and considering voice and data services.

Table 5.1: Common values for downlink dimensioning.

Parameter	**Value**			
	Voice	*Data*		
Service bit rate, R	12.2 Kbps	64 Kbps	128 Kbps	384 Kbps
Eb/N_0 target	6.7 dB	5.3 dB	5.3 dB	5.2 dB
Propagation model	L(dB)=128.1+37.6log[d(km)]			
Chip rate, W	3.84 Mchips/s			
BS max. power, $P_{T,max}$	43 dBm			
Noise power, P_N	-101.15 dBm			
Other-to-own interference, f_{DL}	0.65			
Average orthogonality factor, α	0.5			
Maximum versus average path loss	6 dB (assuming uniform user distribution)			

Table 5.2: Air interface capacity versus the maximum cell path loss.

Maximum Path Loss	Voice 12.2 Kbps	Data 64 Kbps	Data 128 Kbps	Data 384 Kbps
L^{max}=120 dB	707 Kbps	960 Kbps	1024 Kbps	1152 Kbps
L^{max}=130 dB	707 Kbps	960 Kbps	1024 Kbps	1152 Kbps
L^{max}=140 dB	658 Kbps	896 Kbps	896 Kbps	768 Kbps
L^{max}=150 dB	390 Kbps	512 Kbps	512 Kbps	384 Kbps

5.6.1.1. Impact of Users' Position on Air Interface Capacity

The downlink capacity is highly dependent on the multipath propagation environment (e.g., orthogonality between intra-cell transmissions is penalized in richer multipath environments), mobile terminals' speed (different E_b/N_0 values can be required) and user position (affecting the path loss and the ratio of inter/intra-cell interference). Thus, in a given operational period of a CDMA cell, the air capacity offered can differ significantly from the dimensioned capacity.

In this context, this section aims to examine the variation of the air interface capacity due to the distribution of users within the cell. In particular, the capacity variation when having all traffic concentrated in a hot spot is compared to the capacity estimated assuming a uniform spatial distribution. The location of the hot spot is illustrated in Figure 5.3. Its position is parameterized by means of $\Delta L=L^{HS}-L^{max}$ that defines the difference between the path loss attenuation in the hot spot location, denoted as L^{HS}, and the maximum path loss attenuation in the cell edge, denoted as L^{max}.

To calculate the downlink capacity for a hot spot with L^{HS} path loss attenuation, equation (5.6) can be used by substituting the average path loss L_{avg} per L^{HS}. Moreover, instead of using an averaged inter-to-intra cell interference factor f_{DL} in equation (5.6), this value is calculated for a specific hot spot position.

Notice that the value of averaged inter-to-intra cell interference factor plays an important role in the estimation of the pole capacity. In this sense, considering the propagation model given in Table 5.1, and assuming an hexagonal macro-cell deployment with cell radius given to satisfy the maximum path loss L^{max}=140 dB and no shadowing variation, the value of the f_{DL} factor in the straight line joining the center of the cell station with the cell vertex is shown in Figure 5.4.

The relative and absolute capacity variation when considering a hot spot in front of a uniform user distribution versus the location of the hot spot within the cell is shown in Figure 5.5 and Figure 5.6, respectively. A maximum path loss of L^{max}=140dB is considered, along with values of the dimensioning parameters given in Table 5.1. As it is observed in these figures the closer the location of the hot spot from the center of the cell, the higher the capacity that it can be achieved with respect to the one obtained with an uniform distribution of users.

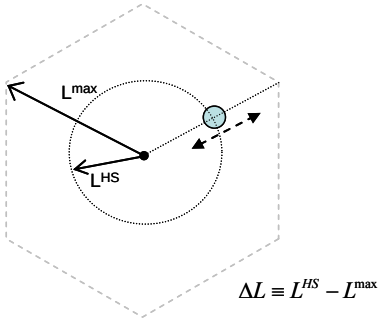


Figure 5.3: Hot-spot location in a cell.

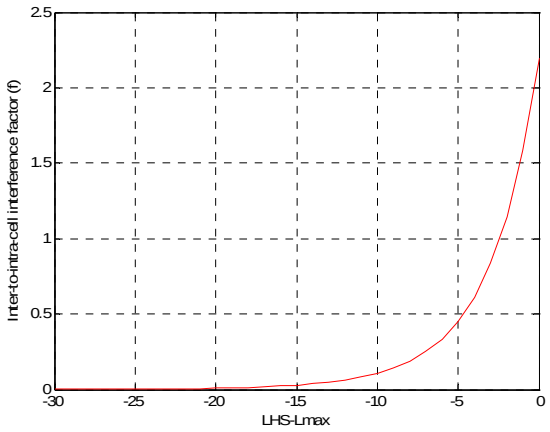


Figure 5.4: Value of the f_{DL} factor in the straight line joining the center of the cell with the cell vertex.

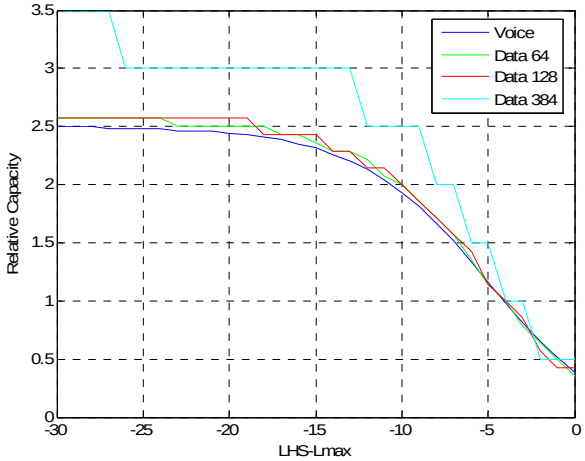


Figure 5.5: Relative air interface capacity of a hot spot in front of a uniform user distribution versus the location of the hot spot.

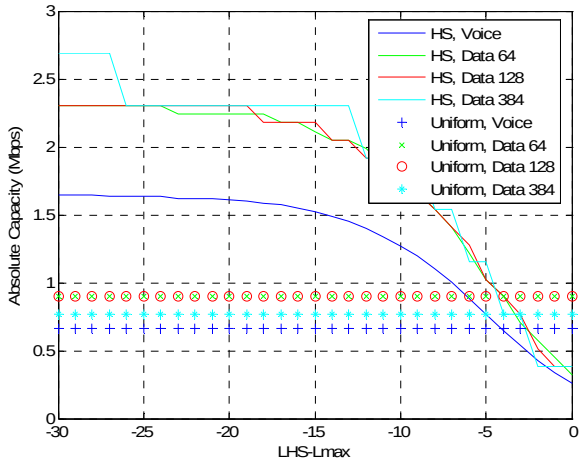


Figure 5.6: Absolute air interface capacity of a hot spot and uniform user distribution versus the location of the hot spot.

5.6.2. Simulation Results

The BS assignment strategies described in section 5.4 are evaluated considering a cellular deployment with 19 hexagonal cells. In order to prevent the so-called border effect (users and cells in the boundaries of the system behave different than others inside the cellular layout, because there

is no interference received beyond the boundaries), a wrap-around technique is used [82]. The channel is characterized by the propagation model introduced in Table 5.3 which follows 3GPP specifications. After the path loss is calculated, log-normally distributed shadowing with 0 dB mean and standard deviation of 10 dB is added. The maximum accepted path loss margin, denoted as Δ, is assumed to be 6 dB. We do not consider macro-diversity in the performance evaluation, so that each user in the system is assumed that is served by a single BS. Furthermore, the analysis is focused on delay sensitive data services where it is particularly important to guarantee a given service rate at both the air interface and transport network segments.

Over such a basis, in simulations we consider single service type scenarios, where each user in the system requests a data service with a given bit rate. Specifically, three different data bit rates are considered: 64 Kbps, 128 Kbps, and 384 Kbps. As detailed in previous section, depending on the data service rate of users under consideration, the downlink air interface pole capacity, referred to as $C_{\text{air,p}}$, is determined assuming an average ratio of other-to-own interference f_{DL}=0.65, and average orthogonality factor α=0.5 [55]. Then, from the air interface pole capacity, the transport capacity of BSs in the system is computed according to equation (5.11). Table 5.3 shows the downlink data bit rates considered in simulations, along with their corresponding E_b/N_0 requirements and pole capacity values. The propagation model is again summarized in this table.

In this context, for a given snapshot of the system (a number of users distributed over the service area), a BS assignment solution is computed using each of the considered BS assignment strategies defined in Section 5.4. The MPL strategy is easily implemented by directly selecting for each user the BS where the maximum utility (i.e., lower path loss value) is obtained. The other two strategies (LBR and JRT) are implemented by means of the defined simulated annealing algorithm, using their own utility function. The provided solution of each strategy is analyzed in order to verify if it fulfills the corresponding radio and transport constraints of each BS in the system. Over such a basis, by performing a large number of independent snapshots, the percentage of feasible solutions is obtained for each algorithm.

Simulation results are presented in terms of the maximum number of supported users when considering a given transport network capacity. Recall that the transport capacity is modeled in terms of the amount of extra capacity needed in the transport network (respect to the air interface pole capacity) to meet a given network availability (e.g., a feasible BS assignment solution can be found in 95% of the performed snapshots). In particular, the extra capacity is modeled by means of the dimensioning factor, denoted as β, introduced in Section 5.3.2, and captured in equation (5.11). At this regard, dimensioning factor values ranging from 1 up to 2.5 have been considered in simulations, implying that the transport capacity of a given BS is equal to the air interface pole capacity (i.e., β=1), or 150% higher than its air interface pole capacity, respectively. The evaluated scenarios are classified as follows:

- Homogeneous transport capacity. The transport capacity of BSs in the system is the same, and the aim is to evaluate the impact of the provisioned transport capacity on the maximum number of supported users in the system in order to provide a given network availability. In this scenario, two different cases are considered: (a) users are uniformly distributed over the service area, (b) non-uniform distribution of users.

Table 5.3: Downlink pole capacity values and CDMA simulation parameters.

Parameter	**Value**		
Downlink data service bit rate, R	64 Kbps	128 Kbps	384 Kbps
Eb/N_0 target	5.3 dB	5.3 dB	5.2 dB
Pole capacity, $C_{\text{air,p}}$	960 Kbps	1024 Kbps	1152 Kbps
Cell radius	1 Km		
Propagation model	L(dB)=128.1+37.6log[d(km)]+S(dB)		
Shadowing standard deviation, S	10 dB		

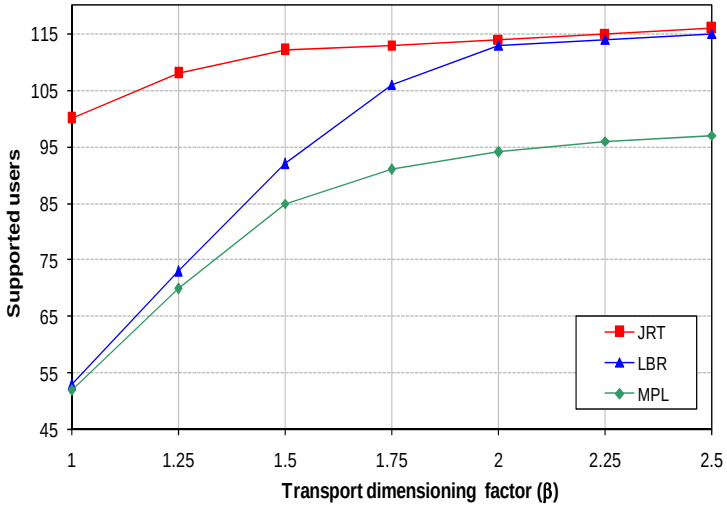


Figure 5.7: Supported users with data service rate of 128 Kbps.

- Partially limited transport network capacity. In this scenario, a number of BSs are assumed to have transport capacity limitations, while the transport of the rest of BSs are assumed to have sufficient transport network so that they do not constitute a bottleneck.

5.6.2.1. Homogeneous transport capacity and users uniformly distributed.

The behavior of the BS assignment strategies is firstly analyzed assuming that all BSs in the system have the same transport network capacity. Users in the system are uniformly distributed. Simulation results in Figure 5.7, Figure 5.8, and Figure 5.9, presents the maximum number of active users in the overall service area for which the BS assignment strategies are able to find a feasible BS assignment solution in 95% of the performed snapshots, when the data service rates of users is 128 Kbps, 64 Kbps, and 384 Kbps, respectively. In these figures, the x-axis represents the transport capacity, expressed in terms of the dimensioning factor, of the BSs in the system.

Simulation results in Figure 5.7 correspond to the case where users request to be served with a data service rate of 128 Kbps, and also meet the E_b/N_0 target given in Table 5.3. As shown in this figure, the classical MPL assignment strategy achieves the poorest performance as it supports lower number of users in the system (in order to obtain a feasible solution) than the other two strategies. On the other hand, the LBR assignment strategy achieves almost the same performance only in scenarios where the transport capacity of BSs is around two times the air interface pole capacity (i.e., $\beta \geq 2$). However, as the available transport capacity in BSs is decreased its performance is considerably deteriorated since it is only able to find a feasible solution if it serves a lower number of users in the system.

On the contrary, the proposed JRT strategy that incorporates transport status in the assignment process is able to increase the number of supported users about 88% with respect to the MPL and LBR assignment strategies, in scenarios where the available transport capacity in the BSs corresponds to the air interface pole capacity obtained in the planning process (see β=1 in Figure 5.7). Furthermore, the benefits of using JRT are also reflected in bandwidth savings, since with less provisioned transport capacity is capable to obtain similar results than non transport-aware strategies such as LBR and MPL. As expected, the LBR and JRT strategies tend to converge as the transport capacity increases, implying that the system is only limited by air interface resources as BSs have enough transport capacity and thus bottleneck situation due to transport shortage do not exist. In any case, both LBR and JRT always are able to achieve some capacity gain over the classical MPL strategy in all the analyzed cases.

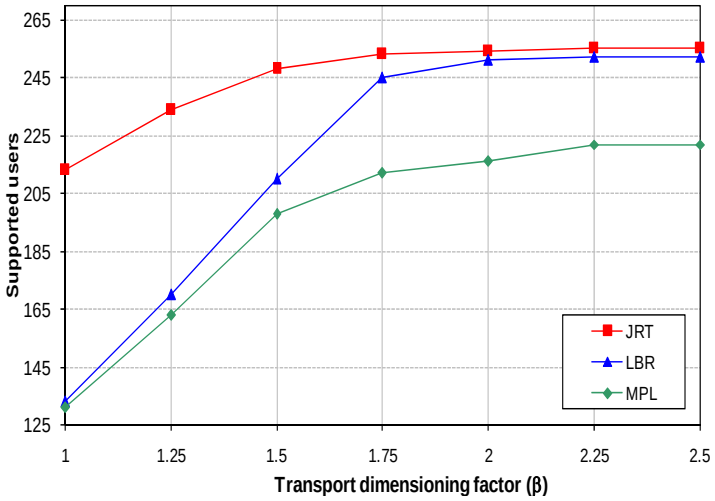


Figure 5.8: Supported users with data service rate of 64 Kbps.

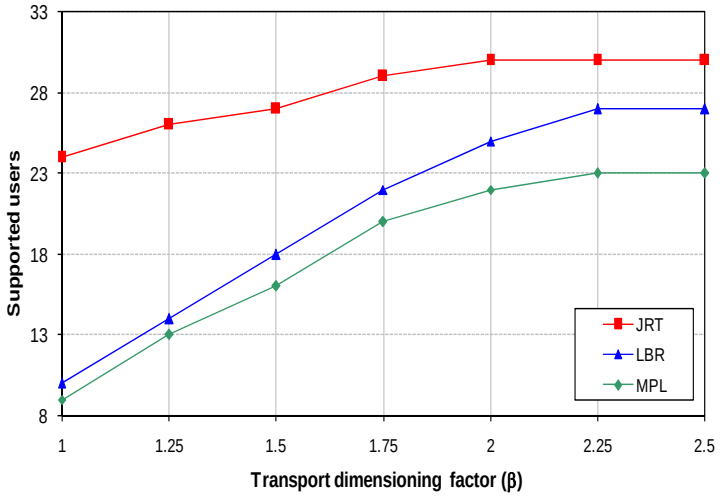


Figure 5.9: Supported users with data service rate of 384 Kbps.

When considering lower data rates such as 64 Kbps (see Figure 5.8), the same trends discussed for 128 Kbps are observed but the obtained gains are lower (e.g., 60 % in front of 88 % for β = 1). The main reason for this decreased gain is that user distribution among BSs becomes more homogenous as the number of active users to be assigned is higher and, consequently, load balancing strategies have less room to improve.

Finally, in Figure 5.9 the maximum number of supported users is shown when considering a data service requirement of users of 384 Kbps. In this case, the gains provided by LBR and JRT are now even higher due to the fact that fewer users can be served per BS and consequently traffic distribution mechanisms between BSs becomes more imperative. It can be seen that JRT can achieve gains of 140 % and 50 % with respect to the LBR strategy for β=1 and β=1.5, respectively.

5.6.2.2. Homogeneous transport capacity and non uniform users' distribution.

In this section we analyze the BS assignment strategies under non-uniform distribution of users in the system. This is characterized assuming the presence of traffic hot spots in the overall service area. A hot spot region could appear due to the existence of a shopping mall or large office which might be 100–200 m across in a macro-cell whose cell radius is a couple of kilometers. In this case,

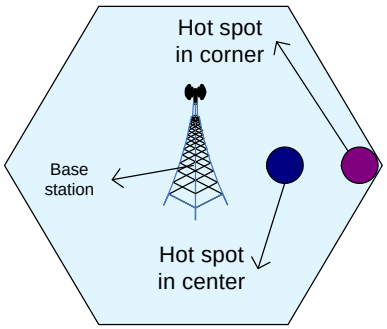


Figure 5.10: Hot spot located in the corner or in the center of the cell.

we assume circular hot spots which could be located in two possible positions within the cell, as illustrated in Figure 5.10. Specifically, in a given hexagonal cell the hot spot can be located in a corner or in center.

Under this analysis, the hot spots are randomly located within the service area of N_{HS} cells, while in the rest of the cells the users are uniformly distributed. Simulation results for this case are depicted in Figure 5.11. We assume that there exists N_{HS}=5 cells with a hot spot region, and that the requested data rate of users in the system is 384 Kbps.

As shown in Figure 5.11, irrespective of the location of the hot spot, the JRT strategy is able to allocate more active users under transport dimensioning factors of $\beta \leq 2$, which is achieved by exploiting the load balancing of radio and transport resources. It is in general observed the presence of the hot spot in the center leads to a slight increase in the number of supported users in the three BS assignment strategies in front of a situation where the hot spot is on located in the corner of the cell.

5.6.2.3. Partially limited transport network

In the previous sections, we have analyzed the performance BS assignment strategies assuming that all the BSs in the system have the same transport capacity. In this section, we assume a partially limited transport network, so that BSs do not necessarily are provisioned with the same transport capacity. To this end, we again consider single service scenarios, where users request to be served with a given data transmission rate in the downlink. Simulation results are shown in

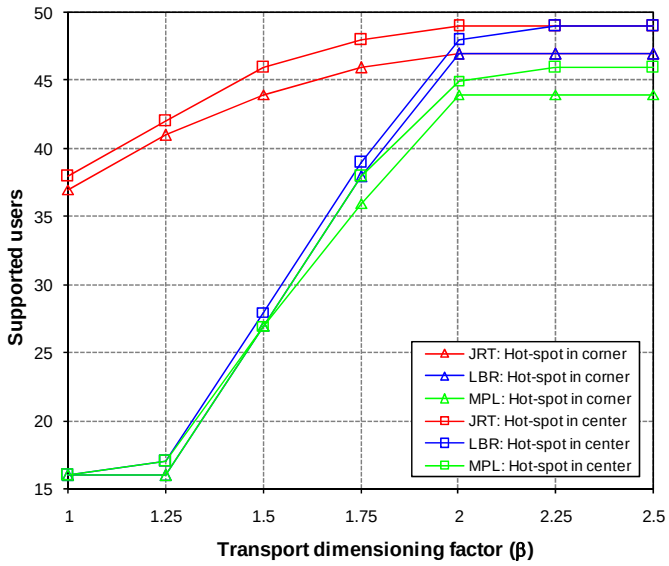


Figure 5.11: Supported users versus transport capacity when N_{HS} cells have a hot spot region.

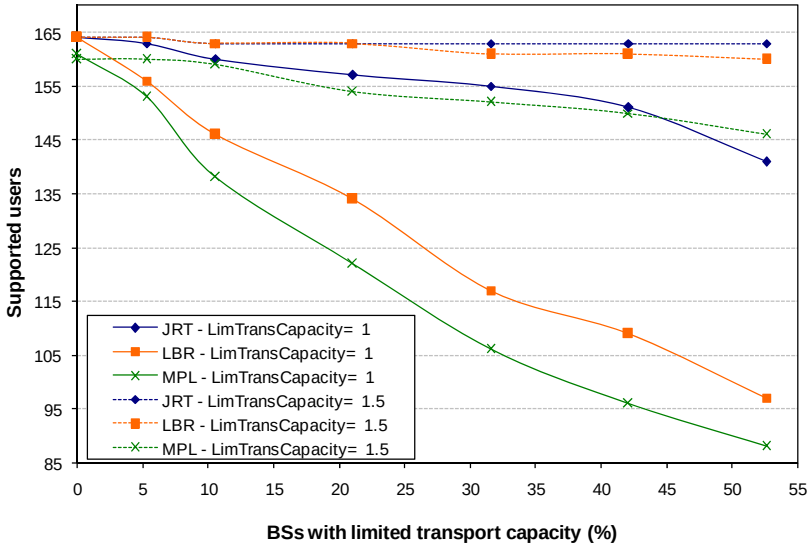


Figure 5.12: Supported users with service rate 128 Kbps in partially limited scenarios (β_{lim}=1 and β_{lim}=1.5). The rest of BSs have unlimited transport capacity (β_{unlim}=3).

Figure 5.12 and Figure 5.13, for the cases where the data service rate of users is 128 Kbps or 384 Kbps, respectively. In these figures, the x-axis shows the percentage of BSs that have a limited transport capacity. To this end, two different limited capacities are modeled by means of a transport dimensioning factor β={1,1.5}, that is the ratio of the transport capacity to the air interface capacity.

The rest of the BSs in the system are assumed to have enough transport capacity (i.e., β=2.5).

Focusing on Figure 5.12, it can be seen that the proposed JRT that is aware of the transport status outperforms the benchmark strategies since it is able to allocate more users that can satisfy both radio and transport constraints. It provides a gain of about 12 % with respect to the RBL strategy in a scenario where around 10 % of BSs have limited transport capacity (β= 1). On the other hand, for a limited transport capacity value equivalent to β = 1.5 it still obtaining gains with respect to radio-based strategy. This is because the JRT strategy takes advantage of the available transport capacity and tries to distribute users among BSs in order to balance the usage of the transport network. It can be seen also that a common assignment approach such as MPL provides lower performance than the JRT and LBR strategies.

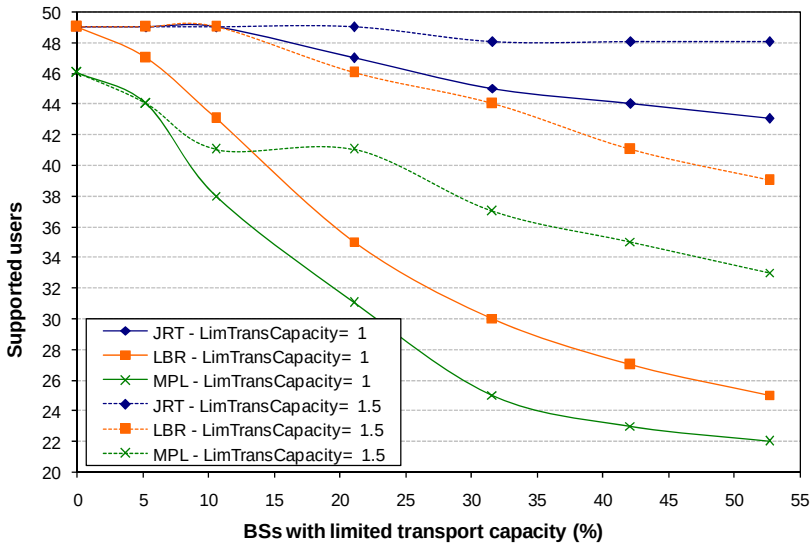


Figure 5.13: Supported users with service rate 384 Kbps respect to percentage of BSs with limited backhaul (β_{lim}=1 and β_{lim}=1.5). The rest of BSs have unlimited transport capacity (β_{unlim}=3).

Table 5.4: CDMA uplink simulation parameters.

Parameter	Value
Uplink bit data rate	128 Kbps
Eb/N_0 target	3.2 dB
Chip rate, W	3.84 Mchips/s
Max. mobile transmission power	21 dBm
Noise power, P_N	-100.15 dBm

When considering higher data rates such as 384 Kbps (see Figure 5.13) the strategies behaves similar that in the previous case, but the obtained gains increase. For instance, the gain achieved by the backhaul-aware strategy is about 14 % when compared to radio-based strategy in a scenario with 10 % of the BSs with limited capacity (β = 1).

5.6.2.4. Impact on downlink and uplink power consumption

The performance gains that the proposed JRT strategy can achieve over LBR and MPL come at the expenses of an increased consumption of radio resources. In the following we evaluate the impact of the proposed JRT strategy on the resulting downlink and uplink power consumption.

We consider a scenario where users in the system are assumed to request to be served with a data rate of 384 Kbps in the downlink. In the uplink direction, the data requirement is assumed to be of 128 Kbps. The transport capacity of BSs is modeled assuming a transport dimensioning factor of β=1.5, indicating that the resulting capacity of BSs is 50% higher than the air interface pole capacity. As seen in previous results this transport capacity value could constitute a bottleneck situation to the transport network. It is also considered a transport dimensioning factor of β=2 (i.e., transport capacity of BSs is 100% higher than the air interface capacity) to illustrate the power consumption under less restrictive transport network conditions. The parameters shown in Table 5.4 are considered in the computation of the uplink transmission power by means an iterative procedure.

Figure 5.14 shows the normalized transmission power in the downlink for the two different transport capacity values. In can be seen in this figure that the downlink power consumption is only increased, with respect to the two reference strategies, when the transport network constitutes the main capacity limitation (i.e., BSs are provisioned with a transport capacity corresponding to

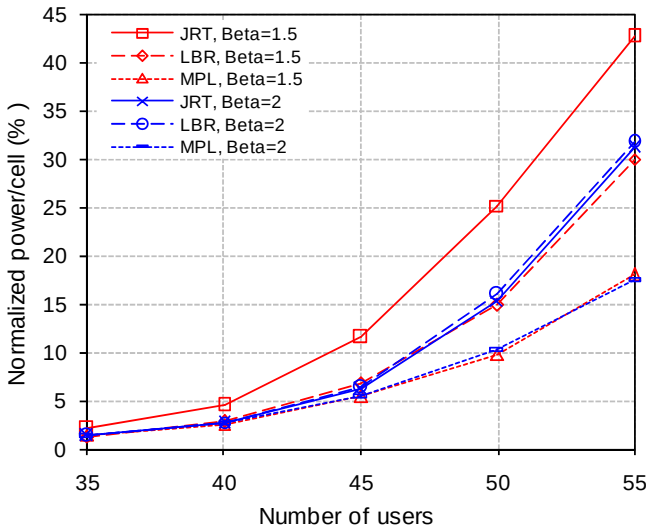


Figure 5.14: Power consumption in downlink as function of active users.

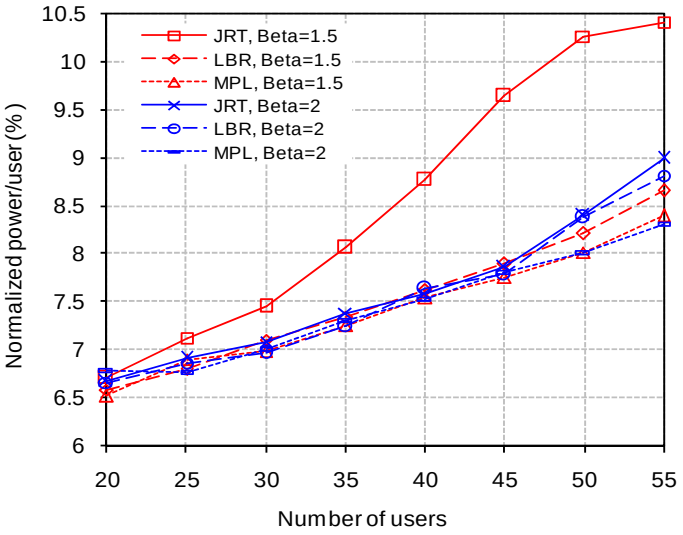


Figure 5.15: Power consumption in uplink as function of active users.

β=1.5). This is because under such situations, the JRT allow that some users can be assigned BSs that are not the best choice from the radio point of view so that the data rate requested by uses could be satisfied and also to prevent congestion situations in the transport network due to the assignment of users to BSs with overloaded transport network. On the other hand, the uplink power consumption (see Figure 5.15) is more critical. It is seen that in the case of β=1.5 where a capacity gain around 55% can be achieved by JRT over LBR and MPL (i.e., 47 active users can be allocated by JRT with a 95% satisfaction level in front of 30 users with the other strategies) uplink power increase is kept below 1.5 dB.

5.7. Summary

This chapter analyzed the BS assignment problem in the context of WCDMA based networks. Besides radio related resource constraints, the formulated problem also accounts for the available capacity at each BS in the transport network. In order to solve the formulated problem a simulated annealing-based algorithm has been developed. Using this algorithm we evaluate two BS assignment strategies that perform user assignments based on radio aspects only, and also a third strategy that besides radio criteria it also incorporates transport constraints within the assignment process.

6 *Cost-based BS Assignment for OFDMA-based Mobile Broadband Networks*

6.1. Introduction

During the last years wireless access technologies have been evolving rapidly to provide higher data rates and pave the way for ubiquitous, high speed broadband wireless coverage. Nowadays, the most outstanding radio technologies to meet such requirements are based on orthogonal frequency division multiplexing (OFDM) schemes that have been successfully adopted for next generation cellular systems such as Long Term Evolution (LTE) and Mobile WiMAX. The usage of an OFDM physical layer enables orthogonal frequency division multiple access (OFDMA) that allows for the exploitation of multiuser diversity by managing both time and frequency components in the radio resource allocation process. In an OFDMA-based cellular system, efficient radio resource allocation techniques are needed to fully exploit OFDMA capabilities [83]. So far, algorithmic solutions to resource allocation problems in OFDMA-based systems are usually developed under the assumption that the capacity bottleneck is always on the air interface. However, as argued in Chapter 1, this assumption proved to be valid for voice-dominant cellular networks where the relatively low data rates achieved at the air interface do not impose stringent capacity requirements to the backhaul.

In this chapter we aim to investigate the impact of possible backhaul resource limitations on the resource allocation decisions in an OFDMA-based broadband communication system. We firstly prove that traditional base station (BS) assignment schemes exclusively based on radio criteria (i.e., minimum path loss and radio load) can fail to find a proper assignment without violating quality of service (QoS) requirements when backhaul can become the bottleneck of the cellular network, even though sufficient network capacity is shown to be available. Based on this proof of concept, we tackle the design of a novel BS assignment algorithm aimed to cope with eventual backhaul capacity limitations in OFDMA-based cellular systems. The performance of the developed algorithm is evaluated and compared to other baseline algorithms. The main idea behind the proposed algorithm is to distribute traffic among BSs according to a load balancing strategy that considers both radio and backhaul load status. This possibility is shown to constitute a tradeoff between reducing backhaul congestion and using radio resources efficiently since some users can be assigned to BSs not being their "best" radio choice but preventing congestion in other BSs. The proposed algorithm is proven to successfully exploit such a tradeoff, turning ultimately into a better overall network performance.

The rest of the chapter is structured as follows. A review of the BS assignment problem in OFDMA systems is presented in Section 6.2. Then, in Section 6.3 details of the system model are provided. In Section 6.4 we formulate a new BS assignment problem that considers radio and

backhaul constraints in the assignment process. The formulated optimization problem is based on utility and resource cost concepts. This section also presents the mapping, after some practical considerations, of the BS assignment problem into a Multiple-Choice Multidimensional Knapsack Problem (MMKP), a well known combinatorial optimization problem arisen in many practical and real life problems. Then, Section 6.5 presents the derivation of a heuristic algorithm to solve the formulated optimization problem. Section 6.6 provides a preliminary evaluation of the BS assignment problem in a simple scenario, and then defines the evaluated BS assignment algorithms and the performance evaluation methodology. Finally, in Section 6.7 the performance of the proposed algorithm is compared respect to classical schemes entirely based on radio conditions, whereas in Section 6.8 summarizes the conclusions of the chapter.

6.2. Related Work

The resource allocation process in OFDMA networks is a complex task that should take care of different aspects covering from the power and rate allocation per user in each individual subcarrier up to the assignment of the serving BS. A common approach followed in the literature to tackle the resource allocation process in OFDMA-based networks consists of decomposing it into several less-complex problems, each to be able to find near-optimal solutions in a computationally efficient manner. In this way, some works [84], [85] follow a three-step approach where the first resource allocation problem to be solved consists in selecting the most appropriate BS to handle radio transmission to/from mobile terminals in the system. This problem is referred to in the open literature as the *BS assignment problem* and constitutes a key component of the overall resource allocation process. Once the BS is decided, next problem deals with the determination of the number of subcarriers to be assigned to each user. Finally, the third problem aims to determining the amount of power allocated to each user in each assigned subcarrier, as well as the selection of the modulation and coding scheme (MCS).

The BS assignment problem in cellular systems has deserved significant research efforts. As pointed out in previous chapter, and irrespective of the multiple access technology, one of the most common BS assignment approaches is the minimum path loss (MPL) that assigns each user to the BS that provides the highest radio link gain [86]. In fact, this approach alone constitutes the core of many BS assignment algorithms used in current 2G/3G cellular systems (where absolute and/or relative received signal level thresholds are used to decide upon the serving BS) and also forms part of more sophisticated approaches in order to exploit, e.g., multiuser detection and multiple antennas [87]. The main disadvantage of the MPL approach comes from the fact that interference and radio load conditions are not considered in the assignment process. Another common approach takes into account the Signal to Interference and Noise Ratio (SINR) in the assignment process, which is particularly important when targeting an aggressive reuse of frequencies throughout the network (e.g., single-channel CDMA networks and OFDMA networks with low reuse factors). In this case, there is a mutual dependency between the SINR values and the BS assignment in the downlink. Notice that in the downlink, moving a user from one BS to another so that the power allocated in the new BS is less than the power required in the previous one, does not necessarily contribute to reduce the interference level to all the users, since users close to the new BS are now even going to see a higher level of interference. So, in the downlink case, a mutual dependency exists between the SINR values and the BS assignment because of co-channel interference (CCI) that further complicates the resource allocation problem.

The aforementioned BS assignment approaches do not explicitly consider capacity constraints in the BS assignment process, so connections would be dropped or their quality deteriorated when assigned to a congested BS. This suggests that, from an overall resource allocation standpoint, these approaches may not lead to the most efficient assignment solution, stressing the fact that potential resource constraints have to be considered within the assignment problem itself. In addition to channel gain and SINR, different constraints such as maximum transmission powers or

minimum guaranteed rates have been also considered under various forms of optimization problems [85], [75], [88]. In this sense, a suboptimal based on SINR conditions and BS downlink radio capacity is investigated in [85]. In order to tackle the mutual dependency between SINR values and the assignments, the BS assignment algorithm devised in [85] sequentially chooses the user with the highest SINR. Specifically, in each step of the algorithm, the user that can be served with the highest SINR is selected and assigned to the correspondent BS. The computation of the SINR is obtained in each step attending to the allocation of users already fixed in previous steps. If the total users' data rate demand (defined as the sum of the traffic of current user and the already assigned users) in the selected BS exceeds its maximum radio capacity, the BS that provides the second highest SINR is tried. This is done for a maximum number of BSs, so that the user would not be served in case that none of them have enough capacity.

More recently [88] analyze the BS assignment problem considering also the case of multiple BS assignments. Irrespective of allowing a single or multiple BS assignments, the core BS assignment problem is modeled as an optimization problem aimed at guaranteeing a minimum rate requirement for each user. Authors in [88] state that the resulting BS assignment problem is NP-hard and propose two algorithms to achieve near-optimal solutions. In any case, it is worth noting that resource consumption modeling in [88] is strictly based on the consideration of which portion of the overall BS capacity a user should receive, and does not consider the mapping of this portion of the BS capacity into the ultimate BS power or time resources that are going to be used and that will definitively impact on the BS assignment. An iterative BS assignment scheme aimed to balance traffic densities is developed in [89], where the assignment decision is based on the MPL criterion and QoS requirements of users. However, most of the works tackling resource allocation in multi-cell OFDMA, e.g., [90], [91], [92], [93], implicitly consider a BS assignment based on a simple MPL criterion, due that these works mainly concentrate on developing algorithmic solutions to the subcarrier and power allocation problems.

6.3. System Model

The BS assignment problem is analyzed attending to the downlink performance of an OFDMA-based cellular network. We focuses on the downlink as it is normally seen as the most restrictive link due to the asymmetric bandwidth demand between uplink and downlink in current networks. This assumption is aligned to most works on BS assignment for OFDMA network found in the literature. The considered system consist of N BSs that cover a geographical area in which there are M active users, as illustrated in Figure 6.1. Each user $i \in \{1,\ldots,M\}$ is assumed to have a minimum bit data rate requirement, denoted as $R_i^{\min}$. The overall network uses a total bandwidth BW divided into K OFDM subcarriers, so that each BS $j \in \{1,\ldots,N\}$ operates a subset of K_j subcarriers attending to a given frequency reuse pattern. The radio and transport resources are assumed to be allocated to each user in a single BS (i.e.., macro-diversity is not considered). Particularly, soft-handover has not been considered mainly because, unlike 3G systems based on CDMA where macro-diversity is a fundamental aspect, future cellular systems based on OFDMA such as LTE do not rely on the exploitation of soft-handover capabilities [94]. It is worth noting, however, that in the context of LTE-Advanced [95] it is being discussed the use of coordinated multi-point transmission and reception (CoMP) [96] in order to improve coverage, and cell-edge throughput by means of tight coordination between the transmissions at different cell sites. Each BS in the system is assumed to be constrained by a limited amount of radio and transport resources. As to radio resources, each BS j is able to allocate simultaneously a maximum of K_j subcarriers and has a maximum transmit power of $P_j^{\max}$. The radio channel gain between BS j and user i is modeled by $\vec{G}_{ij}=\{G_{i,j,1},\ldots,G_{i,j,k}\}$, where $G_{i,j,k}$ denotes the channel gain over subcarrier $k \in \{1,\ldots,K_j\}$. As to transport resources, we assume each BS j has a maximum transport capacity C_j^{trans} (in bits/sec), which refers to the available bandwidth in the path between BS j and the access gateway (aGW) within the mobile

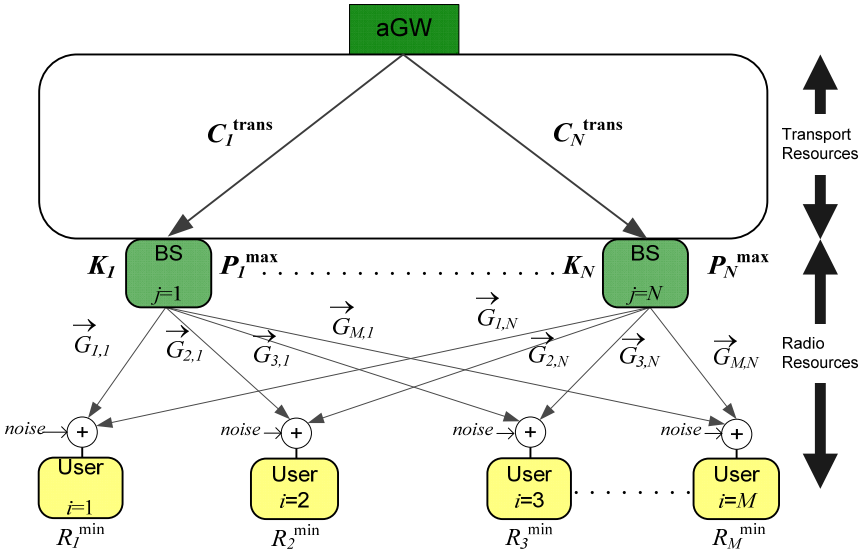


Figure 6.1: System model.

network. Here, the aGW would correspond to the ASN_GW network entity in Mobile WiMAX, or to the Serving Gateway in LTE network architecture.

The system model depicted in Figure 6.1 has not been particularized to any specific backhaul transmission solution. As discussed in Chapter 1, different alternatives for the transport network infrastructure are possible (e.g., microwave radio links, E1 leased lines). Thus, we keep as generic as possible the assumptions regarding the transport network, so that the backhaul capacity assumed here could correspond, for instance, to the available bandwidth of the wired/wireless link used for the last mile connection of a given BS.

The amount of network resources required by user i to meet its rate requirement $R_i^{\min}$ depends on the selected BSs. In order to quantify the resource consumption of user i when assigned to BS j, we define a radio resource cost and a transport resource cost function, denoted as α_{ij} and β_{ij}, respectively. Over such a basis, the BS assignment problem should try to find a feasible assignment (i.e., $R_i^{\min}$ is satisfied for each user i and total radio and transport resources are not exceeded). As well, if several feasible solutions exist (i.e., there are several ways to allocate all the users without exceeding network resources), we are also interested in finding the "best" of these possible solutions. For this reason, a utility function is used to quantify the appropriateness of each assignment in terms of bit rate efficiency of the allocated resources. Details of utility and resource cost functions are given in next subsections.

6.3.1. Radio Resource Cost Function

In a cellular OFDMA system, the computation of the SINR achieved at subcarrier k in the receiver of user i served by BS j, is obtained as follows [85]:

$$\mathrm{SINR}_{i,j,k} = \frac{G_{i,j,k} P_{i,j,k}}{I_{i,j,k} + \eta} \tag{6.1}$$

where $G_{i,j,k}$ is the radio channel gain between BS j and user i over subcarrier k, $P_{i,j,k}$ is the transmit power of BS j on subcarrier k allocated to user i, η is subcarrier thermal noise, and $I_{i,j,k}$ is the co-channel interference power received by user i in that subcarrier. The value of the co-channel interference $I_{i,j,k}$ can be computed as:

$$I_{i,j,k} = \sum_{n=1, n\neq j}^{n=N} G_{i,n,k} P_{m\neq i,n,k} \tag{6.2}$$

where $P_{m,n,k}$ is the transmit power of interfering BS n, on subcarrier k assigned to other user $m \neq i$. Equation (6.1) denotes the channel frequency response of user i on subcarrier k, and the achievable transmission rate $r_{i,j,k}$ on this subcarrier of user i assigned to BS j is given by:

$$r_{i,j,k} = \frac{BW}{K} \cdot \log_2\left(1 + \mathrm{SINR}_{i,j,k}\right) \tag{6.3}$$

Hence, if all resources of BS j were allocated to user i, the maximum achievable rate would be:

$$R_{i,j}^{\max} = \sum_{k=1}^{K_j} r_{i,j,k} \tag{6.4}$$

Over such a basis, considering that a BS dynamically shares transmission resources among its assigned users by allocating, on average, a given amount of subcarriers to user i, denoted as K_{ij} (being $K_{ij} \leq K_j$) during a given amount of transmission time (denoted as ΔT_{ij}, being $\Delta T_{ij} \leq T_s$, where T_s is a scheduling reference period) we could relate the achievable rate to the amount of allocated subcarriers and transmission time to meet user's minimum rate requirement by:

$$\frac{K_{ij}}{K_j}\frac{\Delta T_{ij}}{T_s} R_{ij}^{\max} \geq R_i^{\min} \tag{6.5}$$

From previous expression, the radio resource cost is defined directly as:

$$\alpha_{ij} \triangleq \frac{R_i^{\min}}{R_{ij}^{\max}} = \frac{K_{ij}}{K_j}\frac{\Delta T_{ij}}{T_s} \leq 1 \tag{6.6}$$

where α_{ij}=1 would mean that the assignment of user i to BS j require all available radio resources at the BS to meet its rate requirement. Attending to practical considerations, we consider a limited set of modulation and coding schemes that must be used in each subcarrier, thus reducing the output of expressions (6.3), (6.4), and (6.6) to a set of discrete values. Then, we define the aggregate peak rate over the air interface of BS j, denoted as C_j^{air}, as the highest achievable aggregate data rate when using all subcarriers continuously with the highest rate MCS.

6.3.2. Transport Resource Cost Function

The transport cost, denoted as β_{ij}, related with the assignment of user i to BS j is defined as the ratio of the minimum data rate required by user i, $R_i^{\min}$, to the available transport capacity of BS j, denoted as C_j^{trans}, that is:

$$\beta_{ij} \triangleq \frac{R_i^{\min}}{C_j^{\mathrm{trans}}} \tag{6.7}$$

As a matter of clearly relating the transport capacity C_j^{trans} to the aggregate peak rate of the radio interface C_j^{air}, we define the transport capacity factor ϕ_j as:

$$\phi_j \triangleq \frac{C_j^{\mathrm{trans}}}{C_j^{\mathrm{air}}} \tag{6.8}$$

Therefore, a transport capacity factor ϕ_j =1 would mean that the transport capacity of BS j has been provisioned to support the aggregated peak rate of the air interface. Notice that dimensioning the backhaul capacity to satisfy the air interface peak rate may not constitute a resource efficient solution since not all cell connections can always simultaneously exploit the fastest MCS. However, occasionally, user distribution in the cell (e.g., most served users being close to BS and enjoying good radio conditions) may turn into aggregate data rates close to air interface peak rate.

6.3.3. Utility Function

The concept of utility function has been have been widely used to develop resource allocation algorithms [97]. Commonly, a utility function is modeled as a non-decreasing function of the amount of allocated network resources, where its shape depends on the expected benefit that resource allocation can bring into the system. For instance, a step function can be used to model a system where allocating resources below a given threshold has no utility at all, but the maximum utility is just achieved when reaching such a threshold.

Here, we formulate the utility function to reflect the bit rate efficiency of the allocated resources to support the data transfer of each user assigned to a given BS. Hence, a utility function denoted as u_{ij} captures the suitability of assigning user i to BS j, so $u_{ij} > u_{il}$ would mean that BS j is more appropriate than BS l to serve user i in terms of the bit rate efficiency. Similarly, $u_{ij} > u_{lj}$ would indicate that is better to assign user i to BS j instead of user l.

Over such a basis, the considered bit rate efficiency in the radio interface is directly associated with the spectral efficiency, while in the transport network it's assumed that all assignments lead to the same bit rate efficiency (the resources needed to transport 1b/s of a user between the aGW and the correspondent BS are assumed to be the same for all BSs, noticing here that other assumptions, e.g., based on transport provisioning costs, could be also possible but are out of the scope of this work). Hence, the utility function is defined as:

$$u_{ij} = \frac{1}{K_j} \sum_{k=1}^{K_j} \log_2(1+\mathrm{SINR}_{i,j,k}) \tag{6.9}$$

As a result, assignments to BSs where users have the highest values of SINR are favored.

6.4. Optimization Problem Formulation

The BS assignment problem is formulated in this section as an optimization problem that aims to maximize the total welfare utility, defined as the sum of the utilities of all assignments, subject to BS resource constraints. Let $B = \{b_{ij}\}_{MxN}$, be the BS assignment matrix whose entry b_{ij} is equal to one if user i is assigned to BS j, otherwise it is equal to zero. This problem can be formally written as:

Find B such as: $$\max_{b_{ij}} \left(\sum_{i=1}^{M} \sum_{j=1}^{N} u_{ij} b_{ij} \right) \tag{6.10}$$

$$s.t. \quad \sum_{i=1}^{M} \alpha_{ij} b_{ij} \le 1 \qquad j = 1,\ldots,N \tag{6.11}$$

$$\sum_{i=1}^{M} \beta_{ij} b_{ij} \le 1 \qquad j = 1,\ldots,N \tag{6.12}$$

$$\sum_{j=1}^{N} b_{ij} = 1 \qquad i = 1,\ldots,M \tag{6.13}$$

$$R_i \ge R_i^{\min} \qquad i = 1,\ldots,M \tag{6.14}$$

$$b_{ij} \in \{0,1\} \tag{6.15}$$

The set of constraints in (6.11) and (6.12) assures that no more radio and transport resources than available are used in each BS. Constraints in (6.13) indicate that all users need to be assigned to a single BS, while (6.14) ensures that the expected bit rate of user i, denoted as R_i, meets the minimum data rate requirement of each user. Finally, to avoid splitting or partial assignment of users to BSs, constraint (6.15) is enforced, which in turn leads to the combinatorial nature of the problem with exponentially growing complexity in the degrees of freedom.

The formulated problem in (6.10)-(6.15) is a non-linear combinatorial optimization problem since entries in the assignment matrix B can only take integer values. In addition, notice that utility and radio resource cost functions are non-linear functions that depend on the SINR values, which in turn depend on the BS assignment solution because of CCI (i.e., mutual dependency). Therefore, both utility and radio resource cost function values are coupled with the assignment of the users in the system, making the BS assignment problem very hard to tackle.

6.4.1. Practical Considerations

In order to make the formulated BS assignment problem tractable, we consider some practical considerations. In particular, we consider a fully-loaded system [75], that is, where all BSs are assumed to transmit at maximum power so that the mutual dependency is avoided. Apart from reducing the complexity of the BS assignment problem, the rationale of considering full-load conditions can be also justified by the fact that it is just under such stressed load conditions where resource allocation algorithms are expected to bring out their potential benefits.

We also consider that the maximum transmission power of BS j is distributed uniformly, on average, over the K_j subcarriers. That is, the average of the transmission power allocated to each subcarrier is assumed to be equal for all subcarriers over the time scale at which the BS assignment algorithm operates. Hence, BSs are supposed to make use of all available subcarriers in the same way (i.e., there is no a subcarrier more favored than another). Over such a basis, the co-channel interference value observed by user i under full load conditions can be estimated from (6.2) as:

$$I_{i,j,k}^{\max} = \sum_{n=1,n\neq j}^{n=N_k} G_{i,n,k} \frac{P_n^{\max}}{K_n} \tag{6.16}$$

where P_n^{max} and K_n are, respectively, the maximum transmit power and the number of used subcarriers in interfering BS n. In this way, the computation of SINR under full load conditions by means of (6.1) does not depend on the BS assignment, as neither do utility and radio costs values.

6.4.2. Problem Mapping into an MMKP

Attending to previous practical considerations, in this subsection we show that the BS assignment problem formulated in (6.10)-(6.15) can be mapped into a Multiple-Choice Multidimensional Knapsack Problem (MMKP) [98]. The MMKP is a variant of the 0-1 knapsack problem, a well-known NP-hard combinatorial optimization problem arisen in many practical and real life problems [99], [100].

The MMKP is illustrated in Figure 6.2. The MMKP considers a set of items, classified in I disjoint groups of J_i items each, and a knapsack (to pack some of them) whose available capacity is modeled by means of S distinct resource constraints represented by $(W_1,W_2,\ldots,W_S)$. Packing item j from group i turns into a benefit (utility) given by u_{ij} at the expenses of using a portion of the knapsack capacity given by $W_{ij} = (w_{ij}^1/W_1, w_{ij}^2/W_2,\ldots, w_{ij}^S/W_S)$. The objective of the MMKP is to exactly select one item from each group to maximize the aggregated utility subject to knapsack's capacity. The canonical formulation of this problem is as follows:

$$\max_{b_{ij}} \left(\sum_{i=1}^{I} \sum_{j=1}^{J_i} u_{ij} b_{ij} \right) \tag{6.17}$$

$$s.t. \quad \sum_{i=1}^{I} \sum_{j=1}^{J_i} \left(\frac{w_{ij}^s}{W_s} \right) b_{ij} \leq 1 \qquad s = 1,\ldots,S \tag{6.18}$$

$$\sum_{j=1}^{J_i} b_{ij} = 1 \qquad i = 1,\ldots,I \tag{6.19}$$

$$b_{ij} \in \{0,1\} \tag{6.20}$$

The MMKP problem is equivalent to our original optimization problem given by equations (6.10)-(6.15) attending to the following considerations. The number of groups I corresponds to the number of users M. The set of items J_i within each group i are the set of N BSs where each user can be allocated. The number of limiting resources is S=2N since there are N BSs, each one having two resource constraints. The amount of resources required for serving user i in BS j (choosing item j from group i) is given by $W_{ij} = (\alpha_{ij}^{1},...,\alpha_{ij}^{s},..., \alpha_{ij}^{N}, \beta_{ij}^{1},..., \beta_{ij}^{s},...,\beta_{ij}^{N})$, where α_{ij}^{s} and β_{ij}^{s} are described next. Since the allocation of user i only requires resources in the serving BS j, $\alpha_{ij}^{s} = \alpha_{ij}$ and $\beta_{ij}^{s} = \beta_{ij}$ if s=j, and $\alpha_{ij}^{s} = 0$ and $\beta_{ij}^{s} = 0$ otherwise, being α_{ij} and β_{ij} the radio and transport resources of such an assignment modeled by means of (6.6) and (6.7), respectively. Hence, for s=1,...,N the inner summation in (6.18) reduces to:

$$\sum_{j=1}^{J_i}\left(\frac{w_{ij}^{s}}{W_s}\right)b_{ij} \equiv \left(\frac{w_{is}^{s}}{W_s}\right)b_{is} = \left(\frac{w_{ij}^{j}}{W_j}\right)b_{ij} = \alpha_{ij}b_{ij} \tag{6.21}$$

And for s=N+1,...,2N to:

$$\sum_{j=1}^{J_i}\left(\frac{w_{ij}^{s}}{W_s}\right)b_{ij} \equiv \left(\frac{w_{is}^{s}}{W_s}\right)b_{is} = \left(\frac{w_{ij}^{j}}{W_j}\right)b_{ij} = \beta_{ij}b_{ij} \tag{6.22}$$

In the above expressions, α_{ij} and β_{ij} are the radio and transport resource cost computed by means of (6.6) and (6.7), respectively. Finally, notice that the constraint regarding the minimum user rate requirement considered in (6.14) is implicitly taken into account within the computation of resource costs.

6.4.3. Algorithm Types to Solve the MMKP

There exist two types of solution methods in the literature for solving the MMKP, namely: exact and heuristic algorithms. The former are capable of producing optimal solutions of the MMKP, and they are mainly based on: (a) branch-and-bound search using depth-first search strategy; (b) dynamic programming techniques; and (c) hybrid algorithms combining dynamic programming and branch-and-bound procedures [101]. On the other hand, the second type of methods consists of approximate procedures or heuristics able to produce near-optimal solutions to the problem. Due that the MMKP is an NP-hard problem, producing a globally optimum solution using an exact algorithm is likely to be too time consuming, and hence it is not suitable for most real-time decision-making applications [102]. Therefore, the alternative is to use approximate heuristic approaches with polynomial time complexity to solve the MMKP.

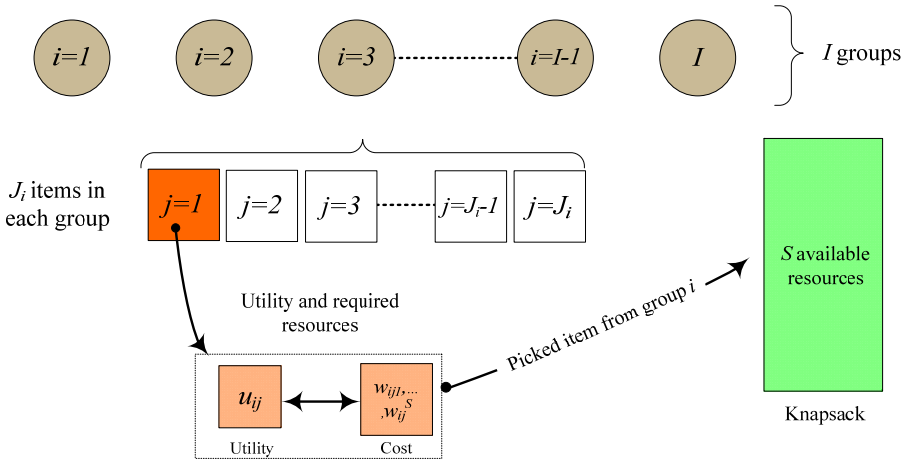


Figure 6.2: Graphical representation of the MMKP

6.5. Heuristic BS Assignment Algorithm

In this section, we develop a heuristic BS assignment algorithm based on the algorithm proposed by Moser et al. [103]. The algorithm uses the concept of "graceful degradation" and relies on a theorem proven by Everett [104] that makes Lagrange multipliers applicable to discrete optimization problems, such as the MMKP. In this regard, algorithm in [103] have already been considered as a useful tool in some works [105] to solve resource allocation problems in OFDMA wireless networks. Therefore, we have adapted the algorithm of Moser et al. to our specific BS assignment problem and introduced some relevant modifications, discussed later on, to the original algorithm. The main underlying concepts behind the adopted approach and the description of the proposed algorithm are provided in next subsections. Then, a detailed description of the algorithm is provided along with an estimation of its computational complexity.

6.5.1. Lagrange Multipliers

According to Everett's Theorem (see [104]), the optimal solution $b_{ij}^* \in \{0,1\}$ of the *unconstrained* maximization problem:

$$\max_{b_{ij}} \left\{ \left(\sum_{i=1}^{M} \sum_{j=1}^{N} u_{ij} b_{ij} \right) - \sum_{j=1}^{N} \lambda_j \sum_{i=1}^{M} \alpha_{ij} b_{ij} - \sum_{j=1}^{N} \mu_j \sum_{i=1}^{M} \beta_{ij} b_{ij} \right\} \tag{6.23}$$

where λ_j and μ_j are two non-negative Lagrange multipliers associated with the radio and transport constraints of BS j, respectively, is also the optimal solution of the constrained optimization problem:

$$\max_{b_{ij}} \left(\sum_{i=1}^{M} \sum_{j=1}^{N} u_{ij} b_{ij} \right) \tag{6.24}$$

$$s.t. \quad \sum_{i=1}^{M} \alpha_{ij} b_{ij} \le \sum_{i=1}^{M} \alpha_{ij} b_{ij}^* \triangleq \pi_j \qquad j = 1,...,N \tag{6.25}$$

$$\sum_{i=1}^{M} \beta_{ij} b_{ij} \le \sum_{i=1}^{M} \beta_{ij} b_{ij}^* \triangleq \tau_j \qquad j = 1,...,N \tag{6.26}$$

The above problem is equivalent to our BS assignment problem except for condition (6.13) discussed later on. From equation (6.23) it is noted that, if Lagrange multipliers λ_j and μ_j are known, the optimization problem can be easily solved. In fact, rewritten equation (6.23) as:

$$\max_{b_{ij}} \left\{ \sum_{i=1}^{M} \sum_{j=1}^{N} (u_{ij} - \lambda_j \alpha_{ij} - \mu_j \beta_{ij}) b_{ij} \right\} \tag{6.27}$$

the optimal solution is given by:

$$b_{ij}^* = \begin{cases} 1 & \text{if } w_{ij} = u_{ij} - \lambda_j \alpha_{ij} - \mu_j \beta_{ij} > 0 \\ 0 & \text{otherwise} \end{cases} \tag{6.28}$$

where w_{ij} is defined as the *weighted utility*, a metric that integrates the utility, radio and transport resource costs and associated Lagrange multipliers. It is worthwhile to note that constraint (6.13) related to group constraints (the user can only be assigned to a single BS among all candidates) can be easily taken into account by selecting, among possible assignments choices in (6.28), the one that provides the maximum weighted utility. Therefore, the BS assignment problem can be solved by computing the set of $2N$ Lagrange multipliers. The assignment solution of all users is feasible if the amount of radio and transport resources allocated in each BS, denoted as π_j and τ_j, in expressions (6.25) and (6.26), respectively, do not exceed available resources, i.e., $\pi_j \le 1$ and $\tau_j \le 1$. Furthermore, the solution is optimal if the following condition is held:

$$\sum_{j=1}^{N}\lambda_j(1-\pi_j)+\sum_{j=1}^{N}\mu_j(1-\tau_j)=0 \tag{6.29}$$

The main difficulty in solving the problem is how to efficiently compute the Lagrange multipliers. In this regard, [103] used the concept of graceful degradation from the most valuable choices based on Lagrange multipliers. First, an initial temporary solution $b_{ij}\in\{0,1\}$ is obtained by assuming all Lagrange multipliers equal to zero, so that the weighted utility equals to the utility, and thus each user is assigned to the "best" BS (i.e., highest u_{ij}) irrespective of its radio or transport load. Then, Lagrange multipliers associated to BSs that would exceed available resources are iteratively increased in a smart way until a feasible solution, if exists, is found. Notice that an increase in the Lagrange multiplier associated to a BS radio/transport constraint results in a reduction of the weighted utility of its served users. As detailed later on, such a reduction in the weighted utility forces that some users could be reassigned to other BSs providing higher weighted utility.

6.5.2. Description of the Algorithm

The developed BS assignment algorithm is shown in Figure 6.4. The algorithm is composed of four phases, namely: initialization, drop, add and relaxation. Firstly, Lagrange multipliers are set to zero (line 01), and then resource costs and user utilities are computed (lines 02-04) for each user. In order to reduce the computational complexity, not all BSs are viewed as potential choices. Instead, each user i is assumed to have a candidate set, denoted as n_i, composed by the BSs having the highest channel gain, which also allows to limit the maximum extent of radio degradation (due to the introduction of backhaul metrics in the BS assignment process). Then, an initial assignment is obtained by selecting the most valuable BS for each user (line 05). The total radio and transport costs at each BS j, denoted by π_j and τ_j, respectively, are computed (lines 06-07), and the resource cost vector $\psi=\{\pi_1,\ldots,\pi_N,\tau_1,\ldots,\tau_N\}$ is conformed (line 08). If the initial assignment is feasible (i.e., none of the elements of ψ is greater than 1.0), that is an optimal solution. Otherwise, the algorithm continues in the drop phase.

Within the drop phase, Lagrange multiplier associated to the most offending constraint violations is repeatedly increased to force user reassignments till a solution not exceeding resource constraints is found. In each iteration of this phase, the BS j^* with the most offending constraint violation s is determined (line 10), where $j^*=s$ if $s=1,\ldots,N$, and $j^*=s-N+1$ otherwise. For each user i currently allocated to the BS j^* (line 12) the Lagrange multiplier of the most offending constraint s required to move user i from BS j^* to another BS j of its candidate set is computed (lines 12-18). This is done so that the weighted utility of user i at the overloaded BS j^*, w_{ij^*}, is decreased to a value less than or equal to the weighted utility on the candidate BS j, w_{ij}. Thus, if the most offending constraint violation at BS j^* is on transport resources, the increment to Lagrange multiplier μ_{j^*} should be such that:

$$(u_{ij^*}-\lambda_{j^*}\alpha_{ij^*}-(\mu_{j^*}+\Delta\mu_{i,j^*\to j})\beta_{ij^*})\le(u_{ij}-\lambda_j\alpha_{ij}-\mu_j\beta_{ij}) \tag{6.30}$$

So, the increment to the transport multiplier, denoted as $\Delta\mu_{i,j^*\to j}$, can be computed as:

$$\Delta\mu_{i,j^*\to j}\ge\frac{u_{ij^*}-u_{ij}-\lambda_{j^*}\alpha_{ij^*}+\lambda_j\alpha_{ij}-\mu_{j^*}\beta_{ij^*}+\mu_j\beta_{ij}}{\beta_{ij^*}} \tag{6.31}$$

Similarly, the increment $\Delta\lambda_{i,j^*\to j}$ to the multiplier of the radio constraint can be computed as:

$$\Delta\lambda_{i,j^*\to j}\ge\frac{u_{ij^*}-u_{ij}-\lambda_{j^*}\alpha_{ij^*}+\lambda_j\alpha_{ij}-\mu_{j^*}\beta_{ij^*}+\mu_j\beta_{ij}}{\alpha_{ij^*}} \tag{6.32}$$

where numerator in (6.31) and (6.32) is the increase of the weighted utility of user i, denoted as $\Delta w_{i,j^*\rightarrow j}$. For each user BS j in the candidate set n_i of users currently allocated to BS j^*, the increase of the corresponding Lagrange multiplier is computed in lines 12-18. Then, as suggested in [103] the user I^* from BS J^* causing the least increase of the corresponding multiplier is chosen (lines 19-25) for exchange as this choice minimizes the gap between the optimal solution characterized by equation (6.28) and the new assignment solution obtained at this point. However, if the multiplier increase is just computed as the equality as done in [103], important convergence problems arise since users tend to have the same weighted utility towards multiple BSs. To avoid this problem, we compute the increment to be added to the corresponding multiplier as the average between the least increase, corresponding to user I^* and BS J^*, and the second least increase obtained with user I' and BS J'. This choice guarantees that only one user is reassigned at each iteration and the next BS assignment solution is stable (equal weighted utilities due to the update of the multipliers are avoided). Furthermore, as the BS assignment problem could have no feasible solution (not enough resources to allocate all the users), condition (6.13) is relaxed at this point by allowing that a user i may not have allocated resources in any BS. We can achieve this by assuming that within the candidate set of each user there is a BS assignment choice with associated resource costs and utility equal to zero. Particularly, we consider that each candidate set has a "null BS". The assignment of a given user to the "null BS" would imply that has not been assigned to any BS.

In line 26 of the drop phase, the reassignment of the selected user is performed, while radio and transport resource costs are updated accordingly in lines 27-28. The process is repeated until a solution not exceeding resource constraints is determined, yet there may be some users not served by any BS. The reassignment procedure performed within the drop phase by means of adjusting Lagrange multipliers is illustrated Figure 6.3. In this example, it is assumed that the most offending constraint violation is on the transport of BS j^*. For each user $i|b_{ij^*}=1$, that is currently allocated to BS j^* the Lagrange multiplier increase $\Delta\mu_{i,j^*\rightarrow j}$ is computed.

The solution arisen from the drop phase may not be the most efficient BS assignment configuration in terms of resource utilization as some BSs could still have available resources. Then, the solution is improved in the add phase by applying the following procedure. For each user i it is verified whether, amongst the BSs in its candidate set, there is an assignment option BS l that provides a higher utility ($u_{il} > u_{ij}$) than current assignment j. The utility increment, denoted as $\Delta u_{i,j\rightarrow l}$, is computed in lines 30-33. Among all user assignments satisfying $\Delta u_{i,j\rightarrow l} > 0$, as well as radio and transport constraints of BS l, the user I'' causing the largest increase in the utility is selected for reassignment (line 34). The exchange is done in line 35 and costs associated with radio and transport constraints are updated in lines 36-37. This process is repeated until no more re-assignments are possible. If achieved solution after the add phase is a feasible solution (all users have been allocated and resources are not exceeded), the algorithm ends, otherwise the algorithm continues on the relaxation phase.

When a feasible solution cannot be found, users without allocated resources after the add phase would have to be dropped or not served temporary (e.g., in case of a joint scheduling and BS allocation problem) in order to guarantee the minimum data rate requirements to the rest of served users. Alternatively, these users can be served at the expenses of allowing some degree of service degradation. This second approach is the one used in this thesis since it allows a fair comparison of the proposed algorithm with other BS assignment strategies in terms of service degradation.

Hence, in the developed algorithm a relaxation phase is considered after the add phase where users without allocated resources are finally allocated to the BS with the highest weighted utility w_{ij} among those of its candidate set. In any case, notice that, as a full load condition has been assumed for the computation of radio resource costs, the resulting BS assignment after the relaxation phase may not necessarily lead to service degradation when real load conditions are accounted. Hence, the output of the presented algorithm is always a complete BS assignment and its feasibility or service degradation caused by exceeding resource constraints is numerically assessed in Section 6.6 by considering accurate load and interference level estimations.

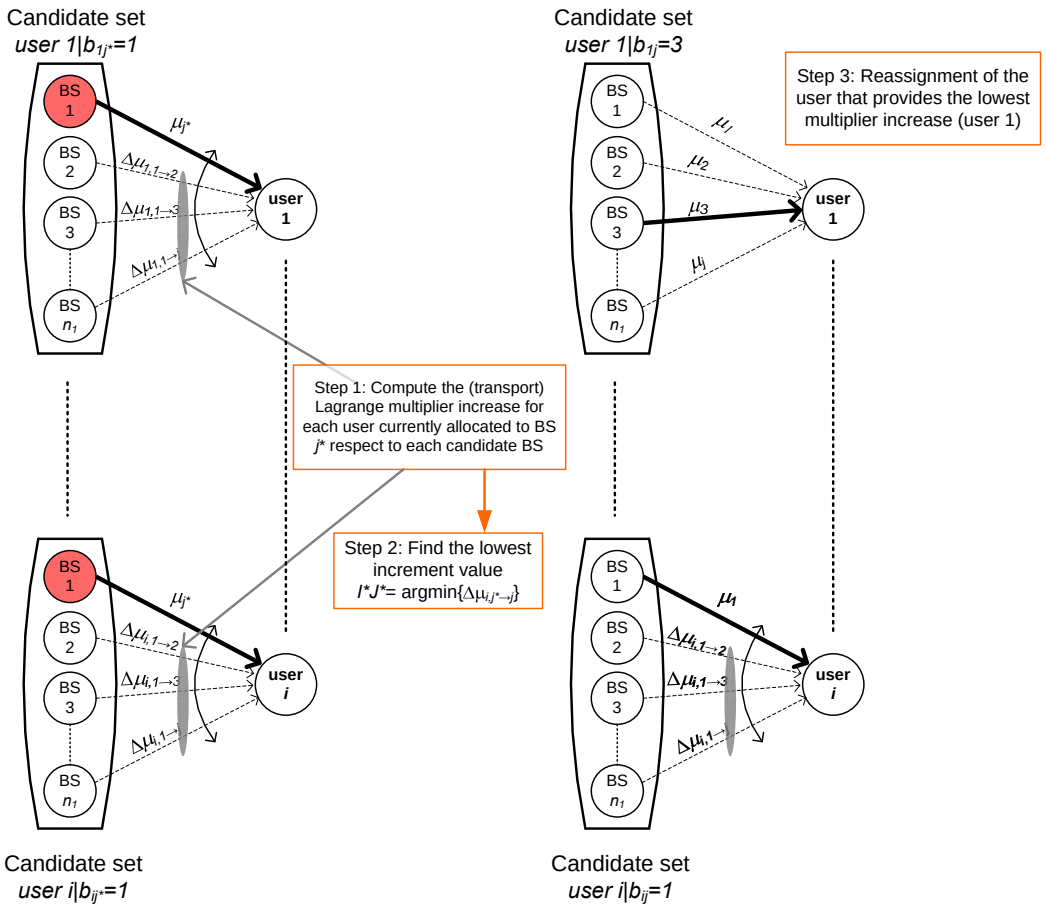


Figure 6.3: Reassignment procedure based on Lagrange multiplier's adjustment.

6.5.3. Complexity Analysis

The algorithm's complexity is determined in this section based on the analysis given in [103]. The initialization phase has a complexity of $O(2N+3\cdot M\cdot n_i)$. In line 09, the `while` loop could be executed up to $O(M\cdot n_i)$ times, as in each iteration one user can be changed from BS j^* to BS J^*. The inner loop (line 12) could perform up to n_i iterations for each user assigned to BS j^*, thus its complexity is $O(M\cdot n_i)$. The computation of the increase of multipliers (lines 13-18) results in a complexity of $O(n_i)$. The complexity of lines 20-25 and lines 26-28 is $O(2M\cdot n_i)$ and $O(3)$, respectively. Thus, in the worst case the complexity order of the drop phase is $O(M^2(n_i)^3 + 2M^2(n_i)^2 + 3M\cdot n_i)$. In the add phase, the complexity of line 34 and lines 35-37 is $O(Mn_i)$ and $O(3)$, respectively. At line 30, for each user we have at most n_i BSs resulting in a complexity of $O(Mn_i)$. The complexity of line 32 is $O(n_i)$, while the while loop (line 29) is executed at most $M\cdot n_i$ times. Thus, the complexity of add phase is $O(M^2(n_i)3+M^2(n_i)^2+M\cdot n_i)$, while the complexity of phase 3 is $O(M\cdot n_i)$. Therefore, the complexity order of the algorithm is given by $O(M^2(n_i)^3)$.

Phase 0: Initialization	01	Set Lagrange multipliers to zero: $\lambda_j \leftarrow 0,\ \mu_j \leftarrow 0$
	02	**for** each user i=1,…,M
	03	**for** each BS j=1,…,n_i on the active set of i compute utility and costs
	04	$u_{ij},\ \alpha_{ij}, \beta_{ij}$
	05	Find the most valuable BS n=argmax$_j${u_{ij}} for each user i and update its assignment accordingly $b_{in} \leftarrow 1$
	06	**for** each BS j compute total radio/transport resource costs
	07	$\pi_j = \sum_{i=1}^{M} \alpha_{ij} b_{ij}, \tau_j = \sum_{i=1}^{M} \beta_{ij} b_{ij}$
	08	Conform total resource cost vector: ψ={$\pi_1,\ldots,\pi_N, \tau_1,\ldots,\tau_N$}
Phase 1: Drop	09	**while** ($\psi_i > 1$ for any j) **do**
	10	Find the BS j* holding the most offending constraint violation s = argmax$_i${ψ_i}, where j*=s if s=1,…,N and j*=s-N+1 otherwise
	11	Compute the increase of the multiplier associated to constraint s (radio/transport) of BS j*
	12	**for** {$i|b_{ij^*} = 1$}
	13	**for** {$j = 1{:}n_i$}
	14	Compute the increase of the weighted utility $\Delta w_{i,j^*\to j} = u_{ij^*} - u_{ij} - \lambda_{j^*}\alpha_{ij^*} + \lambda_j \alpha_{ij} - \mu_{j^*}\beta_{ij^*} + \mu_j \beta_{ij}$
	15	**if** j*=s **then**
	16	$\Delta\lambda_{i,j^*\to j} = \Delta w_{i,j^*\to j} / \alpha_{ij^*}$
	17	**else**
	18	$\Delta\mu_{i,j^*\to j} = \Delta w_{i,j^*\to j} / \beta_{ij^*}$
	19	Find the user to change its assignment and re-evaluate the corresponding Lagrange multiplier
	20	**if** j*=s **then**
	21	I^*J^* = argmin$_{ij}$ {$\Delta\lambda_{i,j^*\to j}$}, $I'J'$ = argmin$_{ij}$ {$\Delta\lambda_{i,j^*\to j}$}, $I' \neq I^*$
	22	$\lambda_{j^*} \leftarrow \lambda_{j^*} + (\Delta\lambda_{I^*,j^*\to J^*} + \Delta\lambda_{I',j^*\to J'}) / 2$
	23	**else**
	24	I^*J^* = argmin$_{ij}$ {$\Delta\mu_{i,j^*\to j}$}, $I'J'$ = argmin$_{ij}$ {$\Delta\mu_{i,j^*\to j}$}, $I' \neq I^*$
	25	$\mu_{j^*} \leftarrow \mu_{j^*} + (\Delta\mu_{I^*,j^*\to J^*} + \Delta\mu_{I',j^*\to J'}) / 2$
	26	$b_{I^*j^*} \leftarrow 0, b_{I^*J^*} \leftarrow 1$
	27	$\pi_{j^*} \leftarrow \pi_{j^*} - \alpha_{I^*j^*}, \tau_{j^*} \leftarrow \tau_{j^*} - \beta_{I^*j^*}$
	28	$\pi_{J^*} \leftarrow \pi_{J^*} + \alpha_{I^*J^*}, \tau_{j^*} \leftarrow \tau_{j^*} + \beta_{I^*J^*}$
Phase 2: Add	29	**while** more assignments can be exchanged **do**
	30	**for** {$i = 1{:}M$}
	31	j= argmax$_j${b_{ij}=1}
	32	**for** {$l = 1{:}n_i$}
	33	$\Delta u_{i,j\to l} \begin{cases} u_{il} - u_{ij} & \text{if } u_{il} - u_{ij} > 0, \pi_l + \alpha_{il} \le 1,\ \tau_l + \beta_{il} \le 1 \\ 0 & \text{otherwise} \end{cases}$
	34	Find the best exchangeable assignment: $I''J''$ = argmax$_{ij}${$\Delta u_{i,j\to l}$} to move user I'' from BS j to BS J''
	35	$b_{I''j} \leftarrow 0, b_{I''J''} \leftarrow 1$
	36	$\pi_j \leftarrow \pi_j - \alpha_{I''j}, \tau_j \leftarrow \tau_j - \beta_{I''j}$
	37	$\pi_{J''} \leftarrow \pi_{J''} + \alpha_{I''J''}, \tau_{J''} \leftarrow \tau_{J''} + \beta_{I''J''}$
Phase 3: Relaxation	38	Each user i with b_{ij}=0 is assigned to the BS with max. weighted utility
	39	**for** {$i|b_{ij}$ =0 for any j}
	40	**for** {$j = 1{:}n_i$}
	41	$w_{ij} = u_{ij} - \lambda_j \alpha_{ij} - \mu_j \beta_{ij}$
	42	j = argmax$_j${w_{ij}}
	43	$b_{ij} \leftarrow 1$
	44	Final BS assignment configuration B={b_{ij}}$_{MxN}$

Figure 6.4: Pseudo-code of the heuristic BS assignment algorithm.

6.6. Performance Evaluation

In this section, the developed BS assignment algorithm is evaluated and compared to two BS assignment approaches that are exclusively based on radio criteria. In the presented analysis, it is considered that the BS decision-making process is able to follow channel variations due to propagation path loss and slow shadowing changes (some authors refers to the time scale dictated by path loss and shadowing changes as a macroscopic time scale, in opposition to a microscopic time-scale dictated by fast fading). Hence, minimum user bit rate requirements and resource costs

Table 6.1: MCS thresholds and maximum achievable data rates.

#	Modulation	Coding rate	$SINR_{min}$ [dB]	Physical data rate [Mbps]
1	BPSK	1/2	3.4	6.99
2	QPSK	1/2	6.4	13.99
3	QPSK	3/4	8.2	20.99
4	16 QAM	1/2	13.4	27.98
5	16 QAM	3/4	15.2	41.98
6	64 QAM	2/3	19.7	55.97
7	64 QAM	3/4	21.4	62.97

considered in the developed BS assignment algorithm would represent average values taken over the time scale dictated by long-term channel variations (e.g., few hundreds of milliseconds). Under such an approach, the mean channel gain in each subcarrier k from BS j to user i, referred to $G_{i,j,k}$, is the same for all subcarriers. Consequently, the computation of $SINR_{i,j,k}$ according to equation (6.1) leads also to the same value for all subcarriers, namely $SINR_{i,j}$, since the interference levels are assumed to be uniformly distributed over the entire bandwidth, as argued in Section 6.4.1 and captured by equation (6.16). So, upon the average $SINR_{i,j}$ that denotes the quality of the channel between user i to BS j, the MCS and consequently the achievable rate at the air interface are taken from the look-up table provided in Table 6.1. The maximum achievable rate (also referred to as physical data rate) values provided in the look-up table are computed according to:

$$R_{ij}^{\max} = \frac{c_r \cdot b_s \cdot K_j}{t_s} \tag{6.33}$$

where c_r is the coding rate, b_s is the bits per symbol, K_j is the number of allocated data subcarriers, and t_s is the symbol duration time. The propagation losses are computed using the COST-231 Hata model. Shadowing is modeled with an 8 dB log-normal standard deviation for shadowing effect and spatial shadowing correlation of 50%. The radius of the cell has been chosen so that a signal to noise ratio SNR_{req} = 3.4 dB is assured at the cell border with a probability of 95%, considering typical sample link budgets for mobile broadband systems [106]. All system parameters are summarized in Table 6.2.

The approach adopted in this work does not preclude the applicability of the proposed algorithm in a problem also tackling fast fading fluctuations in the channel gain, e.g., a joint scheduling and BS assignment problem. However, this alternative approach is out of the scope of the current work that mainly tries to expose the benefits (or needs) to incorporate both radio and transport information in the BS assignment problem. As well, it is worth noting that bit rate values provided in the Table 6.1 could also account for any performance gain associated with the usage of mechanisms exploiting (subcarrier) frequency selectivity as well as multi-user diversity that would operate at shorter time scales than that considered for the BS assignment process.

6.6.1. Preliminary Assessment

The need of including backhaul-related information in the BS assignment process is initially evaluated under a two-cell scenario. Particularly, we demonstrate that traditional BS assignment strategies that are exclusively ruled by radio criteria can fail to find a proper BS assignment solution without violating QoS requirements of connections in scenarios where the backhaul network can become the bottleneck. We prove this fact by means of a simple but illustrative two cell scenario.

We assume N=2 BSs with identical radio and backhaul configuration and M=8 users uniformly distributed in the service area. One possible snapshot of the system is shown in Figure 6.5. A minimum rate requirement of $R_i^{\min} \approx 4.7$ Mbps is assumed for each user i. The rate requirement of each user corresponds to the 7.5% of the peak rate capacity of the BS. This implies that the total requested traffic load per BS would correspond to 30% of the BS peak rate denoted as C_j^{air}, if each

Table 6.2: OFDMA system parameters.

Parameter	**Value**
Max. BS transmission power, P_j^{max}	47 dBm
Transmit antenna gain	18.7 dBi
Cell radius	1060 m
Antenna pattern	Omnidirectional
Operating frequency	2500 MHz
Reuse factor	3
Number of channels	3
Channel bandwidth	20 MHz
Number of data subcarriers, K_j	1440
OFDM symbol duration, t_s	102.9 μs
Path loss model	COST-231 Hata
BS height	32 m
Mobile terminal height	1.5 m
Shadowing standard deviation	8 dB
Shadowing correlation	50%
Shadow fade margin	13.2 dB
Thermal noise	-174 dBm/Hz
Receiver noise figure	7 dB

BS provides service to four users. The backhaul capacity factor of BSs is assumed to be ϕ_j=35% (C_j^{trans}≈22 Mbps). Over such a basis, in Table 6.3 we provide the values of the relevant parameters that determine possible user assignments.

As shown in Table 6.3 and Figure 6.5, the assignment choice for users 1 to 7 is clear because only one BS out of the two have enough radio resources to guarantee the minimum required data rate to each user. Note that either a MPL or a radio load (RL) balancing criterion would provide the same assignment solution, that is, users 1 to 3 assigned to BS 1 while users 4 to 7 assigned to BS 2. On the other hand, user 8 can be allocated to any of the two cells from a radio resource viewpoint as in both BSs it could meet its rate requirement. Here, an MPL criterion would assign user 8 to BS 2, as the path loss to BS 2 is lower than to BS 1 (i.e., 139 dB in front of 142 dB). As well, a RL

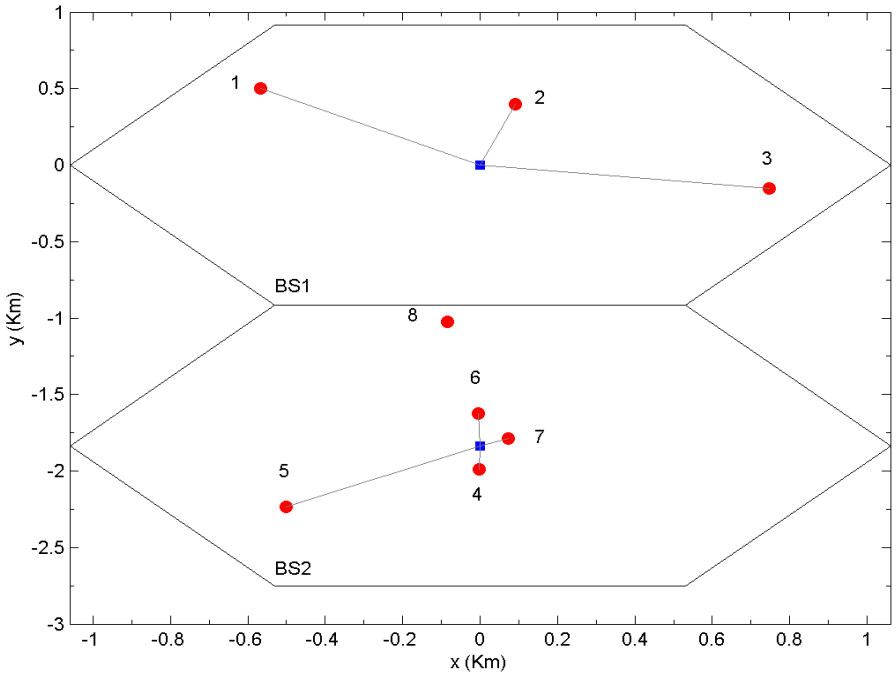


Figure 6.5: Illustrative two-cell scenario.

Table 6.3: Relevant parameters of the assignment process.

Parameter	BS *j*	User *i*							
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Distance (m)	1	750	410	761	1990	2287	1623	1791	1030
	2	2403	223	1840	154	637	213	86	814
Path loss (dB)	1	137.9	128.6	138.0	152.7	154.8	149.6	151.1	142.6
	2	155.5	154.4	151.5	113.7	135.3	118.6	104.9	139.1
SINR (dB)	1	12.8	22.6	12.7	-5.3	-10.3	-1.0	-3.1	7.5
	2	-12.3	-8.0	-5.6	39.2	15.8	31.6	45.0	11.3
Physical data rate (Mbps)	1	20.99	62.97	20.99	0.93	0.29	2.50	1.55	13.99
	2	0.18	0.50	0.87	62.97	27.98	62.97	62.97	20.99
Radio resource consumption, α_{ij}	1	0.225	0.075	0.225	>1	>1	>1	>1	0.337
	2	>1	>1	>1	0.075	0.168	0.075	0.075	0.225
Backhaul resource consumption, β_{ij}	1	0.214	0.214	0.214	0.214	0.214	0.214	0.214	0.214
	2	0.214	0.214	0.214	0.214	0.214	0.214	0.214	0.214
Assigned BS *j*		1	1	1	2	2	2	2	1 or 2

criterion would also select BS 2 to serve user 8 as its radio load is lower than that of BS 1 (i.e., the total radio load of users 1 to 3 at BS 1 is 52.5%, while total radio load of the assignment of users 4 to 7 at BS 2 is 39%). It is important to note, however, that the assignment decision taken by an MPL or a RL criterion would lead to overloading the backhaul capacity of BS 2.

The impact of each assignment criterion on both radio and backhaul load of BSs is depicted in Figure 6.6. As can be seen, the assignment decision regarding user 8, taken by MPL and RL, would deteriorate the QoS of all users connected to BS 2. On the other hand, if user 8 is assigned to BS 1, which is a better choice attending to a backhaul load balancing viewpoint, QoS constraints can be actually satisfied for all active users. This proves that applying only radio aspects to guide the BS selection does not suffice when backhaul can also become the bottleneck. In such situations, assignment criteria combining both radio and backhaul load (RBL) information is needed to be aware of eventual congestion in either the radio or the backhaul and distribute the load among BSs accordingly. Notice that the introduction of backhaul load balancing into the BS assignment might lead to an increase of downlink radio load (e.g., if user 8 is assigned to BS 2, the mean radio load is 50.5%, while assigning it to BS 1 the mean radio load becomes 56%) but otherwise this radio capacity would have remained unused and the QoS of some users deteriorated. Using a two-cell scenario we have demonstrated that traditional BS assignment schemes relying exclusively on radio

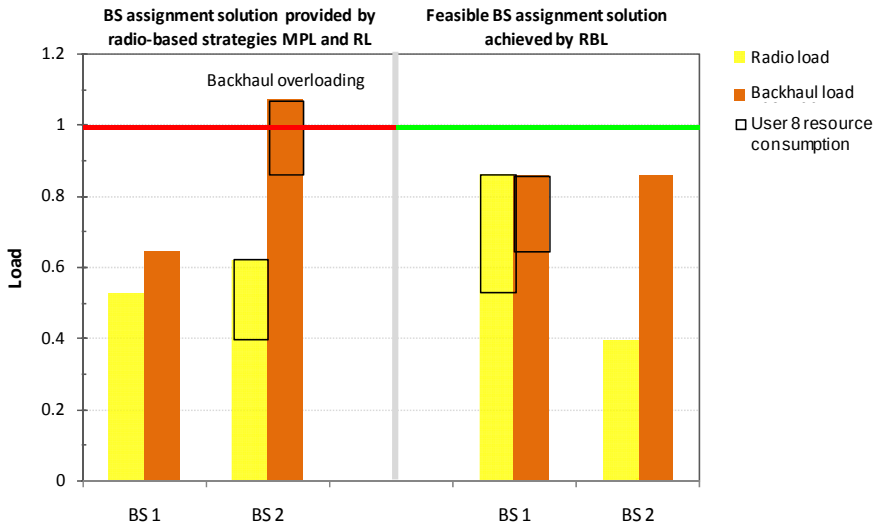


Figure 6.6: BS radio and backhaul load when user 8 is assigned to BS 2 by MPL and RL (left side) or to BS 2 by RBL (right side).

performance metrics, such as path loss and radio load, can fail to find a proper BS assignment without violating users' QoS requirements when backhaul capacity becomes the main resource bottleneck. On the other hand, a backhaul-aware BS assignment scheme is able to fully exploit available cellular network for rate guaranteed services by performing load balancing among BSs according to both radio and backhaul capacity limitations. This can be achieved by introducing backhaul load information in the BS assignment process, so that users are distributed among BSs in order to avoid backhaul overload situations.

6.6.2. Assignment Algorithms Definition

We now define three BS assignment algorithms to implement the BS assignment criteria considered in the example discussed in previous section to solve the BS assignment problem in OFDMA-based cellular networks:

- The first algorithm, called *Algorithm A*, is based on the MPL assignment criterion. This algorithm can be easily implemented by selecting for each user the BS having the highest channel gain. Notice that under full radio load conditions this approach would be equivalent to an algorithm that assigns each user to the BS that provides the highest SINR.

- In the second algorithm, called *Algorithm B*, BS assignment decisions are taken by means of the RL criterion. In particular, the algorithm decides the assignment of each user in the system based on the load level of each BS at the air interface. We have implemented *Algorithm B* by means of a straightforward adaptation of the heuristic algorithm given in Figure 6.4. Specifically, in this case the computation of the more violated constraint violation obtained in the drop phase of the algorithm (see the pseudo-code in Figure 6.4) is obtained from the set of total radio resource costs of BSs (without including total BS transport resource costs). Notice that in this way, the Lagrange multipliers associated to transport constraints of BSs will remain set to zero during the execution of the algorithm. As a result, the weighting utility modeled by means of equation (6.28), and consequently the user assignment procedure, would solely depend on the radio load of BSs.

- Finally, the third assignment approach, referred to as *Algorithm C*, is an algorithm that implements the RBL strategy in order to account for both radio and transport load in the BS assignment decision process. *Algorithm C* is implemented by means of the developed heuristic algorithm detailed in Figure 6.4.

6.6.3. Evaluation Methodology

The process followed to evaluate the considered BS assignment algorithms is illustrated in Figure 6.7. For a given snapshot of the system, where M users are randomly distributed in the service area, we use the three BS assignment algorithms to compute a BS assignment solution for all users in the system under full load conditions. Then, for each obtained solution a more accurate estimation of load and interference levels than the one provided assuming full load conditions is computed. This can be achieved because once the BS assignment solution has been found it is possible to compute the power levels by means of a recursive approach such as the one proposed in [107] that solves power levels under a fixed BS assignment. Such an estimation of load and interference levels is needed to allow a fair comparison of the three different schemes. The real interference levels for a given BS assignment solution would be less than or equal to the maximum one computed by means of equation (6.16), and can be estimated as:

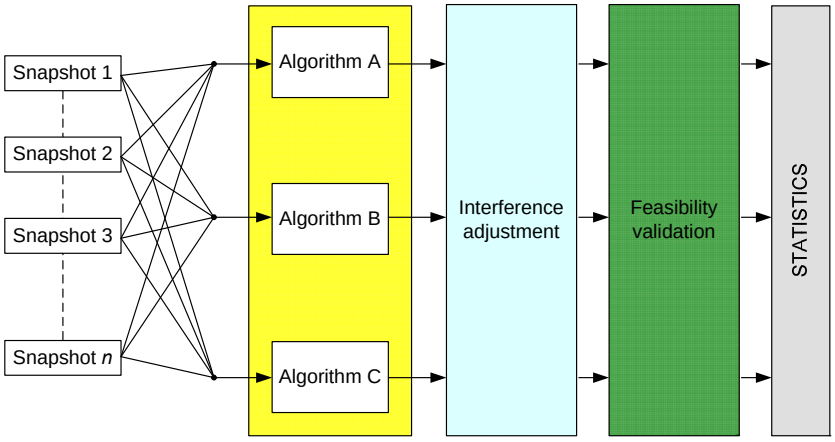


Figure 6.7: Diagram's block of evaluation methodology of BS assignment algorithms.

$$\tilde{I}_{i,j,k} = \sum_{n=1,n\neq j}^{n=N} G_{i,j,k} \frac{P_n^{\max}}{K_n} \theta_n \leq \sum_{n=1,n\neq j}^{n=N} G_{i,j,k} \frac{P_n^{\max}}{K_n} = I_{i,j,k}^{\max} \tag{6.34}$$

where θ_n denotes the radio interface's real load level of the interfering BS n, expressed as:

$$\theta_n = \frac{\sum_{i=1}^{M} \tilde{\alpha}_{in} b_{in}}{\max(1, \sum_{i=1}^{M} \tilde{\alpha}_{in} b_{in}, \sum_{i=1}^{M} \beta_{in} b_{in})} \tag{6.35}$$

where $\tilde{\alpha}_{in}$ denotes the real radio cost that is computed as follows. For the BS assignment solution provided by each algorithm, we compute the real interference levels by means of equation (6.34), in which the real load level at each BS n, denoted as θ_n, is taken into account. Based on the real interference conditions, the SINR for each user is computed and then the maximum achievable rate is obtained from Table 6.1. Once the maximum transmission rate is obtained, and taking into account the minimum user's data rate requirement, the real radio resource cost of each user is computed using equation (6.6).

In order to compute the radio interface's real load level θ_n, we use a recursive algorithm where interference values are adjusted according to the current radio load at each BS. Particularly, starting from full load conditions (i.e., initial values for θ_n are set to one), at each iteration, correspondent radio costs are computed and a new value for θ_n is obtained from equation (6.35) until the algorithm converges (notice that convergence is always achieved by not allowing values for θ_n greater than one). Here we consider two additional tiers of cells in the scenario, so that the M users of a given snapshot are distributed over the 19 inner cells (central cell and the first two tiers of cells), while cells in the third and fourth have a fixed air interface load of θ_n = 0.5 and are considered to avoid the border effect in the characterization of the real interference. It is worth noting that in the simulations performed to solve the BS assignment problem, the cells in the two outer tiers of the cellular layout are assumed to be under full load conditions.

The output of the interference adjustment block provides the real radio resource costs of each of the BS assignment configurations computed by the algorithms. The next step of the evaluation process consists in determining whether the BS assignment solution of each algorithm is feasible or not. This is performed at the feasibility validation block. To this end, we define the feasibility indicator Φ_s, with $s=\{1,\ldots,n\}$, that is equal to 1 if the BS assignment solution for the snapshot s each user meets its rate requirement and each BS $j \in \{1,\ldots,N\}$ meets its resource constraints (radio and transport). Otherwise, the provided BS assignment is not feasible. This can be expressed as:

$$\Phi_s = \begin{cases} 1 & \text{if} \quad \sum_{i=1}^{M} \tilde{\alpha}_{ij} b_{ij} \le 1 \text{ and } \sum_{i=1}^{M} \beta_{ij} b_{ij} \le 1, \quad \forall_j \\ 0 & \text{otherwise} \end{cases} \tag{6.36}$$

The aforementioned process is repeated for each analyzed case, that is, a given distribution of users per cell and under particular user rate requirements and BS transport capacity conditions. Then, we can compute the percentage of feasible solutions that each BS assignment algorithm can provided over the *n* snapshots as:

$$\text{Feasibility } (\%) = \frac{\sum_{s=1}^{x} \Phi_s}{x} \times 100 \tag{6.37}$$

After the execution of the *n* snapshots, different statistics (e.g., percentage of feasibility, radio/transport resource costs, utilities, SINR, user data rates, etc) are finally collected in the statistics block. Furthermore, in order to verify which resource type constitute the main capacity bottleneck in the solution provided by each BS assignment algorithm in the snapshot *s*, it is also defined the radio constraint feasibility, denoted as Φ_s^R, given by:

$$\Phi_s^R = \begin{cases} 1 & \text{if} \quad \sum_{i=1}^{M} \tilde{\alpha}_{ij} b_{ij} \le 1, \quad \forall_j \\ 0 & \text{otherwise} \end{cases} \tag{6.38}$$

and the transport constraint feasibility, denoted as Φ_s^T, and expressed as:

$$\Phi_s^T = \begin{cases} 1 & \text{if} \quad \sum_{i=1}^{M} \beta_{ij} b_{ij} \le 1, \quad \forall_j \\ 0 & \text{otherwise} \end{cases} \tag{6.39}$$

For a given snapshot *s* of the system, $\Phi_s^T = 0$ would mean, for instance, that from the transport point of view there are no sufficient resources at all BSs to serve its assigned users and guarantee the requested minimum data rate. Then, using equations (6.38) and (6.39) it is possible to determine the percentage of solutions satisfying radio constraints and percentage of solutions satisfying transport constraints, respectively, over the *x* performed snapshots. In simulations, the performance evaluation of the BS assignment algorithms is carried out considering *n*=10000 different snapshots for each analyzed case.

It is worth noting that when the solution found by a BS assignment algorithm to the snapshot *n* is not feasible, service degradation will be experienced by each user *i* been served by BS *j* exceeding its radio and/or transport resources. The extent of such service degradation due to insufficient resources at a given BS to guarantee the minimum data rate requirements to its assigned users will be assessed in Section 6.7.3.

6.7. Simulation Results

We study the performance of the BS assignment algorithms defined in Section 6.6.2 in a cellular deployment composed by 19 hexagonal cells (one central cell and its two concentric tiers). The system has three frequency channels with 20 MHz bandwidth and a frequency reuse pattern of 3. The maximum transmit power of a BS is 47 dBm. The transport capacity in BS *j* is expressed in terms of the transport capacity factor ϕ_j according to equation (6.8), and is assumed to be the same for all BSs so that herein we drop index *j*. For each user we consider a candidate set of BSs where the user can be allocated. The candidate set of each user consist of the seven BSs with the highest channel gain. Users are uniformly distributed over the entire service area and all are assumed to have the same downlink data bit rate requirement $R^{\min}$. The list of parameters considered in this study is given in Table 6.2. We also consider a maximum radio cost $\alpha_{ij}^{\max}$ to prevent that a user

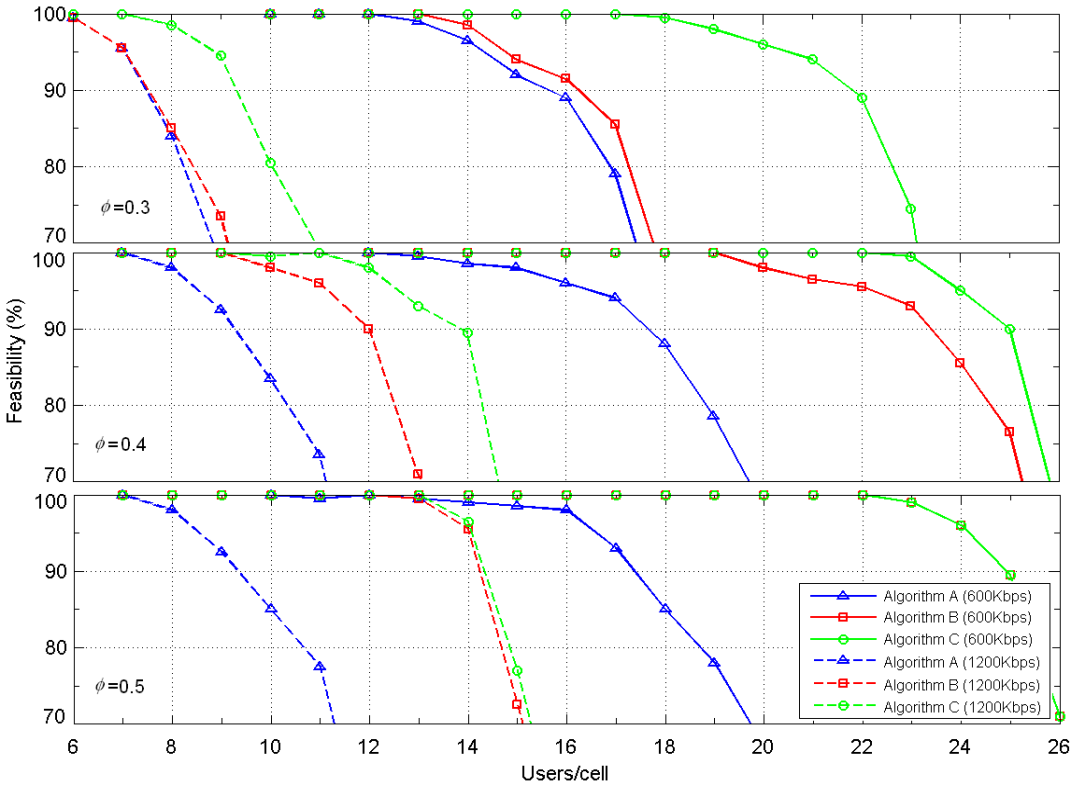


Figure 6.8: Feasible solutions (%) found by each BS assignment algorithm under different transport capacity factors ϕ={0.3, 0.4, 0.5} and data rate requirements Rmin={600, 1200 Kbps}.

may consume an excessive share of overall BS radio resources to meet its requirement. Therefore, the expected data bit rate for user *i*, denoted as R_i, is always limited by:

$$R_i \leq \min(R^{\min}, R_{ij}^{\max} \cdot \alpha_{ij}^{\max}) \qquad (6.40)$$

In the evaluation of the BS assignment algorithms, four different downlink data bit rate requirements are considered $R^{\min}$ ={600, 1200, 1800, 2400 Kbps}. As captured in equation (6.40), the data rate requirement of a given user is satisfied whenever its associated radio resource cost (i.e., computed by means of equation (6.6)) does not exceed a maximum amount of radio resources. In this sense, we assume $\alpha_{ij}^{\max}$=0.2 which implies that a BS can devote a maximum of 20% of total radio resources to a single user in order to meet its downlink data bit rate requirement.

6.7.1. Feasible BS Assignment Solutions

In this section we compare the performance of the three BS assignment algorithms in terms of the percentage of feasible solutions they can find under different transport capacity conditions of BSs in the system and minimum data rate requirements of users. Figure 6.8 presents the percentage of feasible solutions (i.e., all users assigned without service degradation) that each algorithm is able to achieve attending to the mean number of users per cell and considering different minimum rate requirements and transport capacity factors. As shown in Figure 6.8, *Algorithm A*'s performance is always quite poor when compared to load aware schemes. On the other side, *Algorithm C* clearly achieves the highest number of feasible solutions by exploiting both radio and transport load balancing. Notice that, only for transport capacity factors equal to or higher than half the value of the radio peak rate ($\phi \geq 0.5$), *Algorithm B* converges to *Algorithm C* for the considered user bit rates.

Figure 6.9 shows the percentage of feasible solutions for each resource constraint. We consider transport capacities of ϕ={0.3,0.4}, with a minimum data rate requirement of $R^{\min}$=600 Kbps and under different mean load of users per cell. From Figure 6.9 (a) it is in general observed that the

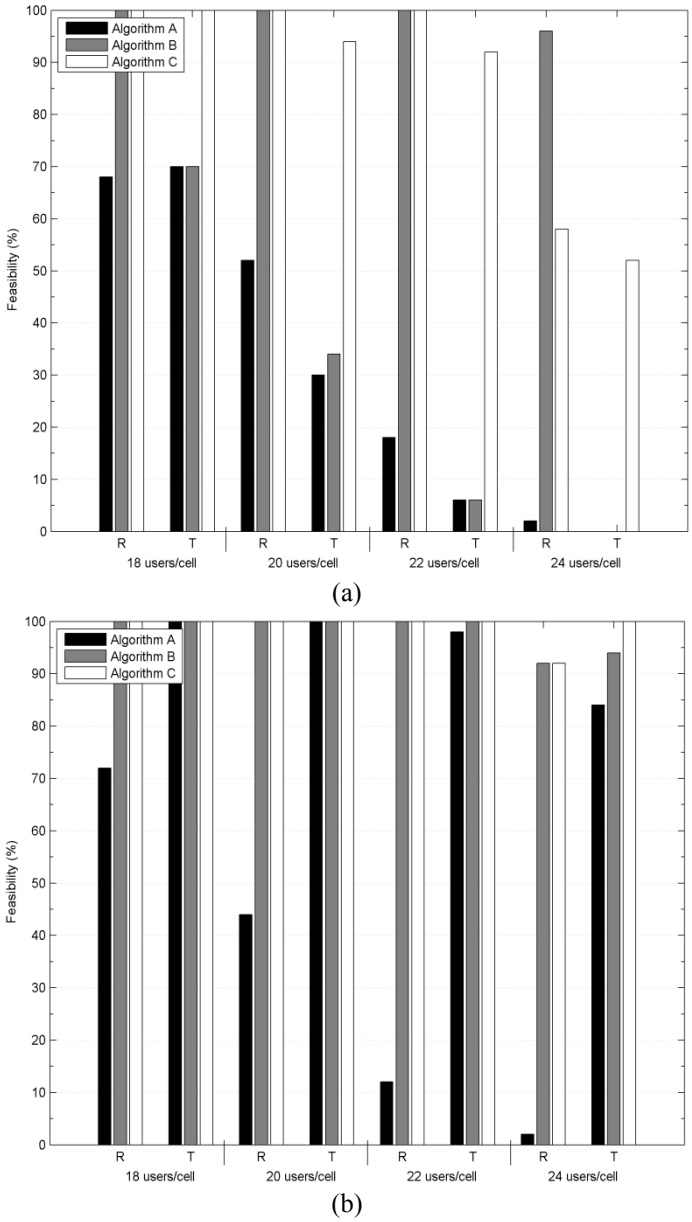


Figure 6.9: Percentage of feasibility for radio (R) and transport (T) constraints under different mean number of users per cell and transport conditions: (a) ϕ=0.3, (b) ϕ=0.4; with rate requirement 600 Kbps.

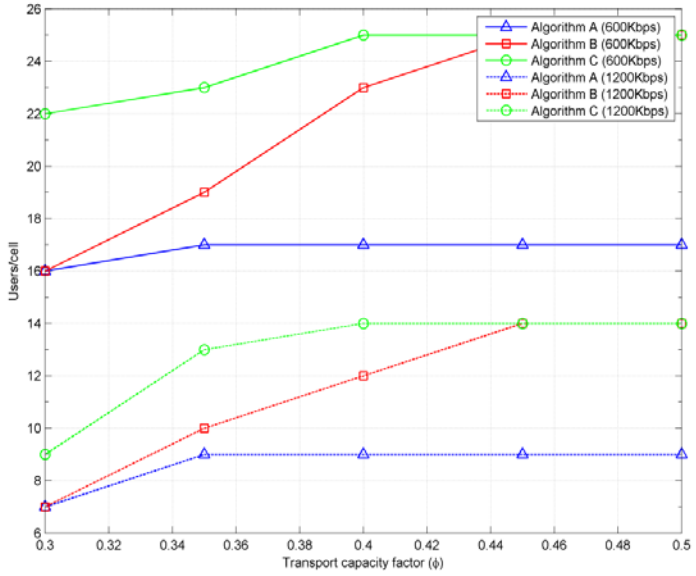
overall percentage of feasible solutions analyzed in Figure 6.8 is mainly limited by the transport (T) constraint of BSs. Although transport resources constitute the main capacity limitation under the considered values of ϕ, *Algorithm C* outperforms the two baseline algorithms. Particularly, for mean load of 20 users per cell, *Algorithm C* is able to provide a feasibility of transport resources of around 95%, whereas algorithms *A* and *B* can obtain a transport constraint feasibility of 30% and 35%, respectively. On the other hand, it is observed that when the transport capacity is increased (see Figure 6.9 (b)) the performance differences between algorithm *B* and *C* are reduced, which implies a less stringent transport capacity restrictions at BSs.

6.7.2. Supported Users versus Transport Capacity

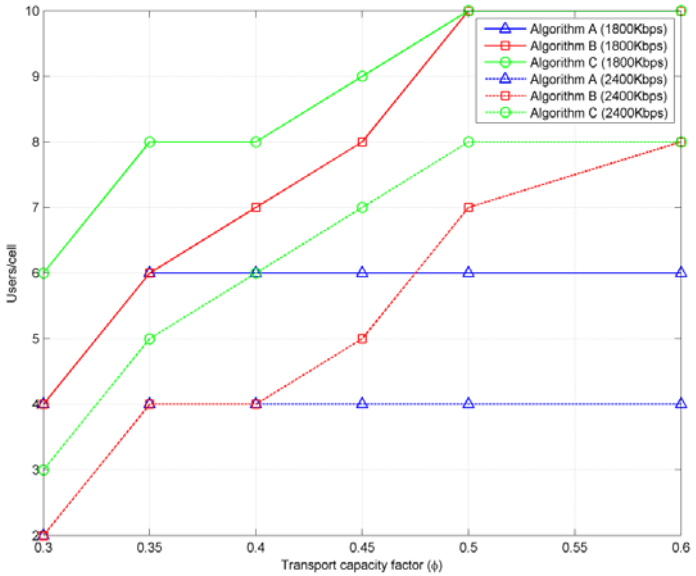
In this section, we aim to analyze the transport capacity required to support a given number of users without violating minimum data rate requirements and considering a given network

availability. In this sense, Figure 6.10 provides the maximum number of users per cell supported by each algorithm when targeting a percentage of feasible solutions equal to 90% (i.e., a BS solution satisfying all user rate requirements and BS resource constraints is found in the 90% of the snapshots). Results are obtained for transport capacities $0.3 \leq \phi \leq 0.6$ and minimum rate requirements R^{min} ={600, 1200, 1800, 2400 Kbps}. Notice that minimum rate requirements are between 1% and 4% when normalized to the BS peak rate.

As shown in Figure 6.10, the relative number of users that can be successfully allocated by algorithms *B* and *C* in front of *Algorithm A* is very noticeable for any transport capacity, specifically under high data rate requirements. For instance, for R^{min}=2400 Kbps and transport capacity factor ϕ=0.5, see Figure 6.10 (b)**,** algorithms *B* and *C* provide capacity gains of 75% and 100%, respectively, over *Algorithm A* (i.e., 4, 7 and 8 users per cell achieved by algorithm *A*, *B*, and *C*, respectively). Under the same transport capacity but a lower data rate requirement R^{min} =1200 Kbps, algorithms *B* and *C* both can support 14 users per cell, in front of 9 users per cell supported *Algorithm A*, which turns into a capacity gain of 56% over *Algorithm A*.



(a)



(b)

Figure 6.10: Supported users/cell for a network availability of 90%. Data rate requirements (a) R^{min}={600, 1200 Kbps}, and (b) R^{min}={1800, 2400 Kbps}.

Notice that under the more limited transport capacity conditions (i.e., $\phi < 0.5$) and the higher data rate requirements, the more is the capacity gain achieved by *Algorithm C*, or, equivalently, the less is the transport capacity needed to support the same number of users in the system. For instance, Figure 6.10 (b) shows that in order to support 8 users/cell with a data rate requirement of 1800 Kbps (i.e., a total aggregated rate of 14.4 Mbps), *Algorithm B* requires around 28 Mbps of backhaul capacity to meet the considered network availability. On the other hand, the level of backhaul resources needed in this case by *Algorithm C* is around 22 Mbps, which turns into a capacity reduction of about 21% in respect to *Algorithm B*.

6.7.3. Impact of Capacity Limitations on User Data Rates

When a feasible BS assignment solution cannot be found due to shortage of BS resources, service degradation is experienced by users. The extent of such degradation is quantified here by considering that a BS j exceeding its radio and/or transport resources proportionally reduces the rate allocated to each served user i. The rate allocated to each user i assigned to BS j is computed as $R_i = R_i \cdot \gamma_j$, where γ_j is the rate reduction factor to be applied, defined as:

$$\gamma_j = \frac{1}{\max(1, \sum_{i=1}^{M} \tilde{\alpha}_{ij} b_{ij}, \sum_{i=1}^{M} \beta_{ij} b_{ij})}$$

In this context, Figure 6.11 illustrates the cumulative distribution of the allocated data rate for a transport capacity factor ϕ=0.3, a distribution of 12 users per cell and data rate requirement R^{min}=1200Kbps. It is shown that *Algorithm A* exhibits the highest degradation, so the requested minimum data rate is only guaranteed to a 74% of the total users. Conversely, the degradation is less pronounced for algorithms *B* and *C*, which can provide the minimum rate requirement to around 80% and 90% of users, respectively.

Furthermore, in order to quantify which is the extent of the service degradation arisen under different configurations of mean traffic and existing backhaul capacity, in Table 6.4 we consider different transport capacity factors ϕ={0.3, 0.4, 0.5}, data rate requirements R^{min} ={1200, 2400 Kbps}, and traffic load conditions (e.g., mean aggregated rates of 14.4, 19.2 and 24.0 Mbps).

Each cell in Table 6.4 provides the percentage of users receiving the minimum requested data rate (upper row), and the percentage of users receiving at least 90% of the requested data rate (bottom row). It is shown that for a mean aggregated rate of 19.2 Mbps at BSs, with R^{min}=1200

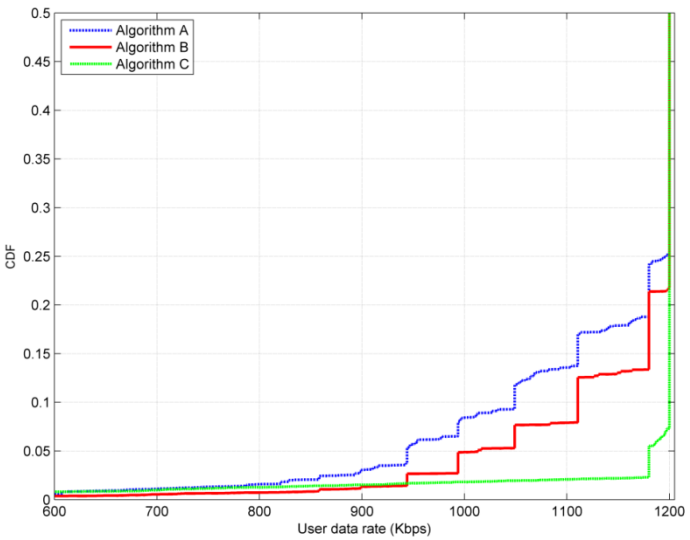


Figure 6.11: CDF of allocated data rate under a scenario with a distribution of 12 users/cell, with data rate requirement R^{min}=1200Kbps, and transport capacity factor ϕ=0.3.

Table 6.4: Satisfaction of users under different mean load of users/cell and transport capacity conditions.

User rate (Kbps)	Users per cell	Mean rate (Mbps)	Algorithm A			Algorithm B			Algorithm C		
			$\phi = 0.3$	$\phi = 0.4$	$\phi = 0.5$	$\phi = 0.3$	$\phi = 0.4$	$\phi = 0.5$	$\phi = 0.3$	$\phi = 0.4$	$\phi = 0.5$
1200	12	14.4	74.6%	88.9%	89.9%	78.1%	96.1%	97.0%	92.7%	96.7%	97.0%
			87.4%	96.5%	96.5%	91.3%	98.9%	98.9%	98.2%	98.9%	98.9%
	16	19.2	33.8%	68.2%	74.0%	34.5%	83.2%	91.9%	38.1%	89.6%	92.3%
			53.7%	81.5%	83.5%	58.8%	94.6%	96.2%	75.3%	95.6%	96.2%
	20	24.0	3.3%	34.2%	42.8%	4.6%	36.6%	66.3%	10.0%	42.6%	66.5%
			17.4%	52.3%	57.9%	17.9%	65.2%	75.3%	31.0%	69.8%	75.6%
2400	6	14.4	63.2%	87.3%	90.8%	67.5%	91.3%	95.8%	78.3%	94.5%	95.8%
			78.2%	89.6%	92.3%	79.0%	94.2%	95.8%	93.0%	95.4%	95.8%
	8	19.2	32.8%	68.0%	78.7%	36.1%	76.3%	92.0%	50.4%	88.2%	92.0%
			52.6%	80.1%	84.0%	52.6%	88.4%	92.4%	84.0%	91.4%	92.4%
	10	24.0	10.4%	42.2%	57.0%	11.4%	45.9%	79.4%	14.15%	59.2%	80.7%
			21.5%	55.1%	68.0%	24.6%	62.6%	86.9%	53.1%	78.6%	87.7%

Kbps and ϕ=0.4, *Algorithm C* guarantees that 89.6% of users receives the minimum requested rate, whereas algorithms *B* and *A* lead to 83.2% and 68.2% of fully satisfied users, respectively. At higher rate requirements, but same mean aggregated rate per BS (i.e., 19.2 Mbps) and transport capacity factor, *Algorithm C* even achieves better performance, where 88.2% of users are fully satisfied in front of 76.3% and 68% for algorithms *B* and *A*, respectively. This is because, under the same aggregated rate, for a higher minimum rate requirement, less users can be allocated in the overall system and, the less the number of users supported per BS, the more important becomes the need to account for (radio and transport) load balancing schemes to properly distribute traffic among neighboring BSs.

Finally, notice that focusing on the percentage of users receiving at least 90% of the requested rate, similar trends are obtained but differences are less noticeable between algorithms *C* and *B* (2% and 3% for previous considered cases), but still quite significant compared to *Algorithm A* (above 10%).

6.7.4. Assessing Resource Consumptions

We have analyzed in previous sections the performance gain that can effectively be attained by *Algorithm C* with respect to algorithms *A* and *B*. We now examine in more detail how each algorithm impacts on the underlying radio and transport resource consumption. We consider a distribution of 8 users per cell, a transport capacity factor in the range $0.3 \leq \phi \leq 0.6$ and a minimum data rate requirement R^{min}=2400 Kbps. Figure 6.12 shows the mean value of BS radio and transport resource costs incurred by each algorithm over all obtained snapshots. We observe that, as expected, the performance gain achieved by *Algorithm C* is realized at the expenses of a higher usage of BS radio resources. Specifically, BS mean radio resource costs with *Algorithm C* are over 23% higher than with *Algorithm B* in the most restricted transport condition (i.e., ϕ=0.3).

Nevertheless, it is worthwhile to note that for ϕ=0.4 where, according to Figure 6.10, *Algorithm C* allows to accommodate up to 6 users/cell in front of 4 users/cell achieved by algorithms *A* and *B*), the mean radio resource cost of *Algorithm C* is only 4% higher than the other strategies (i.e., *Algorithm C* uses around 67.5% while the others 65%). So, *Algorithm C* is able to use this, otherwise unused, radio resources to wisely steer traffic and avoid transport limitations. As a result, a slightly higher transport resource utilization is obtained with *Algorithm C* as it leads to lower data rate degradation. Finally, performance gains achieved by *Algorithm C* are also tightly coupled with its capability to distribute traffic load in a smooth manner among BSs. This fact can be noticed in Figure 6.13 that presents the coefficient of variation of radio/transport resource costs (defined as

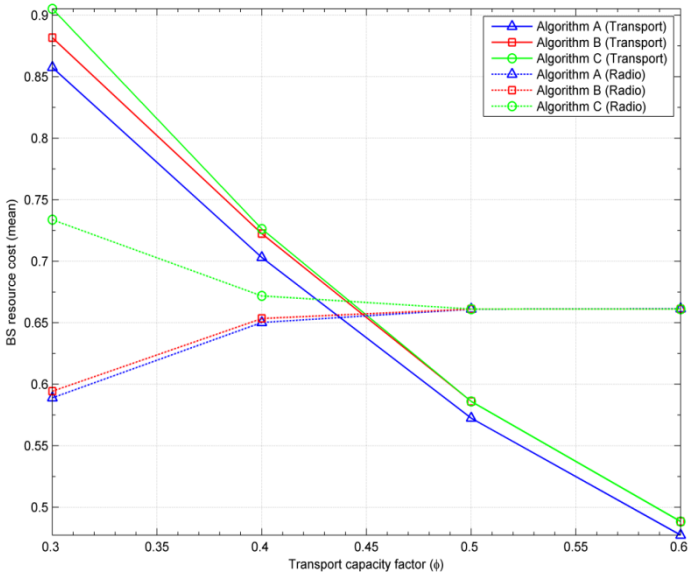


Figure 6.12: Mean of BS radio and transport resource costs for data rate requirement R^{min}=2400Kbps under different transport capacity conditions and a distribution of 8 users/cell.

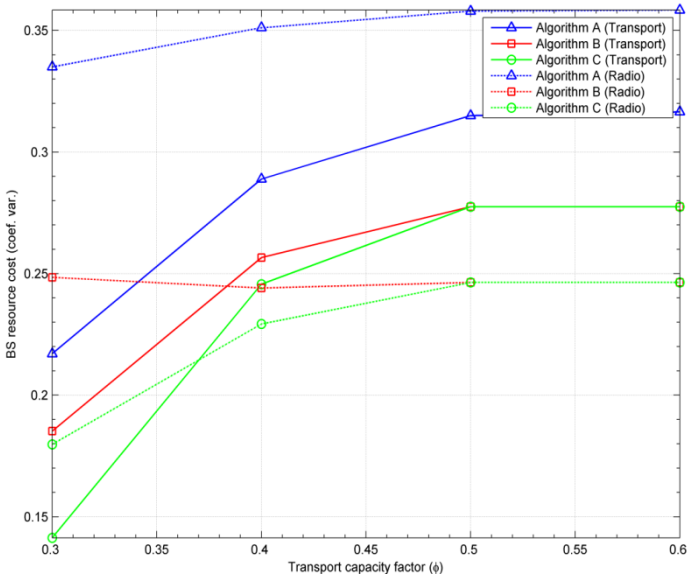


Figure 6.13: Coefficient of variation of BS radio and transport resource costs for data rate requirement R^{min}=2400Kbps under different transport capacity conditions and a distribution of 8 users/cell.

the ratio of standard deviation of radio/transport costs to the mean of radio/transport costs in all BSs). It can be seen that coefficients of variation of *Algorithm C* are always lower than those obtained by the other two algorithms.

6.7.5. Complexity of the Algorithm

As discussed in the complexity analysis given section 6.5.3, the time complexity of the algorithm in the worst case is $O(2N+3\cdot M\cdot n_i) + O(M^2(n_i)^3 + 2M^2(n_i)^2 + 3M\cdot n_i) + O(M^2(n_i)^3+M^2(n_i)^2+M\cdot n_i) + O(M\cdot n_i)$, where N is the total number of BSs in the system, M is the number of users, and n_i is the number of candidate BSs of each user i. The add phase's complexity is $O(M^2(n_i)^3+M2(n_i)^2+M\cdot n_i)$, while the complexity of the drop phase in the worst case is $O(M^2(n_i)^3 + 2M^2(n_i)^2 + 4M\cdot n_i)$. Over such a basis, in a configuration where N=19, M=76, and n_i=7, it can be

determined that the add and drop phases represents around 47% and 52%, respectively, of the total complexity of *Algorithm C*, while remaining corresponds to initialization and relaxation phases.

Furthermore, in the analyzed scenarios the add phase provides a performance enhancement of around 8% with respect to the drop phase in terms of the number of users that can be successfully assigned to a BS (i.e., this enhancement is obtained by comparing the total number of users assigned after the drop and add phases). Lastly, simulation results show that the number of iterations needed to solve a given snapshot fluctuates within a wide margin and are actually quite lower than the upper bound derived from the worst-case complexity analysis. As an example, in case of N=19, M=76, and n_i=7 the upper bound of the algorithm's complexity is 4.8×10^6. We have seen that the number of iterations required for the algorithm to converge is less than 800 iterations in the 95% of the computed snapshots.

6.8. Summary

In this chapter, a BS assignment algorithm developed for OFDMA-based networks has been evaluated. Unlike most of the existing mechanisms, it accounts for potential backhaul network constraints in the BS decision making process. This is motivated by the fact that the rollout of more spectral efficient air interface technologies along with mobile data and multimedia traffic increase are shifting the resource bottleneck from the air interface to the backhaul capacity in certain deployment scenarios. In this chapter we demonstrate that BS assignment strategies based exclusively on radio criteria do not suffice when backhaul capacity can become the bottleneck. Taking this into account, a BS assignment problem that considers both radio and backhaul constraints is formulated as an optimization problem and then mapped into a Multiple-Choice Multidimensional Knapsack Problem (MMKP). Over such a basis, a heuristic BS assignment algorithm is developed to solve the formulated problem. In simulation results we have shown that, in scenarios with limited transport capacity (i.e., scenarios where the transport capacity is less than half of the peak rate in the radio interface), the proposed algorithm brings up significant gains (i.e., around of) with respect to algorithms that are completely based on radio criteria in terms of the number of feasible BS assignment solutions it can determine, as well as in terms of the percentage of users that can be served guaranteeing their minimum bit rate constraints. Therefore, we claim that the proposed algorithm can be used to alleviate potential transport capacity restrictions in cellular system deployments.

7 Conclusions

7.1. Introduction

This chapter concludes this dissertation and is intended to summarize the main contributions of the thesis and also identify future related research possibilities to the work done.

The work presented in this thesis has been mainly focused on the incorporation of backhaul related metrics within a coordinated resource management framework in cellular networks. The motivation behind this goal is that the progressive enhancement of air interface technologies, along with the intensive demand of higher data rate services, are gradually shifting the resource bottleneck from the air interface towards the backhaul network in certain deployments of the cellular networks. Over such a basis, the thesis then concentrates on the evaluation of the base station (BS) assignment, or cell selection, problem as a means of properly distribute traffic among BSs considering both radio and transport network loads. This idea is firstly assessed by means of an analytical model, and secondly under computer simulation. This latter evaluation is performed under two different radio access technologies (RATs), one based on wideband code division multiple access (WCDMA) and a second one based on orthogonal frequency division multiple access (OFDMA). For each considered RAT scenario, we have conceived different algorithmic solutions to implement a cell selection strategy to account for backhaul load status. It has been demonstrated that BS assignment schemes that are exclusively based on radio-related aspects are not able to efficiently cope with backhaul capacity limitations. The next section summarizes the main contributions of the thesis.

7.2. Contribution

The first contribution of the thesis is the analysis of capacity requirements in the transport network of the radio access network (RAN) when IP is used as a transport technology. It has been discussed that although the inclusion of IP as a new transport technology in the RAN brings some cost savings to mobile operators, it also involves significant challenges to transport network in order to meet the stringent quality of service (QoS) requirements. This thesis contributes to this latter issue by conducting a detailed analysis of the bandwidth required in an IP-based transport network for guaranteeing the different delay requirements imposed by either the services or radio related functionalities in the RAN. The transport capacity requirements are evaluated considering the case of best-effort traffic, and the dimensioning approach we have followed is based on an over-provisioning solution. The capacity requirements are evaluated considering the UMTS terrestrial access network (UTRAN), under two different scenarios, that is, dedicated channels (DCHs) and high speed (HS) channels, and two services (voice and web-browsing traffic). It has been shown that transporting voice traffic in the RAN requires a less degree of over-provisioning of transport resources than in the case of web-browsing traffic (i.e., around 70% of over-

provisioning is required to support a mean voice traffic of 2 Mbps, and meet a delay requirement of 5 ms, whereas the extra capacity needed in the case of web-traffic is around 160% to support the same level of mean traffic). This is because the inherent characteristics of each type of traffic. Furthermore, in order to assess the influence of different parameters considered in the transport network and service type, a sensitivity analysis has been also carried out.

This thesis has also contributed to the field of resource management research by proposing a novel resource management framework that, besides radio criteria, it also considers backhaul criteria in the decision making process. The proposed framework, referred to as Coordinated Access Resource Management (CARM), defines some resource management functions that could account for both transport and air interface resources in a coordinated manner. The envisaged functions are: radio access technology (RAT) selection, bearer selection, admission control, congestion control, and cell selection. Another contribution of the thesis is the evaluation of one of the identified CARM functionalities, specifically an enhanced cell selection in a generic mobile network scenario with transport capacity limitations, regardless of the RAT.

In this part of the thesis, the performance evaluation methodology of the cell selection problem has been carried out by means an analytical model based on multi-dimensional Markov chains. This model is used to reproduce the behavior of three different cell selection strategies so that it is possible to determine the impact of transport capacity limitations on decisions taken by each strategy. Numerical results provided from this evaluation have shown that introducing transport status within the cell selection process provides a higher trunking gain, or capacity gain, in the utilization of transport network resources than a cell selection strategy whose underlying selection criteria is completely based on radio aspects only. We have also pointed out that the attained capacity gains achieved by the proposed strategy, referred to as transport prioritized cell selection (TP_CS), leads to a certain level of radio degradation (i.e., in terms of increased path loss per connection), but provided results also demonstrated that the TP_CS constitute a tradeoff between achieving a given capacity gain and reducing the amount of radio degradation.

From the initial performance analysis it has been demonstrated that a cell selection strategy that accounts for both radio and transport status is a suitable technique to cope, at some extent, with resource limitations in the backhaul. The next step in the workflow of the thesis has been to tackle the design of specific algorithms to implement the conceived cell selection, or base station (BS), strategy. To this end, we have analyzed the BS assignment problem under single RAT scenarios, and also based on the downlink performance as it is normally seen the most restrictive link due to the asymmetric bandwidth demand between downlink and uplink. In particular, we have considered two access technologies: WCDMA and OFDMA. The first considered RAT has been selected due to the fact that it is the basis of most of current mobile communication systems (e.g., UMTS). On the other hand, we also would like to turn out our attention to upcoming technologies for the future mobile networks, and hence a natural choice has been to consider one of the most promising access technologies: OFDMA.

Regarding the first RAT scenario, the performance analysis of the BS assignment problem has been addressed considering a single frequency WCDMA network, and assuming that each BS is constrained by amount of resources at the air interface and transport network. This latter constraint is modeled in terms of the air interface pole capacity. The capacity of the transport network, or more precisely the available transport capacity in a given BS, is defined in such a way that it is related to the downlink air interface pole capacity estimated in the planning process. Then, the BS assignment problem is formulated as an optimization problem where the aim is to maximize a given utility and meet radio and transport constraints, as well as data rate constraints of users. Over such a basis, an heuristic algorithm that is based on the simulated annealing technique is proposed to solve the problem. Using the simulated annealing-based algorithm, the BS assignment strategy conceived as part of the CARM framework (referred here as joint radio and transport, JRT) has been evaluated and compared to two reference (baseline) algorithms that rely exclusively on radio criteria (i.e., minimum path loss, MPL; and load balancing radio, LBR). The analyzed algorithms are intended to provide a solution to the BS assignment problem given a snapshot of the system.

The performance evaluation of the algorithms has been done considering a cellular network deployment under different distributions of users and transport network conditions. The main findings of the evaluation are the following:

- In scenarios of limited transport network capacity (i.e., when the transport capacity is lower than two times the air interface pole capacity) both MPL and LBR are able to achieve the considered network availability (i.e., a solution found in 95% of the performed snapshots) for a lower number of users in the system than in the case of the proposed JRT algorithm.

- For a given network conditions and data rate requirements of users, the proposed JRT algorithm is able to find a feasible BS assignment solution for a higher number of snapshots. A feasible solution is achieved when the BS assignment delivered by the algorithm do not exceeds the available radio and transport resources at BSs, and also all served users are provided with the data rate requested.

- In the most restricted transport network scenarios, where the available transport capacity at BSs corresponds to the air interface pole capacity, the proposed JRT approach that incorporates transport network status in the assignment process is able to increase the number of supported users about 88% with respect to MPL and LBR.

- The higher the data rate requested by users in the system, the higher the gain achieved by JRT over MPL and LBR, due that fewer users can be served per BS and hence traffic distribution mechanisms among different BSs in the scenario becomes more imperative.

- The performance gains achieved by the JRT approach come at the expense of a certain amount of radio degradation, in the sense that some users are allowed to be assigned to a BS other than the radio best server (e.g., the power increase in the uplink is less than 1.5 dB with respect to LBR and MPL). Such radio degradation can be limited by means of the maximum accepted path loss margin tolerated with respect to the best server.

- One of the main features of the proposed simulated-annealing algorithm is that it can be used to implement different BS assignment strategies by defining the proper utility function (used within the algorithm's logic) according to the desired assignment criteria of each strategy.

The last part of the thesis investigates the impact of possible transport resource limitations on the resource allocation decisions in an OFDMA-based broadband communication system. Unlike the WCDMA scenario, the task of resource allocation is a more complex that should take care of different aspects covering from the power and rate allocation per user in each individual subcarrier up to the assignment of the serving BS. From this part of the thesis, the main contributions are listed in the following:

- It has been proven that traditional BS assignment schemes exclusively ruled by radio criteria can fail to find a proper assignment without violating QoS requirements in situations where transport capacity can constitute the bottleneck, even there are enough network resources to accommodate all connections. This is proven by a simple counterexample with a two-cell scenario, where it is shown that relying only on radio parameters (e.g,, minimum path loss and radio load balancing strategy) to decide BS assignment is not able to come up with a "good" assignment (all connections served without any reduction in requested data rate) in a situation where it would be really possible to do that.

- The BS assignment problem in OFDMA is formulated as an optimization problem using utility and resource cost concepts, and mapped into a Multiple-Choice Multidimensional Knapsack Problem (MMKP), a well-known NP-hard combinatorial optimization problem arisen in many practical and real life problems. In order to solve the formulated MMKP optimization problem, a novel BS assignment algorithm that is based on the Lagrange multipliers has been derived.

- The proposed algorithm to solve the BS assignment problem considering is able to exploit the benefits of load balancing attending to both radio and transport load status. Furthermore, in situations where the transport network does not limit the achievable network capacity, the proposed algorithm would behave exactly as an algorithm aimed to perform the assignment of users based on the radio load of BSs.

- The evaluation of the proposed BS assignment algorithm has been carried out in cellular network scenarios with limited transport capacity. The proposed algorithm has been evaluated and compared to two of the most common BS assignment algorithms considered in cellular networks which are exclusively based on radio criteria (i.e., minimum path loss and radio load). Simulation results have demonstrated that, for a given network availability, the proposed algorithm is able to support a higher number of users without degrading the requested data service rate (e.g., a capacity gain of 75% and 100%, respect minimum path loss and radio load schemes in scenarios where the transport capacity is equal to the half of the BS peak rate, and user data rate requirements of 2400 Kbps).

- The provided analysis has also explored the impact on radio resources (downlink transmission power) of considering transport load within information in the BS decision-making process. The increase of the downlink transmission power is due to the fact that some users are not to connect to their best radio server, but this allows preventing congestion situations due to the shortage of transport resources.

- The implementation of a solution like the one discussed proposed in the thesis does not add excessive complexity to existing resource management functionality. In this regard, cell selection and control mechanisms already supported in cellular networks can be leveraged by introducing parameters related to transport load into their decision making processes. Complexity analysis addressed in this thesis shows that the number of iterations required for the algorithm to converge is less than 800 iterations in the 95% of the computed snapshots.

7.3. Future Work

The research conducted in this thesis has mainly revealed that the envisaged resource management scheme that account for both radio and transport status information to perform the BS assignment is able to efficiently cope with resource limitations in the transport network. Based on the presented analysis, the objective of this section is to detect the aspects of the work that can be improved, and also describe the evolution of the research line started with this thesis.

The different performance evaluations done in the context of WCDMA and OFDMA (see Chapters 5 and 6, respectively) has been carried out following a Monte-Carlo snapshot analysis, where each snapshot represents the state of the system in an instantaneous point in time. Although this is a widely used and accepted performance evaluation methodology, it does not capture some important features of the cellular networks such as the modeling of the system dynamics. Such dynamism has considerable impact on different aspects of the system, e.g., fast-fading (that leads to short-term variations of link gains between BSs and users) and mobility of users (that causes changes in the propagation environment and BS reassignment). From the simulation optimization viewpoint there is still room to improve the adopted methodology to evaluate the BS assignment algorithm so that the effects of system dynamics could be taken into account.

On the other hand, the evaluation of the BS assignment problem in both WCDMA and OFDMA networks has been addressed focusing on the downlink performance. As discussed in the thesis, we have focused on the downlink as it is normally seen as the most restrictive link due to the asymmetric bandwidth demand between uplink and downlink in nowadays networks. This assumption is aligned to most works on BS assignment problem for cellular networks found in the literature. However, since there could be scenarios where the uplink may become the bottleneck,

we consider that uplink capacity limitations due to either radio or transport load definitively deserve a further analysis in the future work.

The BS assignment algorithms studied in this thesis has been conceived as a centralized approach, which has been shown to be an efficient solution that allows us to quantify the benefits of our proposed "transport-aware" BS assignment strategy in scenarios with two different access technologies. However, due that next generation mobile networks are envisaged to deploy decentralized resource management functionalities, an enhancement of the work presented in this thesis consist in addressing the development of a distributed implementation of the algorithm proposed for OFDMA-based cellular networks. It is foreseen, for instance, that a distributed implementation of the proposed algorithm would be based on a pricing scheme, intended to be independently executed at each BS.

Bibliography

[1] 3GPP TR 25.913, Requirements for evolved UTRA (E-UTRA) and Evolved UTRAN (E-UTRAN).

[2] J. G. Andrews, A. Ghosh, R. Muhamed "Fundamentals of WiMAX: Understanding Broadband Wireless Networking", Prentice Hall, 2007.

[3] S. Pietrzyk, "OFDMA for Broadband Wireless Access", Artech House, 2006.

[4] R. van Nee, R. Prasad, ―OFDM for Wireless Multimedia Communications, Artech House, 2000.

[5] R. Fantacci, D. Marabissi, D. Tarchi, I. Habib, "Adaptive Modulation and Coding Techniques for OFDMA Systems" in IEEE Transactions on Wireless Communications, vol. 8, no. 9, pp. 4876-4883, September 2009.

[6] A. Goldsmith, S. Jafar, N. Jindal, and S. Vishwanath, "Capacity Limits of MIMO Channels," in IEEE Journal of Selected Areas in Communications, vol. 21, no. 5, pp. 684–702, June 2003.

[7] S. Chia, M. Gasparroni, and P. Brick, "The Next Challenge for Cellular Networks: Backhaul", IEEE Microwave Magazine, pp. 54-66, Aug. 2009.

[8] R. Nadiv, T. Naveh, "Wireless Backhaul Topologies: Analyzing Backhaul Topology Strategies", Ceragon White Paper, August 2010.

[9] S. Little, "Is Microwave Backhaul Up to the 4G Task?", IEEE Microwave Magazine, pp. 67-74, Aug. 2009.

[10] http://www.openmobilealliance.org/

[11] MWIF "IP in the RAN as Transport Option in 3rd Generation Mobile Systems", MTR-W-5 Technical Report. Rel. v2.0.0. June 2001.

[12] J. D. Vriendt, P. Lainé, C. Lerouge, X. Xu, "Mobile Network Evolution: A Revolution on the Move", IEEE Communications Magazine, vol. 40, no. 4, April 2002.

[13] P. Newman, "In Search of the All-IP Mobile Network", in IEEE Radio Communications, vol. 42, no. 12, pp. s3-s8, December 2004.

[14] F. M. Chiussi, D. A. Khotimsky, S. Krishnan, "Mobility Management in Third-Generation All-IP Networks", IEEE Communications Magazine, vol. 40, no. 9, pp. 124-135, September 2002.

[15] G. Eneroth, G. Fodor, G. Leijonhufvud, A. Racz, I. Szabo, "Applying ATM/AAL2 as a Switching Technology in Third Generation Mobile Access Networks" in IEEE Communications Magazine, vol. 37, no. 6, pp. 112-123, 1999.

[16] 3GPP, TR 25.933 v5.4.0, "IP Transport in the UTRAN (Release 5)".

[17] A. Samhat, T. Chahed, G. Hébuterne, "Transport in UMTS Radio Access Network: IP versus AAL2/ATM", Proceedings of IEEE WCNC 2004, March 2004.

[18] A. Samhat, T. Chahed, "Modeling and Analysis of Transport of Voice and Data in the UMTS Radio Access Network: IP versus AAL2/ATM", Proceedings of IEEE WCNC, March 2004.

[19] C. Yong Jung, et al, "Performance Comparison of ATM and IP based Transmission Schemes in the UTRAN", Proceedings of IEEE WCNC, March 2004.

[20] 3GPP, TS 25.401, “UTRAN overall description”.
[21] Y. Guo, Z. Antoniu, S. Dixit, "IP Transport in 3G Radio Access Networks: an MPLS-based approach", Proceedings of IEEE WCNC, March 2002.
[22] K. Venken, I. G. Vinagre, J. De Vriendt, “Analysis of the Evolution to an IP-based UMTS Terrestrial Radio Access Network”, IEEE Wireless Communications, Oct. 2003.
[23] A. Samhat, T. Chahed, “Service Differentiation between Voice and Data Traffic in the IP-based Radio Access Network”, Proceedings of IEEE VTC, May 2004.
[24] G. Toth, C. Antal, “On Packet Delays and Segmentation Methods in IP Based UMTS Radio Access Network”, Second Int. Workshop on QoS in Multiservice IP Networks, QoS-IP 2003.
[25] P. Bosch, L. Samuel, S. Mullender, P. Polakos, G. Rittenhouse, “Flat Cellular (UMTS) Networks”, Proceedings of IEEE WCNC, March 2007.
[26] S. R. Sherif, G. Ellinas, A. Hadjiantonis, R. Dorsinville, M. A. Ali, “On the Merits of Migrating From Legacy Circuit-Switched Cellular Infrastructure to a Fully Packet-Based RAN Architecture”, Journal of Lightwave Techology, vol. 27, no. 12, June 2009.
[27] R. A. Junquera, L. Ledesma, “Evolución del Backhaul Celular, El futuro de las Redes Celulares”, available online on http://telesemana.com/reportes, July 2009.
[28] Qualcomm, “UMTS/HSDPA Backhaul Bandwidth Dimensioning”, May 2007.
[29] M. Howard, “Mobile Backhaul Moves to the Forefront”, Special Supplement to Telecommunications Magazine, March 2007.
[30] A. Bolle and H. Herbertsson, “Backhaul must make room for HSDPA”, Wireless Europe, no. 43, pp. 17-18, March 2006.
[31] E. Boch, “High Capacity Ethernet Backhaul Radio Systems for Advanced Mobile Data Networks”, in IEEE Microwave Magazine, vol. 10, no. 5, pp. 108-114, August 2009.
[32] NGMN Alliance, “Optimised Backhaul Requirements”, White Paper v3.0, Aug. 2008, [Online]. Available: http://www.ngmn.org.
[33] K. Ayagari and J. Tang, “Backhaul Solutions for 3G Networks”, Alcatel-Lucent White Paper, 2007.
[34] J. Lakkakorpi and A. Sayenko, “Backhaul as a Bottleneck in IEEE 802.16e Networks”, in Proceedings of the IEEE Global Communications Conference (GLOBECOM), 30 Nov.-4 Dec. 2008.
[35] S. Nádas, S. Rácz, Z. Nagy, and S. Molnár, “Providing Congestion Control in the Iub Transport Network for HSDPA”, in Proc. Global Commun. Conf. (GLOBECOM), 26-30 Nov. 2007.
[36] 3GPP TR 25.902 V7.1.0, “Iub/Iur congestion control” (Release 7), March 2007.
[37] J. Lakkakorpi, “Simple measurement-based admission control for DiffServ access networks”, in Proc. ITCom 2002, July 2002, Boston, SPIE.
[38] K. Nichols, V. Jacobson, L. Zhang, “A two-bit differentiated service architecture for the Internet”, IETF RFC 2638, July 1999.
[39] S. Blake, D. Blak, M. Carlson, E. Davies, Z. Wang, and W. Weiss, "An architecture for Differentiated Services", IETF RFC 2475, 1998.
[40] B. Davie, A. Charny, J.C.R. Bennett, K. Benson, J.Y. Le Boudec, W. Courtney, S. Davari, V. Firoiu and D. Stiliadis, “An Expedited Forwarding PHB,” RFC 3246, Mar. 2002.
[41] J. Heinänen, F. Baker, W. Weiss and J. Wroclawski, “Assured Forwarding PHB Group,” RFC 2597, Jun. 1999.
[42] G. Heijenk, G. Karagiannis, V. Rexhepi and L. Westberg, “DiffServ Resource Management in IP-based Radio Access Networks”, Proceedings of WPMC, Sept. 2001.
[43] H. el Allali, G. Heijenk, "Resource management in IP-based Radio Access Networks", Proceedings CTIT Workshop on Mobile Communications, Feb. 2001.
[44] H. el Allali, “A Measurement Based Admission Control Algorithm for Resource Management in DiffServ IP Network”, Proceedings of PIMRC 2006.

[45] S. K. Kasera, et al, “Congestion Control Policies for IP-based CDMA Radio Access Networks”, IEEE Transactions on Mobile Computing, July-August 2005.
[46] IST-AROMA project web page http://www.aroma-ist.upc.edu/
[47] M. Sagfors, V. Virkki, T. Kuningas, “Overload Control of Best-Effort Traffic in the UTRAN Transport Network”, in Proceedings of the IEEE Vehicular Technology Conference, (VTC 2006-Spring), May 2006.
[48] C. Saraydar, S. Sampath, S. Abraham, M. Chuah, "Enhanced capacity on the lub link through rate control: The single service case", in Proceedings of the IEEE Int Conference on Communications, 2003.
[49] C. Saraydar, S. Abraham, M. Chuah, "Impact of rate control on the capacity of an lub link: Multiple service case", IEEE WCNC 2003.
[50] S. Nadas, S. Racz, Z. Nagy, S. Molnar, “Providing Congestion Control in the Iub Transport Network for HSDPA”, in Proceedings of the Global Telecommunications Conference (GLOBECOM), November 2007.
[51] A. Samhat, T. Chahed, “On the QoS-capable transport in the IP-based UTRAN”, 2nd International Symposium on Wireless Communication Systems, pp. 318-321, Sept. 2005.
[52] 3GPP TS 25.430 v7.0.0, “UTRAN Iub interface: general aspects and principles (Release 7)”
[53] The Business Case for GSM and UMTS Backhaul Optimization, www.cisco.com/go/mobile, 2005.
[54] C. Fraleigh, F. Tobagi, C. Diot, "Provisioning IP Backbone Networks to Support Latency Sensitive Traffic", IEEE INFOCOM 2003, San Francisco, March 2003.
[55] H. Holma and A. Toskala, WCDMA for UMTS, Radio Access for Third Generation Mobile Communications, John Wiley and Sons, 2004.
[56] 3GPP, TR 25.853 v3.0.0, “Delay Budget within the Access Stratum”.
[57] 3GPP TR 25.912 v7.0.0, “Feasibility study for evolved Universal Terrestrial Radio Access (UTRA) and Universal Terrestrial Radio Access Network (UTRAN) (Release 7)”.
[58] OPNET Modeler, http://www.opnet.com/solutions/network_rd/modeler.html
[59] S. Casner, V. Jacobson, “Compressing IP/UDP/RTP Headers for Low-Speed Serial Links”, RFC 2508, February 1999.
[60] M. E. Crovella, A. Bestavros, “Self-Similarity in World Wide Web Traffic: Evidence and Possible Causes”, in IEEE/ACM Trans. on Networking, vol. 5, no. 6, pp. 835-846, December 1997.
[61] X. Pérez-Costa, K. Heinze, A. Banchs, S. Sallent, “Analysis of Performance Issues in an IP-based UMTS”, Proceedings of the Eighth ACM MSWiM, October 2005.
[62] Mooi C. Chuah and Enrique J. Hernandez-Valencia, “A LightWeight IP Encapsulation (LIPE) Scheme”, Internet draft document, draft-chuah-avtlipe-02.txt, IETF, Dec. 2000.
[63] R. Pazhyannur, I. Ali, C. Fox, “PPP Multiplexing”, IETF RFC 3153, August 2001.
[64] D. De Vleeschauwer, et al, “Closed-Form Formula to Calculate the Dejittering Delay in Packetised Voice Transport”, in Proceedings of the IFIP-TC6 / European Commission International Conference Networking, May 2000.
[65] 3GPP TS 23.107, Quality of Service (QoS) concept and architecture (Release 6).
[66] 3GPP TS 23.207, End-to-end Quality of Service (QoS) concept and architecture (Release 6).
[67] IST-EVEREST project web page http://www.everest-ist.upc.es
[68] J. Lakkakorpi, O. Strandberg, J. Salonen, “Adaptive connection admission control for differentiated services access networks”, in Proceedings of the IEEE Global Communications Conference (GLOBECOM), 29 Nov.-3 Dec. 2004.
[69] E. Mykoniati, C. Charalampous, P. Georgatsos, T. Damilatis, D. Goderis, P. Trimintzios, G. Pavlou, D. Griffin, “Admission control for providing QoS in DiffServ IP networks: the TEQUILA approach”, in IEEE Communications Magazine, Januany 2003
[70] P. Trimintzios, I. Andrikopoulos, G. Pavlou, P. Flegkas, D. Griffin, P. Georgatsos, D. Goderis, Y. T’Joens, L. Georgiadis, C. Jacquenet, R. Egan, “A management and control

architecture for providing IP differentiated services in MPLS-based networks", in IEEE Communications Magazine, May 2001.
[71] 3GPP TS 36.300: "E-UTRA and E-UTRAN Overall description - Stage 2".
[72] 3GPP TS 25.331, "Radio Resource Control (RRC); Protocol Specification (Release 7)", available online on http://www.3gpp.org.
[73] R. D. Yates, and C. Y. Huang, "Integrated Power Control and Base Station Assignment", IEEE Trans. on Vehicular Technology, vol. 44, no. 3, Aug. 1995.
[74] F. R. Farrokhi, K. J. Ray Liu, "Downlink Power Control and Base Station Assignment", IEEE Communications Letters, vol. 1, no. 4, July 1997.
[75] J. W. Lee, R. R. Mazumdar, N. B. Shroff, "Joint Resource Allocation and Base-Station Assignment for the Downlink in CDMA Networks", IEEE/ACM Trans. Netw., vol. 14, no. 1, pp. 1 – 14, Feb. 2006.
[76] K. Sipila, K. C. Honkasalo, J. L. Steffens, A. Wacker, "Estimation of Capacity and Required Transmission Power of WCDMA Downlink Based on a Downlink Pole Equation", in Proc. of IEEE Vehicular Technology Conference (VTC 2000), May 2000.
[77] J. P. Castro, The UMTS Network and Radio Access Technology Air Interface Techniques for Future Mobile Systems, John Wiley and Sons, 2001.
[78] M. Xiao, N. B. Shroff, and E. P. K. Chong, "Utility-Based Power Control in Cellular Wireless Systems", in Proc. of IEEE INFOCOM, 2001.
[79] S. Kirkpatrick, C. D. Gelatt Jr., M. P. Vecchi, "Optimization by Simulated Annealing", Science, vol. 220, no. 4598: 671-679, 1983.
[80] C. C. Ribeiro, S. L. Martins, I. Rosseti, "Metaheuristics for Optimization Problems in Computer Communications" Computer Networks, vol. 30, no. 4, pp. 656-69, Feb. 2007.
[81] F. Busetti, "Simulated annealing overview", 2003.
[82] F. Sallabi, A. Lakas, K. Shuaib, and M. Boulmalf, "WCDMA Downlink Simulator with Efficient Wrap-Around Technique", Proc. of Int. Conf. on Wireless and Optical Communications Networks (WOCN 2005), March 2005.
[83] Guoqing Li and Hui Liu, "Downlink Radio Resource Allocation for Multi-Cell OFDMA System", IEEE Trans. Wireless Commun., vol. 5, no. 12, pp. 3451-3459, Dec. 2006.
[84] J. Zander, "Radio Resource Management in Future Wireless Networks: Requirements and Limitations", IEEE Commun. Magazine, vol. 35, no. 8, pp. 30-36, Aug. 1997.
[85] S. Pietrzyk and G. J. M. Janssen, "Radio Resource Allocation for Cellular Networks Based on OFDMA with QoS Guarantees", in Proc. Global Commun. Conf. (GLOBECOM), 23 Nov. - 3 Dec. 2004.
[86] C. U. Saraydar, N. B. Mandayam, and D. J. Goodman, "Pricing and Power Control in a Multicell Wireless Data Network", IEEE J. Sel. Areas Commun., vol. 19, no. 10, pp. 1883-1892, Oct. 2001.
[87] V. K. N. Lau, "On the Macroscopic Optimization of Multicell Wireless Systems with Multiuser Detection and Multiple Antennas - Uplink Analysis", IEEE Trans. on Wireless Commun., vol. 4, no. 4, pp. 1388 – 1393, July 2005.
[88] D. Amzallag, R. Bar-Yehuda, D. Raz, G. Scalosub, "Cell Selection in 4G Cellular Networks", in Proc. of INFOCOM, 13-18 April 2008.
[89] Y. J. Zhang and K. Ben Letaief, "Multiuser Adaptive Subcarrier-and-Bit Allocation With Adaptive Cell Selection for OFDM Systems", IEEE Trans. Wireless Commun., vol. 3, no. 5, pp. 1566 – 1575, Sept. 2004.
[90] I. N. Stiakogiannakis, D.A. Zarbouti, G.V. Tsoulos, and D.I. Kaklamani, "Subcarrier Allocation Algorithms for Multicellular OFDMA Networks without Channel State Information", in Proc. 3rd Int. Symposium of Wireless Pervasive Computing (ISWPC), 7-9 May 2008.
[91] C. Koutsimanis and G. Fodor, "A Dynamic Resource Allocation Scheme for Guaranteed Bit Rate Services in OFDMA Networks", in Proc. of IEEE Int. Conference on Communications (ICC), 19-23 May, 2008.

[92] C. Wengerter, J. Ohlhorst, and A. G. E. von Elbwart, "Fairness and Throughput Analysis for Generalized Proportional Fair Frequency Scheduling in OFDMA", in Proc. of Semi-annual IEEE Vehicular Tech. Conf., 30 May-1 June, 2005.

[93] L. Jorguseski, T. M. H. Le, E. Fledderus, and R. Prasad, "Downlink Resource Allocation for Evolved UTRAN and WiMAX Cellular Systems", in Proc. of Int. Symp. on Personal Indoor and Mobile Radio Commun. (PIMRC), 15-18 Sept. 2008.

[94] A. Racz, A. Temesvary, N. Reider, "Handover Performance in 3GPP Long Term Evolution (LTE) Systems", in Proc. of IST Mobile and Wireless Commun. Summit, 2007.

[95] 3GPP, TR 36.913, "Requirements for further advancements for E-UTRA (LTE-Advanced".

[96] M. Sawahashi, Y. Kishiyama, A. Morimoto, D. Nishikawa, M. Tanno, "Coordinated Multipoint Transmission/Reception Techniques for LTE-Advanced", in IEEE Wireless Communications, vol. 17, no. 3, pp. 26-34, June 2010.

[97] W. H. Kuo and W. Liao, "Utility-based Radio Resource Allocation for QoS Traffic in Wireless Networks", IEEE Trans. Wireless Commun., vol. 7, no. 7, pp. 2714 – 2722, July 2008.

[98] H. Kellerer, U. Pferschy, D. Pisinger, Knapsack Problems. Springer-Verlag, 2004.

[99] K. Navaie and H. Yanikomeroglu, "Optimal Downlink Resource Allocation for Non-real Traffic in CDMA/TDMA Networks", IEEE Commun. Lett., vol. 10, no. 4, April 2006.

[100] G. Chen, S. Khan, K. F. Li, E. Manning, "Building an adaptive multimedia system using the utility model", in Proc. of Int. Workshop on Parallel and Distributed Realtime Systems, San Juan, Puerto Rico, 1999.

[101] A. Sbihi, "A best first search exact algorithm for the Multiple-choice Multidimensional Knapsack Problem", Journal of Combinatorial Optimization, vol. 3, no. 4, pp. 337-351, December 2007.

[102] M. Michrafy, M. Hifi, and A. Sbihi, "Heuristic Algorithms for the Multiple-choice Multidimensional Knapsack Problem", Journal of the Operational Research Society, vol. 55, no 12, pp. 1323-32, Dec. 2004.

[103] M. Moser, D. P. Jokanovic, and N. Shiratori, "An Algorithm for the Multidimesional Multiple-choice Knapsack Problem", IEICE Trans. Fundam. Electron., Commun., Comp. Sci., vol. E80-A, no. 3, March 1997.

[104] H. Everett III, "Generalized Lagrange Multiplier Method for Solving Problems of Optimum Allocation of Resources", Operations Research, vol. 11, pp. 399-417, 1963.

[105] M. Katoozian, K. Navaie, and H. Yanikomeroglu, "Optimal Utility-Based Resource Allocation for OFDM Networks with Multiple Types of Traffic", in Proc. of Semi-annual IEEE Vehicular Tech. Conf., 11-14 May 2008.

[106] K. Ramadas and R. Jain, WiMAX System Evaluation Methodology. Tech. Report, Wimax Forum, Dec. 2007.

[107] R. D. Yates, "A Framework for Uplink Power Control in Cellular Radio Systems", IEEE J. Sel. Areas Commun., vol. 13, no. 7, pp. 1341-1347, Sep. 1995.