Departament de Teoria
del Senyal i Comunicacions

UNIVERSITAT POLITÈCNICA DE CATALUNYA

Ph.D. Thesis

ADAPTIVE RADIO RESOURCE MANAGEMENT FOR OFDMA-BASED MACRO- AND FEMTOCELL NETWORKS

Thesis by

Emanuel Bezerra Rodrigues

In Partial Fulfillment of the
Requirements for the Degree of
Doctor from the Universitat Politècnica de Catalunya

Advisor: Prof. Fernando Casadevall Palacio
Departament de Teoria del Senyal i Comunicacions (TSC)
Universitat Politècnica de Catalunya (UPC)

Barcelona, May 28, 2011

With love to my wife Gladia,
and to our little baby.

Abstract

User and cellular operator requirements and expectations have been continuously evolving, and consequently, advanced radio access technologies have emerged. The International Mobile Telecommunications - Advanced (IMT-Advanced) specifications for mobile broadband Fourth Generation (4G) networks state, among other requirements, that enhanced peak data rates of 100 Mbps and 1 Gbps for high and low mobility should be provided. In order to achieve this challenging performance, Orthogonal Frequency Division Multiple Access (OFDMA) has been chosen as the access technology, and femtocells have been considered for improving indoor coverage.

In order to fully explore the flexibility of these technologies and use the scarce radio resources in the most efficient way possible, intelligent and adaptive Radio Resource Management (RRM) techniques are crucial. There are many open RRM problems in wireless networks in general, and OFDMA-based cellular systems in particular. One of such problems is the fundamental trade-off that exists between efficiency in the resource usage and fairness in the resource distribution among network players.

Several opportunistic RRM algorithms, which dynamically allocate the resources to the network players that present the highest efficiency indicator with regard to these resources, have been proposed to maximize the efficiency in the resource usage. The trade-off between efficiency and fairness appears when the resources have different efficiency indicators to different network players (multi-user or multi-cell diversity). The use of opportunistic resource allocation to explore these diversities causes unfair situations in the resource distribution. On the other hand, schemes that provide absolute fairness deal with the worst case scenario, penalizing players with better condition and reducing the system capacity.

In this thesis, several RRM policies and techniques are proposed to balance this compromise in macrocell and femtocell networks. In the particular case of macrocell systems, we propound a new network management paradigm based on the control of a cell fairness index in scenarios with Non-Real Time (NRT) or Real Time (RT) services. Two fairness control approaches are studied: instantaneous (short-term) control by means of generalized fairness/rate adaptive RRM techniques and average (mid-term) control using utility-based frameworks. For femtocell networks, a novel interference avoidance technique able to balance the trade-off between spectral efficiency in the femtocell tier and fairness among the Femtocell Access Points (FAPs) is formulated. This RRM

strategy is based on a high-level, mid/long-term frequency planning that takes into account the topology of groups of neighboring FAPs.

The RRM techniques considered in this thesis are evaluated by means of extensive system-level and/or numerical simulations. Regarding the macrocell scenario, it is shown that the proposed adaptive RRM techniques are valuable tools for the mobile operators, because they are generalizations of well-known classic strategies found in the literature and they can effectively guarantee different fairness levels in the system and control the trade-off between efficiency and fairness. Furthermore, it is concluded that the utility-based strategies that perform an average fairness control can provide performance results as good as the fairness/rate adaptive techniques, which are based on instantaneous optimization, using less computational resources. Finally, it is demonstrated that the proposed interference avoidance technique for femtocell networks can guarantee a seamless coexistence between neighboring FAPs in any interference topology. Furthermore, this technique can be implemented in both centralized and distributed network architectures and generates very low signaling overhead.

Acknowledgements

Firstly and above all, I would like to thank God for giving me health and all the internal resources needed for the fulfilment of this doctoral endeavor.

All my gratitude to my beloved wife Gladia Lima, who has always supported and encouraged me. Due to her love and patience, this dream has come true.

Many thanks to my family, specially my mom Mércia Bezerra, who raised me with love and helped me to form the human being I am now.

I would like to thank very much my advisor, Prof. Fernando Casadevall, for receiving me with open arms in Barcelona, and for his commitment and generosity in sharing his wisdom. He has been an inspiration in both personal and professional aspects.

My acknowledgments to all the professors and employees of the Department of Signal Theory and Communications (TSC), specially Prof. Jordi Pérez-Romero, for the rich technical discussions and friendship.

I would also like to thank the Network of Excellence in Wireless COMmunications++ (NEWCOM++), which has been a very important forum of academic partnership. Many thanks to Prof. Marco Moretti and Pawel Sroka, who have collaborated with this research.

I am very grateful for the financial support of the Brazilian government by means of the scholarship of the Improvement Co-ordination of Superior Level People (CAPES).

Many thanks to Prof. Rodrigo Cavalcanti, Prof. João César Mota and Gunnar Bark for their recommendation letters. I am also specially grateful for the guidance of Prof. Rodrigo Cavalcanti, who has closely followed the progress of my doctoral work.

I would like to acknowledge the work of the anonymous reviewers of this thesis and the papers submitted to different journals, magazines and conferences. Their valuable comments helped to improve the quality of this work.

Last, but not least, many thanks to the friends I have known here in Barcelona, specially my colleagues at the Mobile Communications Research Group (GRCM), for their help and friendship.

Contents

List of Figures

List of Tables

List of Algorithms

List of Abbreviations

3GPP	3rd. Generation Partnership Project
3GPP2	3rd. Generation Partnership Project 2
3G	Third Generation
4G	Fourth Generation
ADF	Adaptive Delay-Based Fairness
AFR	Available Frequency Reutilization
AMC	Adaptive Modulation and Coding
APA	Adaptive Power Allocation
ASA	Adaptive Satisfaction-Based Allocation
ATF	Adaptive Throughput-Based Fairness
BER	Bit Error Rate
BnB	Branch and Bound
BSC	Base Station Controller
CDF	Cumulative Distribution Function
CDMA	Code Division Multiple Access
CFI	Cell Fairness Index
CFT	Cell Fairness Target
CNR	Channel-to-Noise Ratio
CoMP	Coordinated Multi-Point
CPU	Central Processing Unit
CSI	Channel State Information
DAS	Distributed Antenna System

DFP	Dynamic Frequency Planning
DSA	Dynamic Sub-carrier Assignment
DSL	Digital Subscriber Line
EI	Efficiency Indicator
EPA	Equal Power Allocation
EV-DO	Evolution-Data Optimized
FAP	Femtocell Access Point
FDD	Frequency Division Duplex
FER	Frame Erasure Rate
FFR	Fractional Frequency Reuse
FFT	Fast Fourier Transform
FI	Fairness Indicator
FIFO	First-In-First-Out
FMMR-P	Fairness-Based Max-Min Rate with Proportional Rate Constraints
FSA	Femtocell Sub-carrier Allocation
FSQP	Feasible Sequential Quadratic Programming
FSRM	Fairness-Based Sum Rate Maximization
FSRM-P	Fairness-Based Sum Rate Maximization with Proportional Rate Constraints
FTP	File Transfer Protocol
GA	Genetic Algorithm
GSM	Global System for Mobile communication
HDR	High Data Rate
HOL	Head-Of-Line
HSDPA	High Speed Downlink Packet Access
HSPA	High Speed Packet Access
ICIC	Inter-Cell Interference Coordination
ID	Identification Number
IEEE	Institute of Electrical and Electronics Engineers
IMT-2000	International Mobile Telecommunications for the year 2000

IP	Internet Protocol
IPsec	IP Security
ISI	Inter-Symbol Interference
ITU	International Telecommunications Union
LMMFA	Layered Max-Min Frequency Allocation
LOS	Line of Sight
LTE	Long Term Evolution
LTE-Advanced	Long Term Evolution-Advanced
MAC	Medium Access Control
M-ADF	Modified Adaptive Delay-Based Fairness
M-ATF	Modified Adaptive Throughput-Based Fairness
MBS	Macrocell Base Station
MCS	Modulation and Coding Scheme
MFR	Max Frequency Reutilization
MIMO	Multiple Input Multiple Output
M-LWDF	Modified Largest Weighted Delay First
MME	Mobility Management Entity
MMF	Max-Min Fairness
MMFA	Max-Min Frequency Allocation
MMR	Max-Min Rate
MMR-P	Max-Min Rate with Proportional Rate Constraints
MR	Max-Rate
MT	Mobile Terminal
NRT	Non-Real Time
OFDM	Orthogonal Frequency Division Multiplexing
OFDMA	Orthogonal Frequency Division Multiple Access
PAPR	Peak to Average Power Ratio
PDF	Probability Density Function
PF	Proportional Fairness

PRB	Physical Resource Block
PSC	Packet Scheduling
QAM	Quadrature Amplitude Modulation
QoS	Quality of Service
QPSK	Quadrature Phase Shift Keying
RAT	Radio Access Technology
RRA	Radio Resource Allocation
RRM	Radio Resource Management
RT	Real Time
RTT	Round Trip Time
SA	Simulated Annealing
SC-FDMA	Single Carrier Frequency Division Multiple Access
SFAM	Sum Frequency Allocation Maximization
SINR	Signal-to-Interference plus Noise Ratio
SNR	Signal-to-Noise Ratio
SOHO	Small Office and Home Office
SRM	Sum Rate Maximization
SRM-P	Sum Rate Maximization with Proportional Rate Constraints
TCP	Transport Control Protocol
TDMA	Time Division Multiple Access
TTI	Transmission Time Interval
TU	Typical Urban
UFI	User Fairness Index
UMTS	Universal Mobile Telecommunications System
UTRA	Universal Terrestrial Radio Access
VoIP	Voice over IP
WCDMA	Wideband CDMA
WiMAX	Worldwide Interoperability for Microwave Access
WLAN	Wireless Local Area Network
WWW	World Wide Web

List of Symbols

$\mathbf{A}$	Interference topology matrix
$\mathcal{A}_n$	Subset of available sub-carriers for the nth FAP
b_j^{hol}	Number of bits in the Head-Of-Line packet
$\mathbf{B}$	Interference grouping matrix
$c_{j,k}[n]$	Achievable transmission efficiency of the kth sub-carrier with respect to the jth user
d	Separation distance between the macrocell base station and the user
d_j^{hol}	Head-Of-Line delay of the jth user
$d_j^{\text{hol,filt}}$	Filtered Head-Of-Line delay of the jth user
d_j^{req}	Delay requirement of the jth user
f_{delay}	Filtering constant used in the Head-Of-Line delay average
f_{thru}	Filtering constant used in the throughput average
F_n	Number of frequency resources allocated to the nth FAP
F_n^{prev}	Vector of the number of frequency resources allocated to the predecessors of the nth FAP
F_n^{next}	Vector of the bounded predictions of the number of frequency resources allocable to the succeeders of the nth FAP
$F_n^{\max}$	Maximum number of frequency resources available for the nth FAP that assures perfect interference avoidance
G_m	Set of indexes of the FAPs belonging to a given group of mutual interfering FAPs
J	Total number of users
K	Total number of sub-carriers
L	Number of layers in a given cluster of FAPs
L_j^{path}	Distance-dependent path loss for the jth user
L_j^{shadow}	Propagation loss due to shadowing for the jth user
$L_{j,k}^{\text{fast}}$	Propagation loss due to fast fading for the jth user in the kth sub-carrier
M	Total number of possible groups of mutual interfering FAPs
N	Number of FAPs in a given cluster
$\mathcal{O}_n$	Feasible options of DFP allocation for the nth FAP

p_k	Transmit power of the kth sub-carrier
$p_{j,k}^{\mathrm{rx}}$	Received power of the jth user in the kth sub-carrier
$\overline{p_{j,k}^{\mathrm{rx}}}$	Mean received power of the jth user in the kth sub-carrier
p^{noise}	Noise power
P_{total}	Macrocell base station total transmit power
R_j	Transmission rate of the jth user
R_{cell}	Total cell rate
$\mathcal{S}$	Set of all sub-carriers in the system
$\mathcal{S}_j$	Subset of sub-carriers assigned to the jth user
$\mathcal{S}_n$	Subset of sub-carriers allocated to the nth FAP
t_{tti}	Duration of the Transmission Time Interval (TTI)
T_j^{req}	Throughput requirement of the jth user
$U_j(\cdot)$	Utility function of the jth user
V_l	Set of indexes of the FAPs that form the lth layer
w_j^{nrt}	Utility-based weight for non-real time services
w_j^{rt}	Utility-based weight for real time services
$\mathcal{Z}$	Cluster priority list (set of indexes of all FAPs in order of priority)
$\mathcal{Z}_n^{\mathrm{prev}}$	Set of indexes of all predecessors of the nth FAP
$\mathcal{Z}_n^{\mathrm{next}}$	Set of indexes of all succeeders of the nth FAP
$\mathcal{Z}_{n,\mathrm{direct}}^{\mathrm{next}}$	Set of indexes of the succeeders of the nth FAP that interfere directly with it
$\mathcal{Z}_{n,\mathrm{indirect}}^{\mathrm{next}}$	Set of indexes of the succeeders of the nth FAP that do not interfere with it
α	Fairness controlling parameter of the utility-based alpha-rule framework
β	Fairness controlling parameter of the utility-based beta-rule framework
$\gamma_{j,k}$	Channel-to-noise ratio of the kth sub-carrier with respect to the jth user
Γ	SNR gap
Δf	Sub-carrier bandwidth
Δp	Slice of power to be transferred for fairness adaptation purposes
$\zeta_{j,k}$	Signal-to-noise ratio of the jth user in the kth sub-carrier
$\overline{\zeta_{j,k}}$	Mean signal-to-noise ratio of the jth user in the kth sub-carrier
ζ^{th}	Signal-to-noise ratio threshold for the use of the minimum modulation level
η_{nrt}	Step size for the control loop of the ATF technique
η_{rt}	Step size for the control loop of the ADF technique
ϑ	Noise figure
λ_j	Proportional rate requirement of the jth user
μ	Water-level of the waterfilling problem
ϕ_j	General user fairness index
ϕ_j^{nrt}	User fairness index for non-real time services
ϕ_j^{rt}	User fairness index for real time services

Φ_{cell}	General cell fairness index
$\Phi_{\text{cell}}^{\text{nrt}}$	Cell fairness index for non-real time services
$\Phi_{\text{cell}}^{\text{rt}}$	Cell fairness index for real time services
$\Phi_{\text{filt}}^{\text{nrt}}$	Cell fairness index for non-real time services
$\Phi_{\text{filt}}^{\text{rt}}$	Cell fairness index for real time services
$\Phi_{\text{cell}}^{\text{target}}$	General cell fairness target
$\Phi_{\text{target}}^{\text{nrt}}$	Cell fairness target for non-real time services
$\Phi_{\text{target}}^{\text{rt}}$	Cell fairness target for real time services
Φ_{cluster}	Fairness index in a cluster of FAPs
ξ	Required coverage percentage
$\rho_{j,k}$	Connection indicator that says if the kth sub-carrier is assigned to the jth user
σ	Shadowing standard deviation

Chapter 1

Introduction

1.1. Background and Motivation

Third Generation (3G) cellular networks were designed for multimedia communication. Using these systems, person-to-person communication can be enhanced with high quality images and video, and access to information and services on public and private networks are improved by the higher data rates and new flexible communication capabilities. These characteristics have allowed the creation of new business opportunities not only for manufacturers and operators, but also for the providers of content and applications using these networks.

International Mobile Telecommunications for the year 2000 (IMT-2000) is a worldwide set of requirements for the family of standards for 3G mobile communications [1]. The IMT-2000 specifications were developed by the International Telecommunications Union (ITU). Among other requirements, 3G systems should support the following minimum data rates in the various mobility environments: 144 kbps with high mobility (vehicular); 384 kbps with restricted mobility (pedestrian); and 2 Mbps in an indoor office environment.

Among the family of standards compatible with the IMT-2000 requirements, three of them have become more popular: Universal Mobile Telecommunications System (UMTS)[1], cdma2000, and Mobile Worldwide Interoperability for Microwave Access (WiMAX)[2]. The WCDMA system [2, 3], which evolved from Global System for Mobile communication (GSM), and the cdma2000 system [4], which is the evolution of the IS-95 standard, are both based on FDD with Code Division Multiple Access (CDMA) technology. On the other hand, the Mobile WiMAX system (802.16e-2005) [5], which evolved from the standard 802.16-2001 for fixed broadband wireless access systems [6], is based on the Orthogonal Frequency Division Multiple Access (OFDMA) technology. The

[1] Among the UMTS systems, the most popular is the Universal Terrestrial Radio Access (UTRA)-Frequency Division Duplex (FDD) system, also known as Wideband CDMA (WCDMA).

[2] The Mobile WiMAX system was considered later in the year 2007 as an IMT-2000 compliant technology.

WCDMA system is standardized by the 3rd. Generation Partnership Project (3GPP), cdma2000 by the 3rd. Generation Partnership Project 2 (3GPP2) and WiMAX by the Institute of Electrical and Electronics Engineers (IEEE).

However, user and operator requirements and expectations have been continuously evolving, and competing radio access technologies have emerged. In this meanwhile, the 3GPP, 3GPP2 and IEEE families have standardized transitional systems from 3G to Fourth Generation (4G). Some examples are: 3GPP High Speed Packet Access (HSPA) and Long Term Evolution (LTE), 3GPP2 cdma2000 Evolution-Data Optimized (EV-DO) Rev. A and Rev. B, and IEEE 802.16k-2007 and 802.16-2009.

The requirements for 4G mobile communications were defined by the IMT-Advanced specifications [7]. IMT-Advanced mobile systems include the new capabilities of IMT that go beyond those of IMT-2000. Such systems provide access to a wide range of telecommunication services including advanced mobile services, supported by mobile and fixed networks, which are increasingly packet-based. Some characteristics and requirements of 4G systems are: wider bandwidth, spectrum flexibility, lower latency, improved system capacity and coverage, reduced overall cost for the operator and packet-optimized radio access technology with enhanced peak data rates of 100 Mbps and 1 Gbps for high and low mobility, respectively.

Nowadays, two proposals for 4G standards are accepted by ITU as IMT-Advanced compliant: 3GPP LTE-Advanced [8] and IEEE WirelessMAN-Advanced[3] (IEEE 802.16m) [9]. One important common characteristic of both 4G systems is the use of Orthogonal Frequency Division Multiplexing (OFDM) as the radio interface technology and OFDMA as the multiple access scheme in the forward link. This choice was driven by some important characteristics of OFDM and OFDMA that make them suitable for high speed mobile wireless systems [10].

OFDM divides a broadband channel into multiple parallel narrowband sub-carriers, wherein each sub-carrier carries a low data rate stream. This has two main advantages: firstly, the system becomes resistant to frequency selective fading generated by multipath propagation, because each sub-carrier will perceive an almost flat frequency response; secondly, if the system bandwidth is large enough, the sum of low data rate streams regarding each sub-carrier will lead to a high data rate transmission. Furthermore, the reception of an OFDM signal requires only a Fast Fourier Transform (FFT), which can be implemented with reasonable computational complexity in the mobile terminal.

OFDM-based systems have multi-user diversity, i.e. the users have different channel qualities with respect to the sub-carriers. This diversity can be fully exploited by OFDMA, because it allows multiple users to transmit simultaneously on different sub-carriers. Besides that, OFDMA takes advantage of the frequency diversity, allowing each sub-carrier to use a different Modulation and Coding Scheme (MCS) or a particular power level. Since OFDMA resources are also fragmented in the time domain, users can access the system in different time slots. Finally, the parallel nature of

[3] Also known as Mobile WiMAX Release 2.

the OFDM multiplexing is specially suitable for Multiple Input Multiple Output (MIMO) schemes.

In order to fully exploit the flexibility offered by OFDMA and achieve the challenging requirements of 4G systems, efficient Radio Resource Management (RRM) techniques are of utmost importance. Assuming that the base station knows the Channel State Information (CSI) of the different users, adaptive allocation mechanisms can be used to allocate the limited resources, e.g. bandwidth and power, in an intelligent way in order to maximize some performance metric. Therefore, the problem of allocating time slots, sub-carriers, rates, and power to the different users in an OFDMA system has been an area of active research.

There is another force that can boost the performance of mobile broadband networks towards the data rate requirements of 4G systems. According to [11], the spatial reuse of the frequency spectrum on continuously smaller cell sites has been the main responsible for the huge growth in wireless capacity over the past years. In that sense, femtocells are one of the latest technological contributions that began to attract widespread industry attention in early 2008 [12].

Femtocell Access Points (FAPs) are miniature base stations that provide indoor wireless coverage to mobile stations using some kind of cellular technology and are connected to the rest of the network through a fixed broadband Internet connection, such as Digital Subscriber Line (DSL), cable or optic fiber. Two of the most important benefits of femtocells are capacity increase and signal quality enhancement. The former is due to the higher areal spectral efficiency as a consequence of the reutilization of frequency in smaller cells, while the latter comes from the shorter distance between transmitters and receivers. Other advantages of the use of femtocells are load sharing between femto- and macrocells and infrastructure cost reduction.

Femtocells fill a gap where they are needed most: indoor environments. 4G OFDMA-based macrocells using intelligent RRM techniques are very efficient at providing high data rate multimedia services to outdoor mobile users. However, due to penetration losses, it is very difficult to use high modulation orders in the connections located indoors. Therefore, femtocells have been considered as an integral part of 4G systems in order to achieve the data rate requirement of 1 Gbps for low mobility/indoor users advocated by IMT-Advanced[4].

However, these benefits will only be possible if the mobile operators take special care of interference. The overall Quality of Service (QoS) of the users connected to the 'umbrella' macrocell should not be degraded by the femtocell deployment, mainly in the downlink. In the same way, FAPs located in the cell edge must be able to serve the indoor users without being disturbed by 'loud' macro users, mainly in the uplink. Moreover, the uncoordinated distribution of the femtocells, which are installed by the subscribers in their home or office premises, must be bypassed by self-organizing RRM strategies that are able to guarantee a seamless co-existence between neighbor femtocells. Therefore, it is clear that RRM techniques responsible for interference avoidance/coordination will play a very important role for the successful deployment of femtocells.

[4]In the 3GPP LTE-Advanced system, femtocells are called Home enhanced Node Bs (HeNBs) and their functionalities and requirements are standardized in a set of dedicated technical specifications [13, 14].

Resource allocation for wireless mobile communications systems can have different objectives, such as the maximization of system capacity, cell coverage, user QoS, fairness in the resource distribution, etc. Unfortunately, in general all these objectives cannot be achieved at the same time. Below we list some fundamental compromises that appear in wireless cellular networks:

- ***Coverage vs. QoS***: Due to propagation losses, the QoS of the users located in the cell edge is usually worse than the one perceived by the users that are close to the base station. A procedure used in the planning and dimensioning of cellular systems is to determine the cell radius depending on the required percentage of the users that should use the minimum allowed MCS. The trade-off is also evident in this dimensioning procedure, because the higher the minimum QoS requirement, the smaller the cell coverage will be.

- ***Capacity vs. Coverage***: Excessive capacity can have a negative impact on the coverage of interference-limited systems. This is the case of 3G systems based on CDMA, where the cells shrink when they become heavily loaded (cell breathing phenomenon) [2]. Another aspect is that base stations with high power provide good coverage, but also generate excessive interference to the neighbor cells, which can decrease the overall system capacity.

- ***Capacity vs. QoS***: A clear compromise between system capacity and user QoS is the fact that the existence of more users in the system decreases the QoS per capita, because less resources would be available for each of the users. Furthermore, a common way to evaluate system capacity is to determine the number of admitted users that corresponds to a given percentage of satisfied users. The higher the satisfaction percentage requirement, the lower the capacity (and vice-versa).

- ***Fairness vs. Coverage***: The random user location in the coverage area and the wireless channel variability cause differences in the channel quality perceived by the users. And this quality variability is directly proportional to the cell coverage: the larger the cell size, the higher the variability. Normally, resource allocation algorithms take into account the CSI of the users. So, the higher the variability of the users' CSI, the lower the fairness of the corresponding resource allocation.

- ***Fairness vs. QoS***: Since the wireless resources are limited, the QoS of the users cannot be improved indefinitely. If the QoS of few users is maximized, the others will feel the lack of resources. This imbalance is translated into a fairness decrease. On the other hand, if a high fairness is assured and the users have more or less the same QoS, the maximum achievable QoS in this situation is upper-bounded.

- ***Capacity vs. Fairness***: This compromise is also known as the *efficiency* vs. *fairness* trade-off. In order to maximize system capacity, the wireless resources must be allocated in the most efficient way possible. This is accomplished by using opportunistic resource allocation

algorithms, which assign the resources to the users who have the best channel conditions with respect to these resources. As commented before, mobile cellular systems present a high variability on the channel quality experienced by the users. The use of opportunistic RRM in order to maximize capacity will inevitably concentrate the resources in the hands of the best users, while the ones in worse conditions would starve. This situation is characterized by low fairness. On the other hand, if a high fairness is required, the system is forced to cope with the bad channel conditions of the worst users and allocate resources to them. Since this allocation is not efficient in the resources' point-of-view, the overall system capacity will be degraded.

Notice that the compromises described above are fundamental trade-offs found in mobile cellular systems and most of them are technology-independent. System design, the deployment of specific technologies and the use of suitable RRM techniques can help the network operators to decrease the gap between these opposing factors. If these compromises cannot be solved in a 'win-win' approach, adaptive RRM strategies are still very useful at finding an appropriate trade-off between these objectives.

1.2. Objectives of the Thesis

The concepts that will be treated throughout this thesis stand in the forefront of future broadband cellular systems and is well aligned with the requirements of 4G mobile communication networks. The general objective of this work is to contribute with the state-of-the-art by settling critical and original knowledge about RRM for OFDMA-based macro- and femtocell networks in the ambit of the Medium Access Control (MAC) layer (L2).

The main objectives of this thesis are summarized below:

1. Theoretically conceive and evaluate, using system-level simulations, adaptive RRM solutions for the downlink of OFDMA-based macrocell networks, mainly focusing on the optimization of sub-carrier assignment and power allocation.

2. Propose and assess adaptive RRM techniques for the downlink of OFDMA-based femtocell networks, mainly focusing on interference avoidance strategies.

3. Conceptually study the fundamental trade-off between efficiency (capacity) and fairness in wireless networks and propose ways to manage this trade-off in OFDMA-based systems using RRM.

4. Formulate and evaluate RRM policies with different optimization objectives, such as maximization of system capacity, maximization of fairness in the resource distribution, or a compromise between these factors.

5. Analyze the applicability of different optimization tools in the solution of the RRM problems formulated in this thesis.

1.3. Research Methodology

The methodological approach used to fulfil the objectives presented in the previous section was based on two main fronts: theoretical analysis and system-level simulations.

Firstly, a continuous compilation of bibliographical database was conducted, in order to establish a solid foundation on the research subject. The sources used in this bibliographical research were books, articles (journals, transactions, magazines), papers published in proceedings of international conferences, standardization technical specifications and reports and other databases available in the Internet. This theoretical basis was important because of two reasons. Firstly, it allowed the formulation and proposal of the novel contributions presented in this thesis. Moreover, it was necessary to determine the models that would be implemented in the simulator used in our studies.

The performance evaluation of wireless communications systems using only an analytical / mathematical approach is a very complex task. This is due to the fact that many physical phenomena, like the time-variability of the mobile radio channel and fading in the space and time domains are very difficult to be tackled mathematically [15]. For that reason, system-level simulations are widely used to evaluate aspects such as system planning, dimensioning and adjustment of techniques and algorithms. In fact, simulations make possible the modeling of the system with a high level of detail, which allows the validation of specific algorithms [16]. In the ambit of this thesis, simulations are used with two objectives: 1) validation of the propositions (hypothesis) inherent to the propounded RRM algorithms; 2) the performance evaluation and comparison of the corresponding RRM techniques.

To this end, a discrete-time system-level simulator that models the most important aspects of a cellular network in general, and the OFDMA technology in particular, was developed using the simulation software package Matlab. The RRM techniques and frameworks proposed and studied in this thesis were implemented in this simulation tool. A detailed description of the models used in the simulation tool is given in appendix A.

The performance evaluation of the proposed RRM strategies carried out in this thesis is based on the statistical analysis of the simulation results. The simulation scenarios studied in this work are simulated several times according to the Monte Carlo approach in order to have a reliable statistical confidence. The simulator is flexible in the sense that each snapshot can have different time durations according to our purpose. Some RRM techniques evaluated in this thesis operate on the time basis of one Transmission Time Interval (TTI), and one iteration of the algorithm can be independent of the other iterations. On the other hand, some other RRM strategies need a snapshot with a larger time duration in order to capture the dynamics of the system in the time domain as well as the traffic pattern of the users.

1.4. Contributions of the Thesis

Guided by the goals presented in section 1.2 and using the research methodology described in section 1.3, this thesis makes the following meaningful and novel contributions:

1. ***Performance evaluation tool to analyze the trade-off between efficiency and fairness***: We introduce the concept of efficiency-fairness curves and regions as a visualization tool to complement the performance evaluation of RRM techniques in wireless networks.

2. ***Management of the trade-off between efficiency and fairness in macrocell networks***: A quantitative fairness measure is used to calculate a cell fairness index. We claim that each value of the cell fairness index corresponds to a different performance in terms of efficiency in the resource usage. Therefore, if the mobile operator is able to force the network to operate on a desired fairness level, it can control the trade-off between efficiency and fairness. This original idea is validated using different Radio Resource Allocation (RRA) techniques suitable for macrocell networks, which are evaluated in different simulation scenarios, as confirmed by the following three contributions:

   a) ***Fairness/rate adaptive resource allocation for Non-Real Time (NRT) services***: Rate adaptive is a well-known optimization approach for NRT services in OFDMA macrocell networks. We propose generalizations of some classic rate adaptive policies, where a new fairness constraint is introduced in the corresponding RRA optimization problems. The proposed RRA techniques perform a strict short-term (instantaneous) fairness control and enable the network operator to work on any trade-off point of the efficiency-fairness plane.

   b) ***Utility-based RRA framework for NRT services***: Based on Utility Theory concepts, we formulate a general parametric RRA framework suitable for NRT services named utility-based alpha-rule. Adjusting only a parameter α in its parametric structure, this RRA framework can be designed to work as well-known classic RRA policies. Furthermore, by using an adaptive throughput-based utility function, it is able to perform a mid-term (average) control of the fairness level in the system.

   c) ***Utility-based RRA framework for Real Time (RT) services***: We also propose a novel RRA framework called utility-based beta-rule, whose fairness controlling parameter β can be adapted and make the framework behave as well-known classic RRA policies. Moreover, this flexibility of the framework can be explored and the delay-based fairness in the system can be adjusted so that the cell fairness index can converge to a desired target.

3. ***Interference avoidance technique based on frequency planning for femtocell networks***: Interference avoidance policies based on a high-level, mid/long-term frequency planning and able to balance spectral efficiency and resource-based fairness in femtocell networks

are propounded. The proposed technique can be implemented in a centralized or distributed network architecture and allows the seamless co-existence of several FAPs in any interference topology.

1.5. List of Publications

Some of the contributions of the thesis listed in section 1.4 have been published or accepted for publication in different journals, magazines, international conferences and Spanish national conferences. The complete list of publications associated with this thesis work is presented below.

Journals/Magazines

[J.1] E. B. Rodrigues and F. Casadevall, "Rate adaptive resource allocation and utility-based packet scheduling in multicarrier systems," Majlesi Journal of Electrical Engineering, vol. 5, no. 1, pp. 38-49, March 2011.
[Online] Available: http://ee.majlesi.info/index/index.php/ee/article/view/371

[J.2] E. B. Rodrigues and F. Casadevall, "Control of the trade-off between resource efficiency and user fairness in wireless networks using utility-based adaptive resource allocation," IEEE Communications Magazine, accepted for publication.

Conferences

[C.1] E. B. Rodrigues, M. Moretti, P. Sroka, and F. Casadevall, "Sub-carrier allocation and packet scheduling in OFDMA-based cellular networks," in Proc. NEWCOM++/ACoRN Joint Workshop, pp. 1-5, April 2009.

[C.2] E. B. Rodrigues and F. Casadevall, "Adaptive radio resource allocation framework for multiuser OFDM," in Proc. IEEE 69th Vehicular Technology Conference - VTC Spring, pp. 1-6, April 2009.

[C.3] E. B. Rodrigues, F. Casadevall, P. Sroka, M. Moretti, and G. Dainelli, "Resource allocation and packet scheduling in OFDMA-based cellular networks," in Proc. 4th International Conference on Cognitive Radio Oriented Wireless Networks and Communications - CrownCom, pp. 1-6, June 2009.

[C.4] E. B. Rodrigues, M. L. Walker, and F. Casadevall, "On the influence of packet scheduling on the trade-off between system spectral efficiency and user fairness in OFDMA-based networks," Lecture Notes in Computer Science - LNCS 5733 / Proc. 15th EUNICE International Workshop, pp. 128-137, September 2009.

[C.5] E. B. Rodrigues and F. Casadevall, "*Evaluación de la influencia de los algoritmos de scheduling en el compromiso entre eficiencia espectral y equidad en redes OFDMA*," in Proc. *XXIV Simposium Nacional de la Unión Científica Internacional de Radio - URSI*, pp. 1-4, September 2009.

The work carried out during the doctoral research period has been linked to different research projects, which are detailed in the following:

- "Network of Excellence in Wireless COMmunications++ (NEWCOM++)," funded by the European Union in the Seventh Framework Programme (FP7), Information and Communication Technologies (ICT), Network of Excellence (NoE), period 2008-2010, Ref. INFSO-ICT-216715.

- "Cognitive Spectrum and Radio Resource Management in Heterogeneous Mobile Networks with End-to-End Quality of Service Provisioning (COGNOS)"; original title in Spanish: "*Gestión Cognitiva de Recursos Radio y Espectro Radioelétrico en Redes Móviles Heterogéneas con Provisión de Calidad de Servicio Extremo a Extremo*"; funded by the Spanish Research Council (MICINN) in the framework of the *Plan Nacional de Investigación Científica, Desarrollo e Innovación Tecnológica 2004-2007*, Ref. TEC2007-60985.

The outcome of this project involvement can be measured not only in a couple of joint publications with project partners (see publications [C.1] and [C.3]), but also in the following project deliverables.

Project deliverables

[PD.1] A. Serrador and L. M. Correia (Editors), "Identification of relevant scenarios, use cases and initial studies on JRRM and ASM strategies," NEWCOM++ Deliverable DR9.1, January 2009.

[PD.2] P. Sroka (Editor), "Definition and evaluation of JRRM and ASM algorithms," NEWCOM++ Deliverable DR9.2, January 2010.

[PD.3] Jordi Pérez-Romero (Editor), "Final report of the JRRM and ASM activities," NEWCOM++ Deliverable DR9.3, February 2011.

1.6. Research Structure of the Thesis

The diagram of the thesis research structure is depicted in Fig. 1.1. The blocks in green color correspond to subjects that were addressed in the thesis, whilst the blocks in red were not studied. The remaining of this document is structured in five chapters and four appendices, which are briefly described as follows.

Chapter 2 presents an overall description of RRM problems and solutions for OFDMA-based cellular networks. The topics discussed in this chapter are: the characteristics and benefits of the OFDMA technology and the most common RRM techniques for this kind of system; some optimization tools found in the literature that are able to solve important RRM problems; the system architecture considered in this work; some taxonomy and open research topics associated to RRM for macro- and femtocell networks; and a detailed explanation about the trade-off between efficiency and fairness in wireless networks. Throughout this chapter, we emphasize the choices we have made regarding the possible research directions we could have taken. In this way, it will become clear why we have selected the research topics presented in Fig. 1.1.

There are many RRM problems associated with macrocell OFDMA systems in a single-cell scenario that can be represented using optimization formulations. **Chapter 3** considers one of these possible formulations, which is the well-known rate adaptive optimization, whose goal is to maximize an objective function dependent on the instantaneous users' data rates. As commented in section 1.4, we propose generalized fairness/rate adaptive RRA techniques that are able to perform instantaneous fairness control. In order to achieve the optimization goal, we use heuristics to propose novel Dynamic Sub-carrier Assignment (DSA) and Adaptive Power Allocation (APA) algorithms. Simulation results are presented in order to demonstrate the correctness of the techniques. A detailed description of the simulation tool used in this thesis is presented in **appendix A**. The classic rate adaptive policies that inspired the proposed fairness-aware techniques are described in details in **appendix B**.

Utility Theory is used in **Chapter 4** to propose two adaptive RRA frameworks suitable for macrocell networks providing NRT and RT services, respectively. The former is the outcome of a utility-based optimization problem based on throughput (average data rate), while the latter is based on an optimization with respect to packet delays. Both frameworks are composed of DSA and APA algorithms that rely on the use of utility-based weights to provide QoS differentiation among users and adjust the fairness distribution in the system as desired. The performance of the frameworks is evaluated by means of system-level simulations. In this performance evaluation, various aspects are considered, such as preliminary analysis of classic RRA techniques, convergence of the proposed frameworks, fairness, system capacity, user satisfaction and algorithmic complexity. We also make a comparison between the fairness/rate adaptive RRA techniques described in chapter 3, which consider the optimization formulation based on instantaneous rates, and the utility-based RRA framework suitable for NRT services presented in chapter 4, which considers a time window for the fairness control in the utility-based optimization problem. **Appendix C** propounds a variation of the DSA algorithms of the utility-based frameworks, where a new range of the fairness controlling parameters is considered.

Chapter 5 propounds an interference avoidance technique for the femtocell tier. We propose and evaluate by means of numerical simulations several RRA policies that are able to find different compromises between spectral efficiency in the femtocell tier and fairness in the distribution of

resources among FAPs. We formulate an optimization problem where the variable to be optimized is the number of frequency resources allocated to each FAP and the optimization constraints assure perfect interference avoidance. In order to solve this problem, two RRA algorithms are proposed in the thesis, namely Dynamic Frequency Planning (DFP) and Femtocell Sub-carrier Allocation (FSA). Two versions of the DFP algorithm have been implemented and evaluated: the first using the Branch and Bound (BnB) technique and the second using heuristics. Heuristics is also chosen for the proposal of the FSA algorithm. Aspects related to the implementation of the techniques in centralized and distributed network architectures are also discussed in the chapter. For the interested reader, we present a brief overview of the BnB technique in **appendix D**.

Finally, **Chapter 6** recapitulates the main conclusions of the thesis. We also present a summary of the features of all RRA techniques studied/proposed in this thesis, both in the macrocell and femtocell scenarios, regarding the trade-off between efficiency and fairness. Finally, some perspectives for future work are listed.

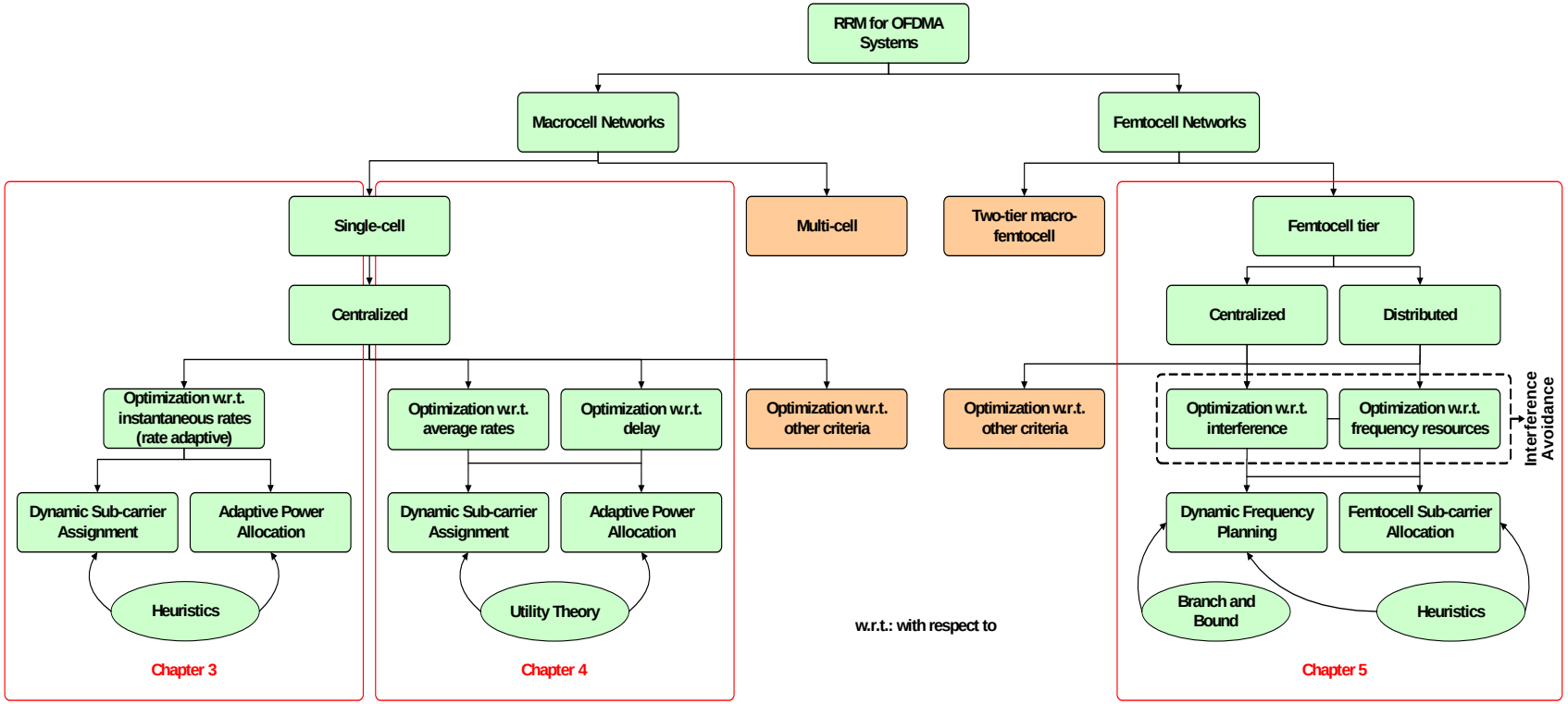


Figure 1.1.: Thesis research structure

Chapter 2

Radio Resource Management for OFDMA Systems

2.1. Introduction

Radio Resource Management (RRM) techniques are responsible for the utilization of the radio resources of the air interface of a given cellular network. RRM functionalities are decisive for the guarantee of Quality of Service (QoS) requirements of different service classes, the optimization of coverage, the maximization of the spectral efficiency (system capacity) and the provision of acceptable fairness in the resource and QoS distribution among the network players.

The Orthogonal Frequency Division Multiple Access (OFDMA) technology is the chosen air interface for 4G networks, such as the Long Term Evolution-Advanced (LTE-Advanced) system standardized by the 3rd. Generation Partnership Project (3GPP) and the Mobile Worldwide Interoperability for Microwave Access (WiMAX) Release 2 system standardized by the Institute of Electrical and Electronics Engineers (IEEE). OFDMA offers a great flexibility for resource allocation, making it possible to propose a variety of interesting RRM policies.

This chapter has the objective of presenting a general panorama of RRM for OFDMA systems and showing where this thesis is located in this broad picture. The chapter begins with a brief description of OFDMA in section 2.2, where we talk about the advantages of using this technology, the types of radio resources available and the kinds of diversity that can be explored. Section 2.3 presents the RRM techniques that are studied in this thesis, while the most common optimization tools used to solve RRM problems for OFDMA systems are listed in section 2.4. The system architecture assumed in the thesis is described in section 2.5. Since we perform studies in both macrocell and femtocell networks, sections 2.6 and 2.7 present an overview of RRM problems and solutions for these two types of network deployments. Finally, section 2.8 describes in detail an RRM problem that will be dealt with throughout this thesis: the fundamental trade-off between efficiency in the resource usage and fairness in the resource distribution.

2.2. Orthogonal Frequency Division Multiple Access

A lot of scientific effort has been devoted to study Orthogonal Frequency Division Multiplexing (OFDM)-based broadband systems motivated by interesting properties of this radio access technology. The following channel, signal and receiver characteristics are worth noting [17]:

- ***Time dispersion***: The use of several parallel sub-carriers in OFDM enables longer symbol duration, which makes the signal inherently robust to time dispersion. Moreover, a guard time may be added to combat further the Inter-Symbol Interference (ISI)).

- ***Spectral efficiency***: OFDM is constructed with fully orthogonal carriers, hence allowing tight frequency separation and high spectral efficiency. The resulting spectrum has also good roll-off properties, given that cross-symbol discontinuities can be handled through time windowing alone, filtering alone, or through a combination of the two techniques.

- ***Reception***: Even in relatively large time dispersion scenarios, the reception of an OFDM signal requires only a Fast Fourier Transform (FFT) implementation in the mobile terminal. No intra-cell interference cancelation scheme is required. Furthermore, because of prefix insertion, OFDM is relatively insensitive to timing acquisition errors. On the other hand, OFDM requires performing frequency offset correction.

- ***Extension to Multiple Input Multiple Output (MIMO)***: Since the OFDM sub-carriers are constructed as parallel narrowband channels, the fading process experienced by each sub-carrier is almost flat, and therefore, can be modeled as a constant complex gain. This may simplify the implementation of a MIMO antenna scheme if this is applied in the basis of sub-carriers or sub-channels (subset of sub-carriers).

In an OFDMA-based system there exist different resources that need to be properly allocated among the users. The resources can be summarized as follows:

- ***Frequency sub-carrier***: Frequency domain adaptation (sub-carrier dimension) achieves large performance gains in cases where the channel varies significantly over the system bandwidth. Thus, frequency domain adaptation becomes increasingly important with an increasing system bandwidth. OFDM transmission straightforwardly supports such frequency-domain scheduling by the dynamic allocation of different sets of sub-carriers for transmission to different Mobile Terminals (MTs).

- ***Time slot / frame***: Exploiting channel variations in the time domain through channel dependent scheduling has been shown to provide a substantial increase in spectral efficiency, as was observed with the High Speed Downlink Packet Access (HSDPA) system standardized by 3GPP. Multiplexing can also be performed in the time dimension of OFDM-based systems, as long as it occurs at the OFDM symbol rate or at a multiple of the symbol rate.

- ***Modulation and Coding Scheme (MCS)***: Using adaptive modulation and coding, the transmitter can send higher data rates over the sub-carriers with better channel conditions to improve throughput and simultaneously ensure an acceptable Bit Error Rate (BER) in all sub-carriers. The MCS used for each sub-carrier can also be changed at a multiple of the OFDM symbol rate. The dynamic modification of the modulation and coding schemes is commonly referred as link adaptation.

- ***Transmission power***: Due to the frequency-selective attenuation of the wireless channel, the transmit power per sub-carrier can be adapted in order to increase the spectral efficiency. The capacity can be maximized if more transmit power is applied to frequency areas with a low attenuation relative to the other frequencies. As different sub-carriers experience different fades and transmit different number of bits, the transmit power levels must be changed accordingly.

- ***Adaptive antennas and MIMO***: In order to fulfil the requirements on coverage, capacity, and high data rates, various multi-antenna schemes need to be supported by Fourth Generation (4G) broadband systems. For example, beamforming can be used to increase coverage and/or capacity, and spatial multiplexing, sometimes referred to as MIMO, can be used to increase data rates by transmitting multiple parallel streams to a single user. The potential of using the spatial domain is large, and the development of new and even more efficient multi-antenna algorithms is expected to continue for a long time into the future. However, the antenna adaptation adds another dimension to the OFDM resource optimization problem, which is to determine the antenna weight vectors.

There are also different sources of diversity in an OFDMA-based cellular system, which must be properly explored by the RRM algorithms. These diversities are:

- ***Time***: The time diversity comes from the time-varying nature of the mobile radio channel. The velocity in which the state of the channel changes can be estimated by the channel coherence time.

- ***Frequency***: Different sub-carriers of a broadband wireless system have a strongly varying attenuation, i.e. the system provides frequency diversity.

- ***Space***: Space diversity, also known as antenna diversity, is any one of several wireless diversity schemes that use two or more antennas to improve the reliability and quality of a wireless link.

- ***Multi-user***: As several terminals are located in the cell, sub-carriers are likely to be in different quality states for different terminals. In other words, the multi-user communication scenario is characterized by spatial selectivity of the sub-carriers. The reason for this spatial selectivity is the fact that the fading process is, in general, statistically independent for different terminals.

- ***Multi-cell***: In a multi-cell scenario (macrocellular or femtocellular), the entire set of sub-carriers in the system can be managed in order to coordinate the inter-cell interference and prevent co-channel users from interfering too much between them.

The OFDMA framework provides a time/frequency grid, as illustrated in Fig. 2.1 [18]. In the case of multiple transmit antennas, there is one such grid for each antenna. This framework provides the possibility of user multiplexing and link adaptation in the time, frequency, and spatial domains. The minimum duration for transmission and resource assignment is the time slot, also called Transmission Time Interval (TTI), consisting of n OFDM symbols. The TTI durations are chosen to be short enough to allow retransmissions without causing unreasonable packet delays, in order to support real-time applications as well as high throughput for Transport Control Protocol (TCP)-based applications.

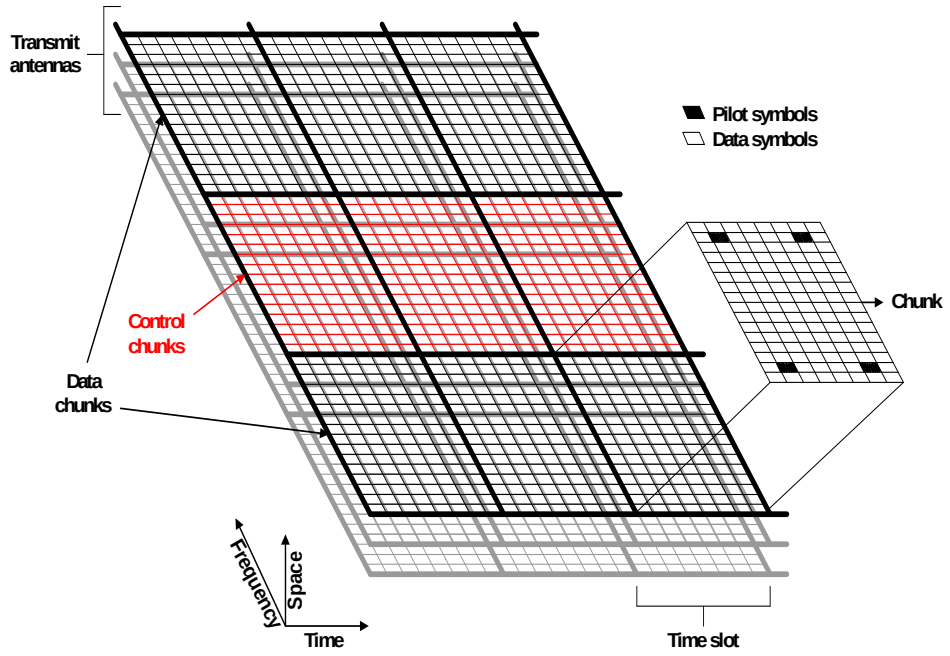


Figure 2.1.: Fundamental OFDMA transmission resources

The time slot is further subdivided in the frequency domain into chunks, also known as Physical Resource Blocks (PRBs), consisting of k adjacent sub-carriers. These chunks are the smallest units for resource allocation and link adaptation. The chunk bandwidths are chosen to be narrow enough so that the channel quality is relatively constant within each chunk, even in rather challenging propagation conditions, while at the same time being large enough to keep the control-signaling overhead reasonable. In Fig. 2.1, the chunk is formed by n x m basic resources, wherein n OFDM symbols and m adjacent sub-carriers are assumed per chunk.

The chunks can be independently assigned for transmission to different MTs to support channel-dependent frequency-domain scheduling. This means that the set of chunks transmitted to a par-

ticular receiver may be fragmented (non consecutively allocated) in the frequency domain. Chunks can also be transmitted with different beams or different antennas, potentially with multiple overlapping layers, thereby supporting scheduling and multiplexing also in the spatial domain. The modulation and coding scheme can be selected independently for each chunk, providing frequency-domain link adaptation. Frequency-domain scheduling and link adaptation with a resolution of a single chunk can provide high performance gains even in severe fading conditions. Besides, OFDMA provides the flexibility of allocating power and rate optimally among narrowband sub-carriers.

Pilots are needed to estimate the channel, both for demodulation and for quality reporting, used for link adaptation. Pilots are inserted in the frequency/time grid by reserving a number of basic resources within each chunk, as can be seen in Fig. 2.1. In addition to chunks for user data, some chunks are pre-allocated for control signaling. These control chunks carry scheduling and transmission format information, i.e., information indicating the intended receiver as well as the transmission format (modulation, coding and pilot pattern) used for each data chunk.

Without loss of generality, we assume in this thesis that the chunks have a frequency granularity of one sub-carrier. Although this is not possible in practical systems due to the required signaling overhead, it is a practice widely used in the research community in order to provide conceptual proofs of the RRM techniques. Notice that usually the number of sub-carriers in a chunk is chosen in the way that all of them perceive a similar (almost equal) frequency behavior. From this point of the document on, the term "sub-carrier" will mean a chunk of one sub-carrier in the frequency-domain that can be allocated to a MT every TTI.

2.3. Radio Resource Management Techniques for OFDMA Systems

There is plenty of room to exploit the high degree of flexibility of RRM in the context of OFDMA. The allocation of resources in OFDMA-based cellular systems corresponds to a multi-dimensional problem, since time, frequency and spatial domains should be efficiently used. An efficient resource allocation strategy is therefore seen as a key aspect able to significantly improve the performance of OFDMA-based systems. In order to take advantage of the finer granularity of OFDMA and the myriad of time/frequency/space resources, it is important to devise efficient allocation techniques to fully exploit all kinds of diversity offered by the system.

Efficient and innovative RRM techniques are essential to provide considerable gains in coverage, capacity and QoS for OFDMA-based broadband wireless networks. In a commercial wireless network, improved coverage, capacity and QoS represent better investment return rates and better radio services. For the customers, this should yield better services, higher fairness and enhanced QoS levels with widespread availability at possibly lower prices. Therefore, it is of utmost importance to investigate innovative means to optimize RRM techniques that deal with the resources in an OFDMA-based system.

RRM techniques intend to exploit the several sources of diversity present in the wireless cellular

systems in order to allocate the available radio resources in the best way possible, considering aspects such as system capacity, cell coverage, user fairness and user QoS. There are many Radio Resource Allocation (RRA) algorithms available for OFDMA-based cellular systems. It is not the objective of this thesis to present an exhaustive list of such techniques, so we present below a brief description of only the ones considered in our research:

- ***Dynamic Sub-carrier Assignment (DSA)***: The spatial selectivity of the sub-carriers, which is related to the multi-user diversity, gives rise to the opportunity of assigning different sub-carriers to different users. The DSA algorithm explores this flexibility of the OFDMA system and determines the pairs users/sub-carriers according to a given RRA policy [19].

- ***Adaptive Power Allocation (APA)***: This algorithm is also called Power Loading in the literature. Each sub-carrier may face a different channel gain depending on which frequency it is related to (frequency diversity), when it is allocated (time diversity) and for which user it is assigned (multi-user diversity). Taking this into account, it is advantageous to dynamically adapt the power of each sub-carrier [17].

- ***Adaptive Modulation and Coding (AMC)***: This technique is also known as bit loading [20]. It exploits the time and frequency diversities in order to allocate the most suitable MCS to each sub-carrier according to its Signal-to-Noise Ratio (SNR). In this thesis, we use a simplified AMC algorithm that is executed after the DSA and APA algorithms, once the SNR of each sub-carrier has been calculated. For more details about the link adaptation scheme used in this research, refer to section A.5 of appendix A.

- ***Dynamic Frequency Planning (DFP)***: An example of frequency planning approach that was proposed to mitigate the inter-cell interference in macrocell networks is Fractional Frequency Reuse (FFR) [21]. It is also possible to use frequency planning in order to determine the amount of sub-carriers to be allocated to users or base stations as a prior step before the final assignment of sub-carriers' subsets to the end-users [22, 23].

- ***Femtocell Sub-carrier Allocation (FSA)***: This is a special case of the DSA algorithm applied for femtocells. In this thesis, we assume that the final sub-carrier assignment to the users is done by the DSA algorithm, whilst the sub-carrier allocation for the Femtocell Access Points (FAPs) is performed by the FSA algorithm.

- ***Interference avoidance/coordination***: This is a broad, "high-level" class of RRM strategies that comprises other "low-level" RRM algorithms in order to manage the interference in the system. Considering the general case of a two-tier macro- and femtocell network, the possible types of interference to be managed are: macro-to-macro, macro-to-femto, femto-to-macro and femto-to-femto.

As commented in section 1.6 of chapter 1 (see Fig. 1.1), chapters 3 and 4 propose RRA strategies suitable for macrocell networks and comprised of DSA and APA algorithms, while an interference

avoidance technique suitable for femtocell networks and composed of the DFP and FSA algorithms is formulated in chapter 5.

2.4. Optimization Tools

There are many RRM problems in OFDMA-based cellular networks. Some possible objectives of RRM techniques are: better capacity, optimized coverage, lower energy consumption (green radio), enhanced QoS, higher fairness, better usage of the spectrum (cognitive radio), etc. Most of these RRM approaches can be formulated as RRA optimization problems, with an objective function that must be maximized or minimized and optimization constraints that correspond to physical restrictions of the network.

In order to solve these problems, some optimization tools are more suitable than others. This choice depends on some factors, such as the formulation of the optimization problem itself, the nature of the variables involved, the possibility of linearization of the objective function and the possibility of relaxation of the optimization constraints. Furthermore, for the same RRM problems, different optimization formulations lead to different RRA policies.

When the optimum solution of the optimization problem is too difficult to be found, some sub-optimum approaches can be used. Two of the most common approaches found in the literature are listed below [17, 19]:

- ***Relaxation of constraints***: The idea is to relax some optimization constraints in order to ease the solution of the problem. An example is to relax the integer constraint on the sub-carrier or bit assignments, so that each sub-carrier can be assigned to multiple different MTs simultaneously and/or can carry a non-integer amount of bits during one TTI. By doing this, the optimization problems may become linear programming problems, which can be solved efficiently. However, one must have in mind that after solving the relaxed problem, the relaxed solution has to be re-evaluated because only integer solutions are feasible from the network's point of view.

- ***Problem splitting***: This approach uses the concept "divide to conquer", i.e. split the complex problem into two or more simpler steps so that a sub-optimum solution close enough to the optimum can be found. This approach is widely used in the RRA techniques proposed in this thesis.

A great variety of optimization tools exist deriving from both mathematics and computer science. These tools are based on several methods - ranging from linear to evolutionary programming techniques. It is out of the scope of this thesis to present a rigorous classification of optimization tools suitable for the solving of RRM problems. Therefore, we only list below the tools used in the present research, and next we present a brief description of some common optimization tools existent in the OFDMA technical literature.

- ***Heuristics***: It refers to experience-based techniques for problem solving, which are used to find sub-optimum solutions hopefully close to the optimum with much less complexity than other conventional combinatorial optimization techniques. It has been widely used by many works in the literature.

- ***Utility Theory***: This theory was originally conceived for application in economics [24, 25], but has been attracting a lot of attention of researchers of the area of communication networks recently. Utility theory provides a flexible means to formulate quantitatively the relations between user experience and various network performance metrics. Utility functions are able to capture the satisfaction level of users for a given resource assignment.

- ***Branch and Bound (BnB)***: BnB is an approach developed for solving combinatorial optimization problems [26–29]. This optimization tool combines enumeration of all possible solutions by means of "branches", and the process of "pruning" some of them. The BnB approach is not an approximating procedure, but it is a deterministic optimization technique that finds the optimum solution. In [30], a BnB method was applied in an OFDMA system in order to solve power minimization and rate maximization problems.

In the scope of this thesis represented in Fig. 1.1, we use heuristics to implement the DSA and APA algorithms evaluated in chapter 3. The same RRA algorithms are completely reformulated using Utility Theory in chapter 4, which makes it possible to face a similar problem from a different RRM perspective. Finally, it can be seen in chapter 5 that heuristics and the BnB technique are used to formulate two variants of the DFP algorithm, while heuristics is also used in the proposal of the FSA algorithm. For a brief description of these RRA algorithms, see section 2.3.

As commented before, several optimization tools can be used to solve other specific RRA problems. In the following, we present some of the tools found in the literature.

Convex optimization is one of the most common tools used to propose RRA techniques for OFDMA systems. For example, the Lagrangian's method of multipliers, which is a classical tool for nonlinear optimization problems with constraints, was utilized to solve analytically the well-known Finite Tones Water Pouring problem [31]. The Simplex method is an effective tool to solve an important class of optimization problems known as linear programming, in which the objective function and all constraints are linear. In [32], it was applied in an OFDMA multi-cell system in order to solve a linear programming problem that maximizes the total throughput of the network while guaranteeing the QoS of the users. The Feasible Sequential Quadratic Programming (FSQP) method is a class of efficient algorithms for solving nonlinear constrained optimization problems. It has received much attention and an example of its application in OFDMA multi-cell systems aiming bit rate maximization using power allocation algorithms can be found in [33].

Probabilistic optimization (stochastic programming) is another possible optimization tool. This class of optimization comprises Genetic Algorithm (GA) and Simulated Annealing (SA), among others. GA is an established stochastic search method based on the Charles Darwin's theory of

natural selection. GA, unlike other traditional methods of optimization (as gradient-based method, random search and enumerative schemes), is robust and effective in combinatorial optimization problems. They are effective because of their ability to exploit favorable characteristics of previous solutions and successively produce better solutions. Other advantages of GA are: it is not necessary to know if the objective function is continuous or differentiable; and they are easy to implement. An example of application of GA in an OFDMA system was proposed in [34]. This work considers the problem of minimization of the total power needed for an OFDM symbol, subject to QoS restrictions in a single-cell scenario. SA is a random-search technique which exploits an analogy between the way in which a metal cools and freezes into a minimum energy crystalline structure (the annealing process) and the search for a minimum in a more general system. Not only does SA accept solutions with improved cost, but it accepts solutions with deteriorated cost with a given probability as well. This feature gives to the algorithm the "hill climbing" capability and, consequently, the ability to avoid becoming trapped in local minima. Different from GA, SA works with only one solution instead of a set of solutions. The work [35] proposes the application of SA in an OFDMA system, considering the problem of maximizing a utility function subject to power and QoS restrictions in a single-cell scenario.

The *Hungarian method* is a classic tool for solving the assignment problem [36], a special sub-class of transportation problems. In its standard form, the assignment problem is formulated to minimize the sum of ratings (or costs, in analogy to transportation problems), when a cost matrix is defined. There are several publications exploiting the Hungarian method as an optimization tool towards the RRM in OFDMA systems, e.g. [37–41].

The *Vogel's method* is one of the efficient algorithms marked for application in the transportation problem. This is an iterative approach based on a heuristic, giving an initial feasible basic solution that can yield to a solution very close to the optimum. It is shown in [42] that the Vogel's method can be applied for sub-carrier allocation in OFDMA systems, even though it is a sub-optimum approach.

The application of the *Eigenvector method* in multi-cell wireless communications is well known, mainly in power control [31]. In this problem, a central controller tries to maximize the minimum Signal-to-Interference plus Noise Ratio (SINR) among all co-channel links (links using the same channel in different cells). This max-min problem leads to a balancing of SINRs in the set of these links, promoting the fairness among the mobiles. This method was applied in a multi-cell OFDMA system, where each co-channel link corresponds to a sub-carrier allocated in different cells for distinct MTs [43]. However, before applying this technique, it is necessary to determine the assignment of MTs to the sub-carriers in order to identify the sets of co-channel links with their respective channel gains. Finding the best user assignment for each sub-carrier involves an exhaustive search over all users. In order to avoid this, the work [43] proposes a sub-optimum approach to find the user allocation.

Game Theory models strategic situations, or games, in which an individual's success in making

choices depends on the choices of others. It has been initially used in social sciences, most notably in economics, but has been successfully used in several areas of wireless communication problems. A tutorial article about dynamic spectrum sharing in cognitive networks is [44]. This work analyzes the network users' behaviors and proposes an efficient dynamic distributed design. A distributive non-cooperative game is proposed in [45] to perform sub-channel assignment, adaptive modulation, and power control for multi-cell OFDMA networks. The goal was to minimize the total transmission power under the constraints of rate requirement and maximum transmitted power.

2.5. System Architecture

In this thesis, we assume a flat all-IP access network, as indicated in Fig. 2.2. This figure presents a high-level system architecture that comprises both macrocell and femtocell networks. This system architecture is well aligned with the current standards of the 3GPP LTE-Advanced [13, 14, 46] and IEEE Mobile WiMAX Release 2 [9, 47] systems.

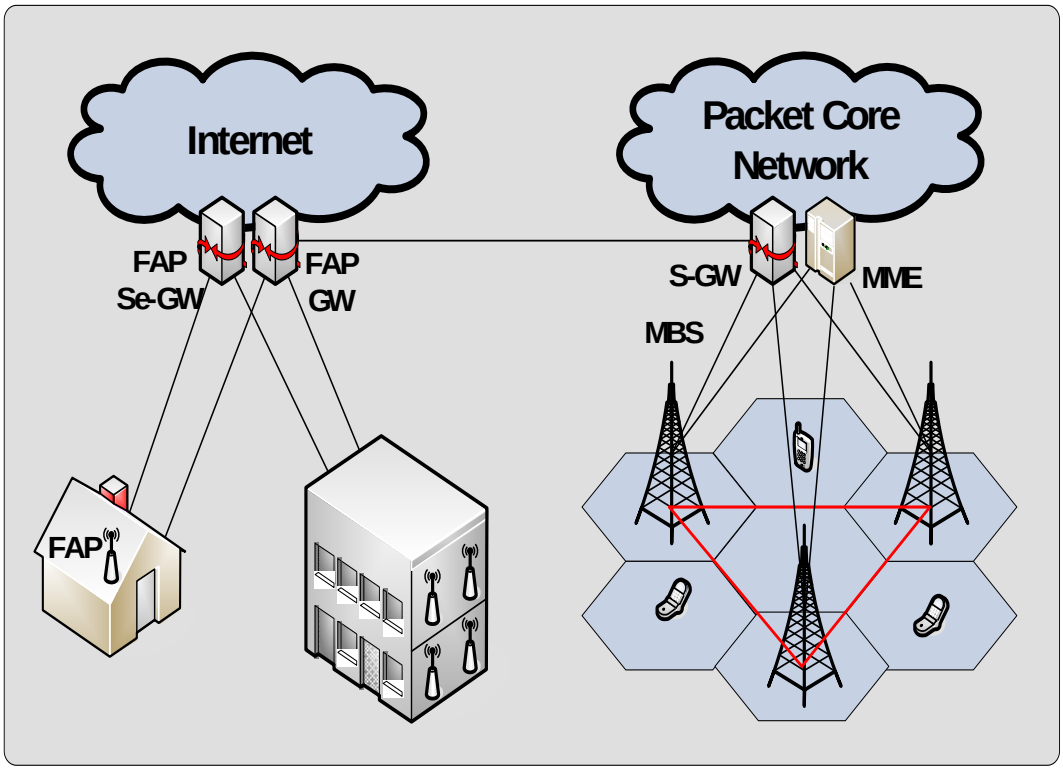


Figure 2.2.: System architecture

Regarding the macrocell network, it can be seen in Fig. 2.2 that several advanced Macrocell Base Stations (MBSs) provide multimedia services to MTs in the macro coverage area. These MBSs are connected between them in order to exchange information related to mobility control and inter-cell interference coordination. In 4G cellular networks, the RRM techniques are executed in the MBSs.

The MBSs are linked directly to the packet core network of the mobile operator, which has two main components: the Mobility Management Entity (MME) and the Serving-Gateway (S-GW). Among other functionalities, MME is responsible for idle mode MT tracking and paging procedure, bearer activation/deactivation process, user authentication, ciphering/integrity protection, security management, roaming, etc. The S-GW routes and forwards user data packets, while also acting as the mobility anchor during inter-BS and inter-Radio Access Technology (RAT) handovers. It also manages and stores MT contexts, e.g. parameters of the Internet Protocol (IP) bearer service, network internal routing information, etc.

Now, let us consider the femtocell network. A FAP located in a house or building connects to the outer world by means of an Internet fixed broadband connection, such as Digital Subscriber Line (DSL), cable, or optic fiber. In order to reach the mobile network, this FAP must be connected to a FAP Gateway (FAP GW). In general, Internet links are not secured, so the connection between the FAP and its gateway goes through a FAP Security Gateway (FAP Se-GW), which provides an appropriate security mechanism. The FAP Se-GW provides IP Security (IPsec) tunnels for the FAPs and is responsible for their authentication and authorization. The FAP is connected to the FAP GW and other functional entities in the network through this IPsec tunnel. The FAP GW acts as a concentrator to aggregate the traffic of a large number of FAPs to the packet core network. We assume that the FAP GW supports femtocell-specific RRM functionalities, such as admission control, handover control and interference management.

In this thesis, we consider the downlink of an OFDMA-based cellular system using Frequency Division Duplex (FDD). The reason for choosing to study the downlink direction is twofold. Firstly, it is widely assumed in the mobile communication research area that most of the data traffic in cellular networks is transmitted in the forward link. This is due to the fact that the data generation of multimedia services like audio and video streaming, web browsing and File Transfer Protocol (FTP) is highly asymmetric. Secondly, the OFDMA technology is usually used in the downlink. For example, the 3GPP LTE-Advanced system adopts the Single Carrier Frequency Division Multiple Access (SC-FDMA) technology for the uplink, because it has the advantage of presenting a Peak to Average Power Ratio (PAPR) lower than the OFDMA case, which greatly benefits the mobile terminal in terms of transmit power efficiency and terminal costs [13, 48].

In the following sections 2.6 and 2.7, a brief revision of the state-of-the-art and some common RRM problems and approaches for the downlink of OFDMA-based macro- and femtocell networks are presented.

2.6. RRM for the Downlink of OFDMA Macrocell Networks

Reviewing the state-of-the-art about RRM for OFDMA macrocell networks, we observe two main scenarios: single-cell and multi-cell. The former is characterized by the presence of one transmitter, which is the MBS, and multiple receivers, which are the MTs. In this scenario, the inter-cell

interference is not modeled, and so the quality measure of the users is the SNR. The latter is characterized by the presence of multiple MBSs (transmitters) and multiple MTs (receivers). In this case, the quality measure of the users is the SINR because now the inter-cell interference has to be taken into account.

Most of the works that evaluate the multi-cell scenario can be categorized in three main schemes: centralized, distributed and hierarchical. These schemes are divided according to the network nodes in which the resource allocation is performed, as indicated below:

- ***Centralized***: Centralized RRM schemes are characterized by the presence of a central entity, for example an RRM server, which is connected to all MBSs and knows the Channel State Information (CSI) of all users in all sub-carriers available in the system. This central controller is responsible for the allocation of all system resources, such as groups of sub-carriers, and the power and MCSs that should be used in each sub-carrier. Of course, this kind of architecture can not be deployed in current systems due to some reasons: the time and uplink channel capacity needed to feedback the CSI of all MTs in all sub-carriers to the central controller; the computational effort necessary to generate the resource allocation; and the signaling overhead to send the assignment information to the users. However, centralized schemes can be used as a benchmark for comparison with more realistic schemes as they are expected to provide the best solutions. Some examples of works that considered the centralized approach are [49–51].

- ***Distributed***: Centralized RRM produces near-optimum channel allocation at the expense of high signaling overhead, so distributed schemes appeared as an alternative to the centralized ones. In distributed RRM algorithms, there is no central entity responsible for receiving measurement reports or distributing resources to MTs or MBSs. In addition, the RRM decisions performed by the MBSs may or may not be supported by information obtained through communication among them. Works [52–54] formulated distributed RRA algorithms for OFDMA systems.

- ***Hierarchical***: There is another class of RRM algorithms in which the allocation procedure is sub-divided in tasks that are performed by different network nodes that are in different hierarchical levels in the system, for example an RRM server and the MBSs. Due to this characteristic, this class of RRM algorithms tends to be a good trade-off between the centralized and distributed ones. For instance, the hierarchical network approach was studied in [23, 55].

Inter-Cell Interference Coordination (ICIC) techniques are very important in a multi-cell scenario in order to mitigate the inter-cell interference and improve the cell edge user throughput. In that sense, a new cellular architecture called Coordinated Multi-Point (CoMP), which provides coordination among multiple cell sites, has been proposed for the 3GPP LTE-Advanced system [56]. In order to accomplish that, CoMP transmission and reception schemes need to use proper interfaces that connect the MBSs in order to exchange the necessary information, e.g. the X2

interface in 3GPP LTE-Advanced. CoMP schemes are potentially efficient in managing interference and transforming a multi-cell scenario into a single-cell one, where there is no inter-cell interference. Besides that, there is a trend in next generation mobile communication networks that the RRM techniques should be executed in the MBSs, not in the radio network controllers anymore, as was the case for the Third Generation (3G) systems. We also have that all the information needed by the RRM techniques proposed in this thesis is available in each MBS locally. Therefore, the RRM techniques for macrocell networks studied/proposed in this thesis, while not focused specifically on interference coordination, can be extended to CoMP scenarios. The reasons explained above support our decision of evaluating the RRA techniques proposed in this thesis in a single-cell scenario.

Among the works found in the literature that considered the single-cell scenario, two main optimization approaches were found [17, 19]:

- ***Margin Adaptive***: The objective of the margin adaptive problem is to minimize the overall transmit power of the MBS while guaranteeing the individual rate requirements of the users.
- ***Rate Adaptive***: The objective is to maximize an objective function based on the users' instantaneous bit rates at each TTI, subject to a constraint on the maximum transmit power of the MBS.

The employment of the relaxation of constraints in order to solve the margin adaptive problem was first presented by [57], which relaxed the integer constraint of the sub-carrier assignment. A sub-optimum solution to this problem was proposed, in which the sub-carrier is assigned to the user that has the greatest share of that sub-carrier. Then, a power loading algorithm is utilized to perform the power distribution among sub-carriers in order to fulfil the rate requirements of the MT with minimum power. Due to the assumed simplifications, the obtained solution gives a lower bound estimation of this minimum required transmit power. However, it is shown in [57] that this sub-optimum solution is very close to the optimum one. This initial work described in [57] served as comparison basis for multiple other subsequent studies.

Some works were based only on heuristics [37, 58]. In [37], the authors propose a heuristic real-time algorithm which consists of two phases: first, an initial sub-carrier allocation is obtained via a constructive algorithm, which is based on ordered lists of sub-carriers; and then iterative swapping of the sub-carrier between users is done on the initial allocation in order to minimize the objective function. Experimental results showed that the performance of this real-time algorithm was close to the one provided by the optimal allocation. The work described in [58] proposed a cross-layer adaptive resource allocation algorithm for packet-switched OFDM systems composed of two parts: virtual clock scheduling, which provides guaranteed performance and fairness from the data link layer's perspective; and adaptive sub-carrier assignment and power allocation, which exploit the system diversities and provide guaranteed transmission efficiency and proportional fairness in the physical layer. Regarding the sub-carrier assignment and power allocation algorithms, the

authors proposed to apply the power allocation algorithm presented in [37] and to use linear integer programming to solve the sub-carrier assignment problem.

The problem splitting approach and heuristics were used in conjunction in [22]. The first step was responsible for determining the number of sub-carriers each terminal should receive and was done using a greedy algorithm. Once the resource allocation is determined for each terminal, the specific assignment of the sub-carriers is done in the second step. Simulations results show that the power requirements of the proposed algorithms are only slightly higher than those of [57] while Central Processing Unit (CPU) run times are smaller by a factor of 100. Other works that employ the combination of the problem splitting approach and heuristic-based optimization are [32, 59, 60].

Now, let us discuss about the rate adaptive approach. There are variants of the rate adaptive optimization depending on the objective function and constraints being considered. For example, [61] formulated an integer programming problem whose objective was to maximize the instantaneous total cell rate and proposed a computationally-efficient greedy algorithm. The algorithm assigns each sub-carrier to the terminal with the highest channel gain. Then, a power loading algorithm is applied in order to distribute the transmit power with respect to the objective function.

Another variant of the rate adaptive approach tries to maximize a lower bound of all terminals' instantaneous rates, which introduces a fairness component into the RRA problem. One of the first efficient algorithms that solved this rate adaptive variant was proposed by [62], which used the problem splitting approach in order to solve it. The first step consisted of combined power and sub-carrier allocation based on the average channel gain calculated over all sub-carriers for each user. The second step was to solve the so-called assignment problem, which decides on the best sub-carrier/terminal pair and has the bipartite weighted matching problem as its graph-theoretic counterpart [17]. The authors on [62] suggested using the Hungarian method [36] to solve the assignment problem. The last step was to apply adaptive power allocation to each terminal in order to distribute its allocated power among the sub-carriers assigned to it. Another work that uses the problem splitting method is [63]. In this paper, a proportional rate adaptive resource allocation method for multi-user OFDM is proposed, where sub-carrier assignment and power allocation are carried out sequentially to reduce the complexity, and an optimal power allocation procedure is derived.

Some proposals to solve the rate adaptive problem were based only on heuristics [64, 65]. Apart from proposing integer relaxation or problem splitting, the authors in [64] base their heuristic algorithm on a constant power assignment for all sub-carriers. They claim that since most sub-carriers assigned to the users will be in quite a good state, the constant power distribution does not reduce the system performance too much. It was shown in [17] that, at least for wireless point-to-point connections, constant power distribution achieves almost the same performance as adaptive power loading. On the other hand, the authors in [65] proposed a heuristic algorithm that performs joint sub-carrier and power allocation considering the frequency selective nature of users' channels.

Some works have used more than one optimization approach to cope with the rate adaptive

problem. For example, a solution combining integer relaxation and problem splitting was proposed in [42]. The optimal sub-carrier and bit allocation problems, which were formulated in [42] as nonlinear optimizations, were converted into linear ones and solved by integer programming. Based on this, a sub-optimum approach that separately performs sub-carrier allocation and bit loading was proposed.

In this thesis, we are interested at studying the fundamental trade-off between efficiency in the resource usage and fairness in the resource distribution among network players. Therefore, we evaluate the rate adaptive optimization with respect to instantaneous rates in chapter 3, as indicated in Fig. 1.1 (see page 12). This is due to the fact that the aforementioned trade-off appears naturally in the rate adaptive problem when different policies concerning capacity or fairness maximization are contemplated.

One flexible optimization tool successfully used for many RRM problems in macrocellular networks is Utility Theory [66]. Using utility-based objective functions in the RRA optimization problems, we are able to formulate RRA problems with respect to several network metrics or resources. Utility-based optimization is specially suited to RRA problems that involve different types of traffic, such as Non-Real Time (NRT) and Real Time (RT) services.

A scheduling algorithm suitable for best-effort data users that maximizes total utility of the system was proposed in [67]. In this work, exponential and logarithmic utility functions were used. The work [35] formulated an RRA optimization problem based on a utility function that takes into account the following aspects: the relation between rate and power, QoS, and priority among users. The goal was to ensure QoS requirements among heterogenous services and improve the system throughput. The utility-approached RRA technique proposed in [35] was able to allocate sub-carriers, transmission power and rate according to the formulated optimization problem. Traffic prioritization is again tackled in [68]. Distinct utility functions dependent on the specific traffic class are combined and maximized, so that different treatment is provided to each data flow. This approach was applied to allocate sub-carriers and time slots of an adaptive OFDMA scheme for services with heterogeneous QoS requirements quantified as delay, jitter or throughput. In [69], a joint transmit scheduling and dynamic sub-carrier and power allocation method was proposed to exploit multi-user diversity in an OFDM wireless network with mixed real-time and non-real-time traffic patterns. The authors considered a utility function that is dependent on the user's experienced delay, the corresponding channel quality and the user's delay requirements.

As commented in section 1.6 of chapter 1, we also use Utility Theory in the work presented in this thesis (see Fig. 1.1). In chapter 4, we propose two utility-based RRA frameworks based on optimization problems that take into account the average data rate (throughput) and the delay of NRT and RT services, respectively.

2.7. RRM for the Downlink of OFDMA Femtocell Networks

RRM for femtocell networks is a nascent area that has thrown up several interesting problems which offer immense opportunities for contributing both from a research as well as an engineering perspective [11].

FAPs are small cellular base stations that are deployed by the end-users on their home or office premisses. Some characteristics of the FAPs are: reduced power and coverage area, low-cost, and direct connection to the backbone IP-based network. They were initially designed by the mobile operators to extend indoor coverage, aiming to solve the problem of coverage holes, improve overall system capacity and offload data traffic from the MBS, allowing the mobile operator to be mainly focused on outdoor and mobile users. Some other key benefits of femtocells are infrastructure cost reduction and signal quality enhancement.

Although femtocells can use any cellular access technology, the object of study of this thesis are femtocell networks based on OFDMA, due to the great flexibility of RRM offered by this system.

Nowadays, cellular operators face a remarkable challenge to provide good coverage to indoor users, specially those ones that use high speed data services. The classic solution of using outdoor macrocells to serve these users is not efficient due to some reasons, among others [12]: 1) this is an expensive approach because a large amount of the MBS transmit power will be used to overcome the indoor penetration losses; 2) high data rate services require good channel conditions in order to be able to use the highest MCSs; in an "outside in" approach, this is only possible if the indoor users are located near windows facing macrocell sites.

Several indoor solutions have been proposed to solve this coverage problem [12]:

- ***Repeaters***: They are components that amplify the outdoor signal coming from the MBS and retransmit it inside a building. In this way, the attenuation of the walls of the building is avoided.

- ***Distributed Antenna System (DAS)***: The main idea is to locate separated antenna elements on different floors of a building and split the power between them. Doing that, a homogeneous coverage and an improved efficiency in the system can be achieved, mainly if the coverage areas of the different antennas do not overlap and fit as much as possible the shape of the building.

- ***Radiating cable***: It is a metallic wire that works as a long antenna. If it is connected directly to a base station, it can receive and transmit the electromagnetic energy all along the cable, which can be deployed in any environment inside the building.

- ***Picocells***: It is a small base station very similar to an Wireless Local Area Network (WLAN) access point. The main advantage of picocells is their price compared with a standard base station. Although they are installed by the cellular operators, the installation cost is also lower compared with the outdoor base stations.

Since repeaters, DAS and picocells are deployed by the network operator, they are more suitable for hotspots in large business centers, office buildings and shopping malls. Moreover, their deployment needs special attention regarding their location and configuration parameters. Therefore, these solutions are not scalable to be used in some scenarios such as Small Office and Home Office (SOHO) and home users (for personal communications and entertaining) due to their high cost. Femtocells are a suitable solution for such scenarios. They can solve the indoor coverage problem and have the advantage for the operators that they already have the functionalities of a Base Station Controller (BSC) and they are in principle paid and deployed by the customers themselves. However, this big scalability must be guaranteed by proper self-organizing RRM techniques that are able to guarantee the seamless co-existence between the macro and femto tiers as well as performing interference management within the femtocell tier.

From an operational point of view, several femtocell access modes have been defined [70]:

- ***Open access***: All users can get access to the network through the femtocell. This is a typical operator-deployed scenario where public access femtocells are deployed in public zones.
- ***Closed access***: Only the users attached to the FAP (e.g. the inhabitants of the home) can get private access to the network through the femtocell. This is a typical user-deployed scenario.
- ***Hybrid access***: Priority is given to the owner of the FAP but additional users can be tolerated.

According to [71], due to security issues, the subscribers of the femtocells in the home environment are not keen to open their access to nonsubscribers unless they obtain some kind of benefit/revenue. In this thesis we assume that the femtocells are deployed in a SOHO and home environments, so we consider closed access mode.

One of the main challenges in a two-tier macro- and femtocell network is how to allocate spectrum between the tiers. Two main approaches are possible [71]:

- ***Spectrum sharing***: In this case the same spectrum is used by both macro and femtocells. That is, the sub-channels of the whole licensed spectrum may be allocated to femto and macro communications. However, sophisticated techniques are required to mitigate cross-tier interference and new radio planning rules have to be elaborated to account for the impact of the random femtocell deployment on the macrocell performance.
- ***Spectrum splitting***: In this case macro and femtocells use orthogonal frequency bands. From the interference point of view, this approach is the simplest one as no cross-tier interference is expected. In such a case, macrocell and femtocell tiers can be considered as totally separated networks.

When femtocells share the same spectrum of the macrocell and are deployed in the closed access mode, coverage holes or dead zones in the operator's macrocell network can appear [11, 71]. In

order to illustrate that, let us consider Fig. 2.3, which depicts a two-tier macro- and femtocell network. The first tier comprehends the traditional macrocellular network, while the second one incorporates several shorter range cells (femtocells) that are distributed in a random manner inside the same geographic area covered by the larger cellular network (also called umbrella macrocell).

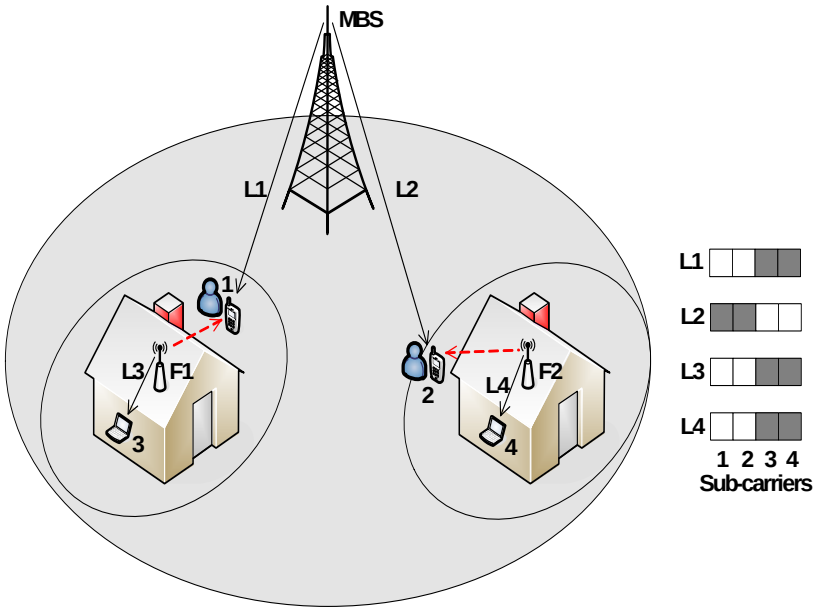


Figure 2.3.: Cross-tier downlink interference in an OFDMA femtocell network using the spectrum sharing approach and closed access mode

In this example, we have two outdoor MTs (1 and 2) connected to the MBS through links L1 and L2, respectively, and two indoor MTs (3 and 4) served by their corresponding FAPs (F1 and F2) through links L3 and L4, respectively. Let us assume that the OFDMA system has 4 sub-carriers to be shared among the macrocell and the femtocells. We assume that the DSA algorithm of the MBS assigns sub-carriers 3 and 4 to MT 1 and sub-carriers 1 and 2 to MT 2. Suppose that FAP F1 is serving MT 3 with sub-carriers 3 and 4, and meanwhile, the outdoor MT 1 is walking down the street and passes by the vicinity of the home of MT 3. In this situation, MT 1 falls inside the coverage area of FAP F1. Since MTs 1 and 3 simultaneously use the same sub-carriers, and the outdoor MT 1 cannot handover to FAP F1 due to the closed access mode, we have that the outdoor MT 1 connected to the MBS is heavily interfered. In this case, it is said that the femtocell created a dead zone or coverage hole around itself, and this problem is worse if the femtocells are in the macrocell edge. Notice that this cross-tier interference can be avoided if proper power control and sub-carrier assignment algorithms are used. For instance, each femtocell could optimize their transmit power so that a small amount of power is leaked outside the home building. Furthermore, intelligent DSA algorithms in the MBS and FAPs could coordinate the sub-carriers assignment to the macro and femto users, just like the successful case of MTs 2 and 4 in Fig. 2.3.

Most of the works found in the literature that address RRM problems for femtocell networks

try to optimize the co-existence between the macrocell and femtocell tiers in the spectrum sharing case. Some works that proposed to use power control to mitigate the cross-tier interference are [72, 73]. A power control method for pilot and data that ensures a constant femtocell radius in the downlink is described in [72]. In order to do that, the authors propose to set the transmit power of each femtocell to a value that is on average equal to the power received from the closest MBS at the target cell radius, subject to a constraint of maximum femtocell transmit power. The work [73] proposed a utility-based power control algorithm that provides link quality protection to an active macrocell user by progressively reducing the SINR targets at strong femtocell interferers when the macrocell user is unable to meet its SINR target.

Some other works use sub-carrier assignment algorithms to mitigate the cross-tier interference, for example [71, 74–76]. Distributed and centralized interference avoidance techniques were evaluated in [71]. The idea was that each femtocell had a sub-carrier priority list, where the ones that suffered the least interference had priority in the allocation. The work [74] proposed a dynamic resource partitioning to mitigate the destructive downlink femto-to-macro interference. The algorithm forces that femtocells are denied access to downlink resources that are assigned to macrocell MTs in their vicinity. It was shown that the interference to the most vulnerable macro MTs is effectively controlled at the expense of a modest degradation in the femtocell capacity. Game Theory was used in [75] and [76] to propose RRM techniques in a scenario where the MBS is underlaid with multiple FAPs and sharing the same spectrum. The former formulated a decentralized interference control strategy that takes into account all kinds of interference in a two-tier network: macro-to-femto, femto-to-femto and femto-to-macro. The proposed solution is comprised of two steps: 1) resource block allocation, which selects the best resource allocation policy for the femto users; and 2) transmission power allocation, which selects the proper power levels for each resource block. The latter compares two game-theoretic approaches: 1) a non-cooperative case, where the MBS and the FAPs (players) behave selfishly aiming at improving their respective achievable rates; and 2) a hierarchical model, where the MBS has priority over the FAPs and they collaborate in order to improve the overall network efficiency. The authors showed that the hierarchical model outperforms the non-cooperative approach in terms of achievable rate and optimal number of deployed FAPs.

Several works and standardization forums [77] agree that the downlink femto-to-macro interference in a co-channel (spectrum sharing) and closed access deployment can be potentially destructive in some cases. One simple strategy to eliminate this problem is to deploy the macro and femto tiers in separated frequency bands (spectrum splitting). It is likely that the mobile operators will choose this solution in early stages of femtocells deployment in order to protect the QoS of the existing macrocell access network. This is the approach assumed in this thesis.

In the spectrum splitting approach, the co-tier interference between neighboring femtocells is still an RRM problem to be solved. Fig. 2.4 illustrates an example of femto-to-femto interference in an OFDMA femtocell network. Co-tier interference among femtocells occurs mainly between immediate neighbors due to low isolation between houses and apartments. In a suburban scenario,

this interference would occur horizontally in adjacent terraced houses, whereas in a dense urban scenario, the FAPs can have horizontal and vertical neighbors in blocks of apartments. Since the deployment of femtocells is uncoordinated, it is likely that several femtocells would be installed in locations close to each other, which would cause mutual interference.

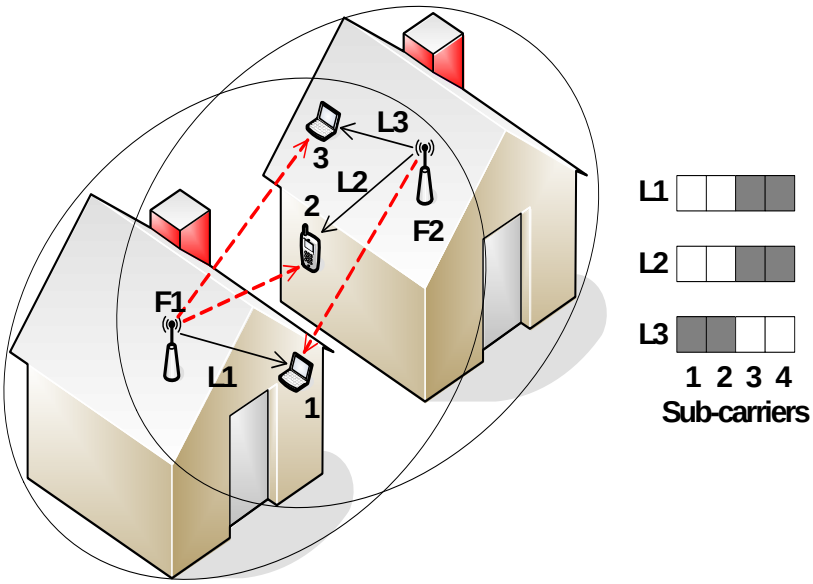


Figure 2.4.: Co-tier downlink interference in an OFDMA femtocell network

Fig. 2.4 presents two FAPs F1 and F2 that are installed in two neighboring houses. FAP F1 serves MT 1 through link L1, whilst FAP F2 serves MT 2 and 3 through links L2 and L3, respectively. Due to the proximity of the FAPs' locations, each MT experience interference links from the non-serving FAP. In this situation, the sub-carrier allocation plays a decisive role in the final impact of the interference. Let us assume that there are 4 sub-carriers available for allocation in each femtocell. Since MTs 2 and 3 are served by the same FAP F2, they are assigned different sub-carriers (see Fig. 2.4). If there is not a coordination between FAPs F1 and F2, it may happen that FAP F1 decides to assign sub-carriers 3 and 4 to MT 1, which will cause (suffer) downlink interference to (from) MT 2. Therefore, interference avoidance techniques based on frequency partitioning among FAPs are necessary to guarantee the seamless co-existence of neighboring femtocells.

Some works have studied interference management in the femtocell tier, for example [78–81]. More details about the state-of-the-art about this specific problem can be found in section 5.3 of chapter 5.

2.8. Fundamental Trade-Off between Efficiency and Fairness

We have chosen a fundamental RRM problem that underlays the investigation presented in this thesis: the trade-off between efficiency and fairness in wireless networks. We further elaborate on

this problem in the following.

It is well-known that wireless communications are characterized by the scarcity of radio resources. Due to this reason, an efficient usage of the available resources becomes mandatory. The resource allocation can follow different criteria. For example, the resources can be assigned to the users that present the best channel quality, or allocated to the FAPs that suffer the least interference. In this case, the channel quality and the received interference can be considered as efficiency indicators of the resources. Opportunistic RRA algorithms were proposed to maximize the efficiency in the resource usage [82]. The term "opportunistic" means that the resources are dynamically allocated to the network players that present the highest efficiency indicator with regard to the radio resources. The trade-off between efficiency and fairness appears when the resources have different efficiency indicators to different network players (multi-user or multi-cell diversity). The use of opportunistic resource allocation to explore these diversities causes unfair situations in the resource distribution.

Without loss of generality, let us consider the case of opportunistic RRA that take into account the channel quality of the users. The objective of such opportunistic techniques is to allocate more resources to the users with better channel conditions, which leads to a higher resource utilization and system capacity. However, an opportunistic strategy benefits the users closer to the MBS, i.e. the ones with highest SNR, which can cause a starvation of the users with worse channel conditions. This can severely degrade some users' experience as a result of unfair resource allocation and increased variability in the scheduled rate and delay. Moreover, long delays in scheduling of packets coming from bad channels can cause severe degradation in the performance of the overall system for higher layer protocols, such as TCP. On the other hand, schemes that provide absolute fairness deal with the worst case scenario, penalizing users with better condition and reducing the system efficiency. It is clear in this case that there is a fundamental trade-off between resource efficiency and user fairness.

From a network operator perspective, it is very important to use the channel efficiently because the available radio resources are scarce and the revenue must be maximized. In the users' point of view, it is more important to have a fair resource allocation in a way that they are not on a starvation/outage situation and their QoS requirements are guaranteed[1]. Then the question is: how can the network operator manage this trade-off? This thesis tries to answer this question and highlights important clues towards this goal.

In order to better understand the aforementioned trade-off, it is indispensable to define what fairness means. There are two main fairness definitions: resource- or QoS-based [83]. In the former, fairness is related to the equality of opportunity to use network resources, for example the number of frequency resources a FAP is allowed to use or the amount of time during which a MT

[1] Mobile operators are becoming increasingly more worried about fairness issues in their networks. According to specialized magazines, 80% of Internet traffic are coming only from 5% of end-users, thereby congesting the network for the rest of the users. A small number of customers use their broadband service inappropriately, for example when sending or downloading very large files, or using 'peer to peer' and file sharing software. In order to solve this problem, network operators are implementing 'Fair Use Policies' in order to manage inappropriate use and make sure the service can be used fairly by everyone.

is permitted to transmit. In the latter, fairness is associated with the equality of utility derived from the network, e.g. flow throughput. Resource and QoS-based fairness is related to the notion of how equal is the number of resources allocated or how similar is the service quality experienced by the players, respectively. If all players in a given instant approximately have the same number of allocated resources, or perceive more or less the same QoS level, we can say that the system provides a high fairness. On the contrary, if the resources are concentrated in the hands of few players, or few of them experience a very good QoS while the others are unsatisfied, the resource allocation can be considered unfair.

Regarding QoS-based fairness, it is well known that the inherent characteristics and transmission requirements of RT traffic differ from those of NRT data traffics. RT services, such as Voice over IP (VoIP) and video conference, require a low and bounded delay, while NRT services, such as World Wide Web (WWW) and FTP, are not delay-sensitive but require an overall high throughput. Due to these factors, the packet delay can be used as fairness indicator in a scenario with RT services, while rate or throughput can be used as fairness indicators in a scenario with NRT services.

Fig. 2.5 depicts a conceptual view of the trade-off between resource efficiency and QoS-based user fairness in a simplified scenario of two users in a wireless system. This conceptual analysis is also valid for the case of resource-based fairness.

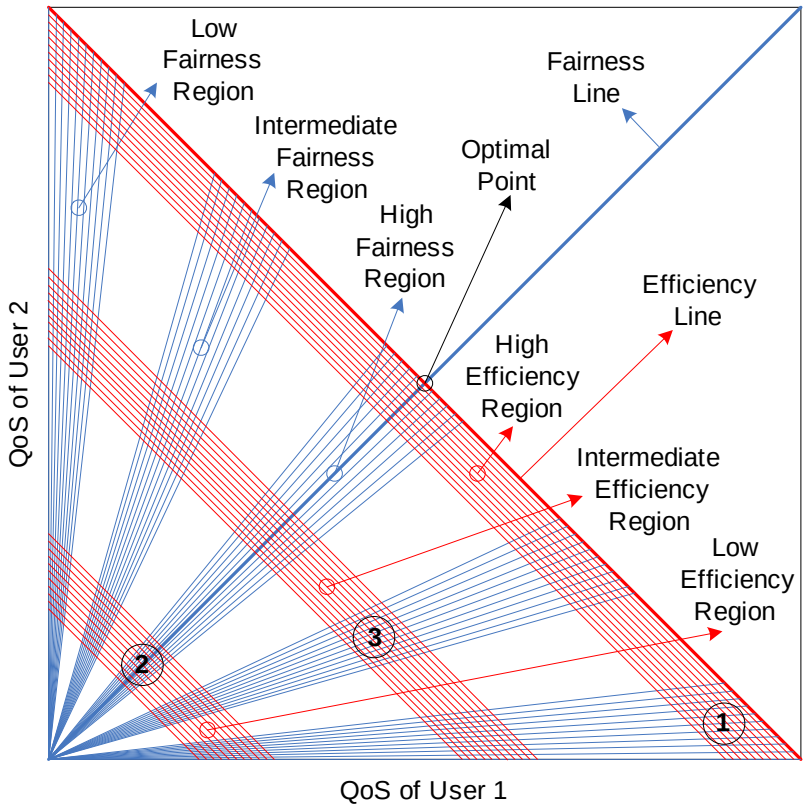


Figure 2.5.: Trade-off between resource efficiency and QoS-based fairness in wireless networks

The axes on the figure present the QoS experienced by the two users after the resource allocation. One can notice that there are two main lines on the figure: fairness and efficiency. The fairness line indicates that the QoS of the users are the same in any point along this line, i.e. the fairness is maximum. Since the radio resources in the wireless system are limited, the efficiency line delimits a capacity region. The crossing between these lines is the optimal network operation point, which characterizes a resource allocation with maximum efficiency and fairness. In the figure, one can see regions of low, intermediate and high fairness and efficiency. Wired networks can effectively work near the optimal point due to the implementation of congestion control techniques, such as TCP [84]. However, the frequency and time-varying wireless channel poses significant challenges to the solution of this problem, and the optimal RRA technique that always provides maximum efficiency and fairness in wireless networks is still an open problem. In fact, in most of the times the optimal point may be unfeasible due to the channel quality of the users.

In order to illustrate that, a scenario is considered where user 1 has better channel conditions than user 2. Region 1 would be the result of an opportunistic RRA policy that gives importance only to the efficiency in the resource usage. In the considered scenario, the majority of the resources were allocated to user 1, while user 2 would starve, causing an unfair situation. On the other hand, region 2 characterizes an RRA policy that provides absolute fairness but causes a significant loss in efficiency since it has to deal with the bad channel conditions of user 2. Finally, region 3 is an example of how an RRA policy can balance these two opposing factors.

As commented before, the fundamental trade-off between efficiency in the resource usage and fairness in the resource distribution is used as a background RRM problem that inspires the RRA techniques proposed in this thesis. Table 2.1 presents the efficiency and fairness indicators that compose the aforementioned trade-off studied in each of the chapters of the thesis.

Table 2.1.: Indicators for the evaluation of the trade-off between efficiency and fairness

	Efficiency indicator	**Fairness indicator**
Chapter 3	Achievable sub-carrier transmission rate	Instantaneous user data rate
Chapter 4	Achievable sub-carrier transmission rate[a] Ratio between achievable sub-carrier transmission rate and user throughput[b]	Average user data rate[a] Head-Of-Line packet delay[b]
Chapter 5	Number of interfering FAPs	Number of frequency resources allocated to the FAPs

[a] Suitable for NRT services.
[b] Suitable for RT services.

Chapter 3

Fairness/Rate Adaptive Resource Allocation for Macrocell Networks

3.1. Introduction

The fundamental problem of the trade-off between resource efficiency and user fairness in the downlink of cellular systems, which was described in section 2.8 of chapter 2, is studied in this chapter for the case of an Orthogonal Frequency Division Multiple Access (OFDMA) macrocell network serving Non-Real Time (NRT) services.

Optimization-based Radio Resource Allocation (RRA) algorithms are one of the best candidates for the solution of this problem. Two RRA algorithms are studied in this chapter: Dynamic Sub-carrier Assignment (DSA), which is responsible for the dynamic allocation of subsets of sub-carriers for different Mobile Terminals (MTs), and Adaptive Power Allocation (APA), which adapts the power level for each sub-carrier according to the instantaneous channel conditions.

As commented in section 2.6 of chapter 2, the optimization-based RRA strategies for OFDMA systems found in the literature typically follow two approaches: *margin adaptive* and *rate adaptive*. The former formulates the dynamic resource allocation with the goal of minimizing the transmitted power with a rate constraint for each user [57, 60]. The latter aims at maximizing the instantaneous data rate with a power constraint [61, 64, 85]. Since the aforementioned trade-off is an explicit consequence of the use of opportunistic rate adaptive RRA algorithms, this latter approach is the one adopted in this study.

The chapter is organized as follows. Section 3.2 presents the state-of-the-art on the management of the trade-off between resource efficiency and user fairness using rate adaptive algorithms and highlights the novel contributions of this thesis regarding this topic. Generalized rate adaptive policies able to cope with fairness constraints are proposed in section 3.3, while the corresponding RRA techniques that implement these policies are presented in section 3.4. The impact of these

RRA techniques on the aforementioned trade-off is evaluated by means of extensive system-level simulations in section 3.5. Finally, the conclusions are drawn in section 3.6.

3.2. Management of the Trade-Off between Resource Efficiency and User Fairness Using Rate Adaptive Optimization

The compromise between efficiency and fairness was conceptually described in section 2.8 of chapter 2. The present chapter extends this conceptual analysis presenting Fig. 3.1, where a novel visualization of the trade-off between these two opposing factors in a wireless network is shown. We introduce the concepts of efficiency-fairness planes, curves and regions. The efficiency-fairness plane is formed by two axis: Fairness Indicator (FI) and Efficiency Indicator (EI), which are any quantitative measures of the fairness among users or the efficiency in the usage of the system resources, respectively. Without loss of generality, we can assume that FI and EI are limited in a range defined by a minimum and a maximum value. Examples of such indicators are the so-called Jain's fairness index proposed in [86], which is the one chosen in this thesis (see section A.7.3 of appendix A), and the system spectral efficiency (capacity), respectively. Fig. 3.1 also shows the optimal network operation point, which characterizes a resource allocation with maximum efficiency and fairness. However, this Radio Resource Management (RRM) problem is difficult to be solved because of the limited number of available radio resources and the variability of the wireless channel in the frequency and time domains.

Now, suppose that there exists an RRA policy that is able to provide several trade-off points, like the ones depicted in Fig. 3.1: low fairness and high efficiency (point 1), intermediate fairness and efficiency (point 2), and high fairness and low efficiency (point 3). If we connect all these points, we draw the efficiency-fairness curve. We define the area below this curve as the efficiency-fairness region. If two RRA policies are compared using the same fairness and efficiency indicators, the one that provides the largest region is the best because it is able to achieve a better balance between these two opposing factors.

Once the concept of the efficiency-fairness trade-off used in this chapter has been clarified, now the different rate adaptive approaches considered in the literature are reviewed. There are three main classic approaches to cope with the rate adaptive optimization problem: Max-Min Rate (MMR) [42, 64], Sum Rate Maximization (SRM) [61], and Sum Rate Maximization with Proportional Rate Constraints (SRM-P) [65, 85, 87–92].

Reference [64] was the first to propose a rate adaptive approach, whose objective was to maximize the minimum rate of the users. The authors proposed a sub-optimum heuristic solution comprised of separate sub-carrier assignment and equal power allocation. This is the fairest policy in terms of data rate distribution because after the resource allocation the users have almost the same rate. The MMR approach was also addressed by [42], which reformulated the optimization problem in order to be solved by Integer Programming techniques. Notice that such a policy considers a worst

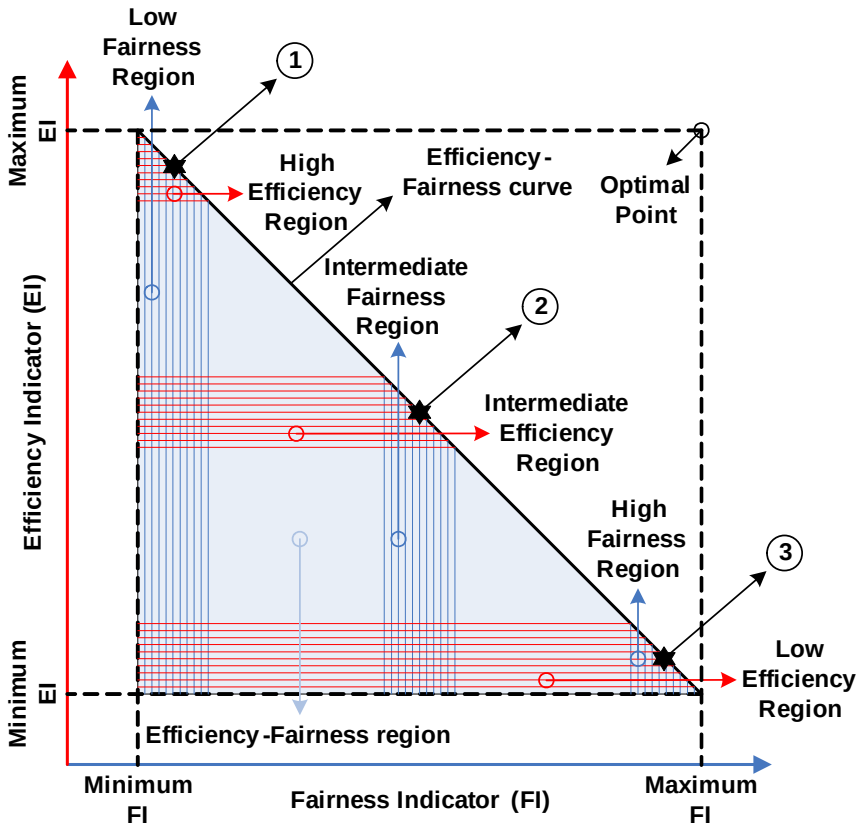


Figure 3.1.: Efficiency-fairness plane in wireless networks

case scenario, in which the user fairness is maximum but the overall system capacity can be very degraded (see point 3 in Fig. 3.1).

The solution of the SRM problem presented in [61] is the classic opportunistic rate adaptive policy, which maximizes the system capacity no matter the Quality of Service (QoS) of the individual users. The sub-carriers are assigned to the users who have the highest channel quality, and next the power is allocated among the sub-carriers following the waterfilling procedure. This resource allocation benefits the users close to the base station and ignores the users with bad channel conditions, who may not receive any resources. According to Fig. 3.1, this policy is located in point 1.

The SRM-P optimization problem was initially proposed by [85] attempting to be a trade-off solution between system capacity and user fairness. The idea was to consider the same objective function as the problem described in [61] and add a new optimization constraint of rate proportionality for each user in order to rule the rate distribution in the system. This new optimization problem is suitable for a scenario where there are different service classes with different proportional rate requirements. The solution was divided in two separate steps: a sub-optimum sub-carrier assignment based on [64] and an optimal power allocation.

The SRM-P problem was further addressed by other references, as explained in the following.

Reference [87] linearized the power allocation problem avoiding the solution of a set of non-linear equations that was required by the solution proposed in [85]. A joint sub-carrier and power allocation was considered in [65], while a priority based sequential scheduling in the sub-carrier assignment procedure was used in [88]. Reference [89] proposed a sub-carrier assignment based on [87] followed by a proportional greedy bit allocation inspired in the margin adaptive approach, which allocates a new bit to the sub-carrier that needs less power to carry an extra bit. A cross-layer technique described in [90] extended the work in [89] by considering the buffer state information in the sub-carrier assignment and power loading algorithms. The two-step approach was also used in [91], where the authors proposed a sub-carrier assignment similar to [87] and a low-complexity iterative power manipulation procedure that provides a feasible solution with the required proportionality. Finally, reference [92] presented a sub-carrier assignment algorithm similar to the one presented in [87], with the difference that the initial number of sub-carriers for each user was calculated taking into account also the user channel gain. Next, a power allocation algorithm that performed rate proportionality tracking was described, where slices of power were re-allocated among sub-carriers until the difference between proportional rates was small enough.

The most relevant and novel contributions of the thesis regarding the control of the trade-off between efficiency and fairness using rate adaptive resource allocation are presented in the following:

1. ***Efficiency-Fairness plane, curves and regions***: This is a powerful performance evaluation tool that launches a new perspective on the analysis of RRM strategies for wireless networks. The efficiency-fairness regions are somewhat similar to the capacity regions, which is a visualization tool widely used in the performance evaluation of cellular systems [93].

2. ***Generalization of classic rate adaptive policies***: To the best of our knowledge, there is not in the literature a rate adaptive policy that provides a controllable trade-off between resource efficiency (maximum system capacity) and user fairness in OFDMA networks. In the present work, three fairness/rate adaptive policies are proposed, which enable the network operator to work on any trade-off point of the efficiency-fairness plane. These proposed policies are called Fairness-Based Sum Rate Maximization (FSRM), Fairness-Based Sum Rate Maximization with Proportional Rate Constraints (FSRM-P) and Fairness-Based Max-Min Rate with Proportional Rate Constraints (FMMR-P). They are generalizations of the classic rate adaptive policies SRM, SRM-P and MMR, respectively. The proposed policies are implemented by RRA techniques composed of DSA and APA algorithms.

3. ***Management of the trade-off between resource efficiency and user fairness***: All proposed fairness/rate adaptive policies rely on the concept of a cell fairness index, which was based on the quantitative fairness measure proposed in [86]. The network operator sets a target fairness index and the RRA techniques will dynamically allocate the resources in order to operate at the desired trade-off point. With this original network management paradigm, the network operator will be able to answer the following question: which network

performance in terms of capacity can be expected under the constraint of, let say, 90% fairness? Or in other words, how much efficiency can be improved by compromising on fairness to a certain extent?

4. ***Fairness-based DSA and APA algorithms***: By means of sub-carrier re-assignment among users and power re-allocation among sub-carriers, the proposed DSA and APA algorithms are able to increase or decrease the fairness in the system according to a given criterion defined by the corresponding fairness/rate adaptive policy.

3.3. Fairness/Rate Adaptive Resource Allocation for OFDMA Systems

In this section, three new proposed fairness/rate adaptive policies are described: FSRM, FSRM-P and FMMR-P. They are generalizations of three classic rate adaptive policies found in the literature: SRM [61], SRM-P [87] and MMR [64]. The generalizations of the aforementioned classic policies take into account a new way to control the trade-off between resource efficiency and user fairness. This control is applied on a cell fairness index and is formulated as a new constraint in the optimization problems. The proposed fairness/rate adaptive policies are described in sections 3.3.1, 3.3.2 and 3.3.3. The classic rate adaptive policies are described in details in appendix B.

All proposed policies share a common framework that can be formulated as an RRA optimization problem, as follows:

$$\max_{\rho_{j,k},p_k} \; f\left(\rho_{j,k},p_k\right), \tag{3.1}$$

$$\text{subject to } \rho_{j,k} = \{0,1\}, \quad \forall j,k, \tag{3.2}$$

$$\sum_{j=1}^{J} \rho_{j,k} = 1, \quad \forall k, \tag{3.3}$$

$$p_k \geq 0, \quad \forall k, \tag{3.4}$$

$$\sum_{k=1}^{K} p_k \leq P_{\text{total}}, \tag{3.5}$$

$$\Phi_{\text{cell}} = \Phi_{\text{cell}}^{\text{target}}, \tag{3.6}$$

where J and K are the total number of MTs and sub-carriers, respectively; $\rho_{j,k}$ is the connection indicator, whose value 0 or 1 indicates whether the kth sub-carrier is assigned to the jth MT or not; p_k is the power of the kth sub-carrier; P_{total} is the Macrocell Base Station (MBS) total transmit power; $f\left(\rho_{j,k},p_k\right)$ is a general objective function that depends on the sub-carriers assignment and power allocation; Φ_{cell} is the instantaneous Cell Fairness Index (CFI); and $\Phi_{\text{cell}}^{\text{target}}$ is the Cell Fairness Target (CFT), i.e. the desired target value of the CFI.

Constraints (3.2) and (3.3) say that each sub-carrier must be assigned to only one user at any instant of time, while constraints (3.4) and (3.5) state that the sub-carriers' powers must be non-negative and the sum of powers among all sub-carriers must be lower or equal to the MBS total

transmit power. A new fairness control mechanism is explicitly introduced into the optimization problem of the fairness/rate adaptive policies by means of the fairness constraint (3.6). This constraint requires that the instantaneous CFI Φ_{cell} must be equal to the CFT $\Phi_{\text{cell}}^{\text{target}}$ at each Transmission Time Interval (TTI). In this way, a short-term (instantaneous) fairness control can be achieved.

The fairness/rate adaptive optimization (3.1)-(3.6) is a mixed binary integer programming problem, because it involves both binary variables $\rho_{j,k}$ and continuous variables p_k. This problem is not convex because the integer constraint (3.2) makes the feasible set nonconvex.

Constraint (3.6) is the main novelty in comparison with the classic rate adaptive policies. It has a deep impact on the design of the RRA techniques used to solve the optimization problem (3.1)-(3.6), as will be shown in section 3.4. In order to better comprehend the importance of this constraint, let us further elaborate on the concept of the fairness index.

In this work, it is assumed that the MTs may belong to different user classes, like bronze, silver and gold. In order to model this fact, it is considered that each MT is associated with a proportional rate requirement $\lambda_j \in [0, 1]$, which is a constant that indicates the proportion of the total system rate that this MT requires. In order to evaluate how close the MT transmission rate is from its required rate proportion, the User Fairness Index (UFI) is defined as

$$\phi_j = \frac{R_j}{\lambda_j}, \tag{3.7}$$

where R_j is the transmission rate of the jth MT. More details about the calculation of R_j can be found in section A.5 of appendix A.

In order to measure the fairness in the rate distribution among all MTs in the macrocell, the CFI is calculated by

$$\Phi_{\text{cell}} = \frac{\left(\sum_{j=1}^{J} \phi_j\right)^2}{J \cdot \sum_{j=1}^{J} \left(\phi_j\right)^2}. \tag{3.8}$$

This index is based on the fairness index proposed by [86] (more details can be found in section A.7.3 of appendix A). In (3.8) we have that J is the number of MTs in the cell and ϕ_j is the UFI of the jth MT given by (3.7). Notice that $1/J \leq \Phi_{\text{cell}} \leq 1$. A perfect fair allocation is achieved when $\Phi_{\text{cell}} = 1$, which means that the instantaneous transmission rates allocated to all MTs are equally proportional to their requirements λ_j (all UFIs are equal). The worst allocation occurs when $\Phi_{\text{cell}} = 1/J$, which means that all resources were allocated to only one MT.

The general RRA problem formulated in (3.1)-(3.6) is particularized in different fairness/rate adaptive policies by considering different objective functions, as will be shown in the following.

3.3.1. Fairness-Based Sum Rate Maximization (FSRM)

The objective function of the FSRM policy is the maximization of system capacity, as indicated below:

$$\max_{\rho_{j,k},p_k} \sum_{j=1}^{J} \sum_{k=1}^{K} \rho_{j,k} \cdot \log_2 \left(1 + p_k \cdot \gamma_{j,k}\right), \tag{3.9}$$

where $\gamma_{j,k}$ is the Channel-to-Noise Ratio (CNR) of the kth sub-carrier with respect to the jth MT (see section A.4.2 in appendix A for more details). The other variables have been already introduced.

The objective function (3.9) is the same of the classic rate adaptive SRM policy [61]. In the original SRM problem, constraint (3.6) does not exist. Therefore, SRM is a pure channel-based opportunistic policy, where the resources are allocated to the users with better channel conditions and the other users are neglected. Such a solution maximizes the cell throughput, but on the other hand it is extremely unfair. Looking at Fig. 3.1, the classic SRM policy could be represented as the point 1 in the efficiency-fairness plane: high resource efficiency and low fairness.

Even though the objective function of the proposed FSRM policy seeks the maximization of capacity, the fairness constraint (3.6) acts as a counterpoint, provoking the explicit appearance of a trade-off. FSRM tries to answer the following question: *how can a given fairness level be achieved while keeping the system capacity as high as possible?* Guided by this criterion, the FSRM policy can achieve different fairness levels and draw a complete efficiency-fairness curve (see Fig. 3.1).

3.3.2. Fairness-Based Sum Rate Maximization with Proportional Rate Constraints (FSRM-P)

The objective function of the FSRM-P policy is also given by (3.9), which means that this policy also aims the maximum efficiency in the usage of the resources. However, this policy considers two constraints that shape the objective function. The first one is the hard fairness constraint given by (3.6). The second is a soft constraint related with rate proportionalities, as indicated in (3.10).

$$R_i : R_j = \lambda_i : \lambda_j, \quad \forall i, j = 1 : J, i \neq j, \tag{3.10}$$

where R_i and λ_i are the instantaneous transmission rate and the proportional rate requirement of a generic MT i, respectively. This constraint was also proposed in the classic rate adaptive SRM-P policy [85, 87]. The soft constraint (3.10) states that the fairness/rate adaptive FSRM-P policy should guarantee as much as possible that the rates allocated to the users follow the proportions stipulated by the constants λ. For example, suppose that we have three MTs with $\lambda_1 = 1$, $\lambda_2 = 2$ and $\lambda_3 = 4$. If the soft constraint (3.10) was totally satisfied, we would have $R_3 = 2R_2 = 4R_1$.

Expression (3.10) is a soft constraint because it does not take part formally of the optimization problem of the FSRM-P policy; it is used only as a design guideline for the corresponding RRA technique. Constraint (3.10) is somewhat incompatible with the hard fairness constraint (3.6). If

the proportional rate requirements λ are met exactly, we have $\Phi_{\text{cell}} = 1$ according to (3.8) and (3.7), and therefore Φ_{cell} cannot be adapted to any desired value of $\Phi_{\text{cell}}^{\text{target}}$, as required by the hard constraint (3.6). Then, the FSRM-P problem has to deal with a scenario where the proportional rate constraint (3.10) has to be fulfilled only partially in order to vary fairness and allow the network operator to have a controllable trade-off between resource efficiency and user fairness. However, the violation of the rate proportionalities in favor of fairness variation has to be done in such a way that system capacity is maximized, as required by the FSRM-P objective function.

Therefore, the FSRM-P policy is guided by the following question: *how can a given fairness level be achieved while keeping the system capacity as high as possible and the users' rates as proportional as possible?*

Notice that the fairness/rate adaptive FSRM-P policy is able to achieve several Cell Fairness Targets (CFTs) and cover a specific region of the efficiency-fairness plane by relaxing the soft constraint (3.10).

3.3.3. Fairness-Based Max-Min Rate with Proportional Rate Constraints (FMMR-P)

Differently from the FSRM and FSRM-P policies, whose objective function was the maximization of the total sum of rates in the system, the objective function of the FMMR-P policy is the maximization of the minimum user's capacity, which is mathematically expressed by

$$\max_{\rho_{j,k},p_k} \min_{j} \sum_{k=1}^{K} \rho_{j,k} \cdot \log_2\left(1 + p_k \cdot \gamma_{j,k}\right). \tag{3.11}$$

When maximizing the worst user's capacity, it is assured that all users achieve a similar data rate, and hence maximum fairness among users is provided.

For this policy, we also assume the general case where the MTs have different proportional rate requirements that should follow the soft constraint (3.10). However, as explained before, the soft constraint (3.10) must be fulfilled to the extent allowed by the hard fairness constraint (3.6).

The fairness/rate adaptive FMMR-P policy is a generalization of the classic rate adaptive MMR policy, which was originally proposed in [64]. MMR provides very high fairness among users at the expense of low system capacity. Such a policy can be hypothetically represented as the point 3 in Fig. 3.1. The consideration of the fairness constraint (3.6) by the FMMR-P policy completely changes the RRA paradigm and allows the possibility of building a complete efficiency-fairness curve in the plane.

The design of the FMMR-P policy follows the reasoning: *how can a given fairness level be achieved while keeping the worst user capacity as high as possible and the users' rates as proportional as possible?*

We will answer in the next section the design guideline questions formulated for each fairness/rate adaptive policy. We present in that section suitable RRA techniques to solve each of the aforementioned RRA problems.

3.4. Proposed Resource Allocation Techniques

The underlying concept behind the fairness/rate adaptive policies is the fact that resource allocation can be based on two possible approaches:

- ***Resource-centric / efficiency-oriented***: the RRA policy allows the resource to "choose" who is the best user to use it;

- ***User-centric / fairness-oriented***: the RRA policy allows the user to choose which is the most adequate resource to him.

Whether the RRA policies use the former, the latter, or both approaches, will determine their ability to control the intrinsic trade-off between resource efficiency and user fairness found in wireless networks.

Three "actors" play an important role in the proposed techniques: the "richest" user (the one with the maximum proportional rate), the "poorest" user (the one with the minimum proportional rate), and the resource.

The proposed fairness/rate adaptive RRA policies are able to increase or decrease the fairness in the system. This process is illustrated in Fig. 3.2. In this hypothetical example, we have the distribution of the QoS among 20 users. The user IDs are ordered in such a way that the users with best QoS are given IDs around 10, and the users with worst service quality are given the extreme IDs (close to 1 or 20). A situation of high fairness is represented by the solid red curve, in which the users have almost the same QoS. If we want to decrease fairness, a policy in which the resources are concentrated in the hands of few best (rich) users must be used, i.e. few users have high QoS while the others are neglected. On the other hand, a QoS distribution depicted by the dashed pink curve shows an unfair resource usage. If fairness is to be increased from that point, the resources, and consequently the QoS, should be divided more equally among the users. This is accomplished removing resources from the rich and giving them to the poor.

As said in section 3.3, the general fairness/rate adaptive problem as formulated in (3.1)-(3.6) is a nonconvex optimization problem, which makes it very difficult to find a joint optimum solution. A sub-optimum approach usually adopted by other works in the literature to solve the rate adaptive problems is to consider the joint optimization problem as two separate parts: first, perform DSA with equal power allocation, and next perform APA with the previously fixed sub-carrier assignment.

This thesis also adopts this problem-splitting approach to propose RRA techniques able to solve the proposed fairness/rate adaptive policies. Each of the proposed policies is implemented by a sequence of three heuristic RRA algorithms, as explained in the following.

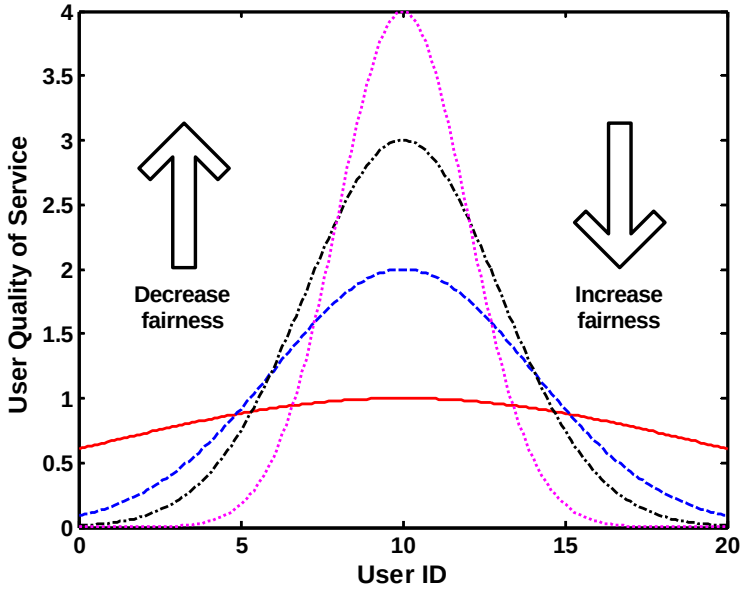


Figure 3.2.: Relation between QoS distribution and fairness adaptation

1. ***Initial Dynamic Sub-carrier Assignment (DSA)***: an initial fairness level (CFI) is achieved after the execution of a classic rate adaptive DSA algorithm. Therefore, an initial positioning on the efficiency-fairness plane is determined. This algorithm is described in details in section 3.4.1.

2. ***Fairness-based DSA***: depending whether the initial CFI is higher than the desired CFT or not, the fairness in the system is decreased or increased by means of sub-carrier re-assignments among users. A rough approximation of the CFT is achieved.

3. ***Fairness-based Adaptive Power Allocation (APA)***: depending on the difference between the CFI and CFT after the fairness-based DSA, the fairness in the system is decreased or increased by means of power re-allocations among sub-carriers of different users. In this way, an accurate approximation of the CFT is achieved.

There is a common philosophy behind both fairness-based DSA and APA algorithms, which is:

1. Fairness variation is only possible if resources are moved between different users. The first step is to decide who is the user from whom a resource will be removed. This choice depends on the RRA policy being considered, and if the fairness must be decreased or increased.

2. Next, remove a small amount of resource (sub-carrier or power) from this user. In the case of sub-carrier re-assignment, the sub-carrier with worst channel quality is removed. In the case of power re-allocation, the power is removed from the sub-carrier of the chosen user that will experience the smallest rate decrement.

3. Finally, give this resource to another user that can take the most benefit of it, or in other words, give this resource to another user who can use it in the most efficient way possible. In the DSA case, this means to assign the removed sub-carrier to the user that has the highest channel gain on it. In the APA case, we give the slice of power to the user that will experience the greatest rate increment among all users.

More details about the fairness-based DSA and APA algorithms are given in sections 3.4.2 and 3.4.3, respectively.

3.4.1. Initial Dynamic Sub-carrier Assignment

The adaptation of the CFI in order to meet the desired CFT has to start with an initial point. The initial DSA algorithm determines this start point using a classic rate adaptive DSA algorithm, namely SRM, SRM-P or Max-Min Rate with Proportional Rate Constraints (MMR-P). In fact, the rate adaptive policy MMR-P is proposed for the first time in this thesis. It is a variant of the classic rate adaptive MMR policy. For more details about the classic rate adaptive policies, see appendix B.

Each proposed fairness/rate adaptive policy will use a different initial DSA algorithm, and so will adapt the CFI starting from different points on the efficiency-fairness plane, as will be explained in section 3.4.4.

3.4.2. Fairness-Based Dynamic Sub-carrier Assignment

The general block diagram of the fairness-based DSA algorithm is depicted in Fig. 3.3 while its detailed pseudo-code is presented in Algorithm 3.1.

It is an iterative heuristic algorithm that adapts the cell fairness index by means of a sub-carrier re-assignment procedure. The algorithm starts with the calculation of the CFI according to (3.8) and tests if it is higher or lower than the cell fairness target. If the CFI is higher than the CFT, the fairness in the system must be decreased; otherwise it must be increased. This is accomplished by an iterative procedure that stops when a rough approximation of the CFT is achieved.

As can be seen in Algorithm 3.1, the fairness decrease procedure (lines 9-21) have some actions in common with the fairness increase (lines 23-34). Below we summarize these actions and highlight some differences between these procedures:

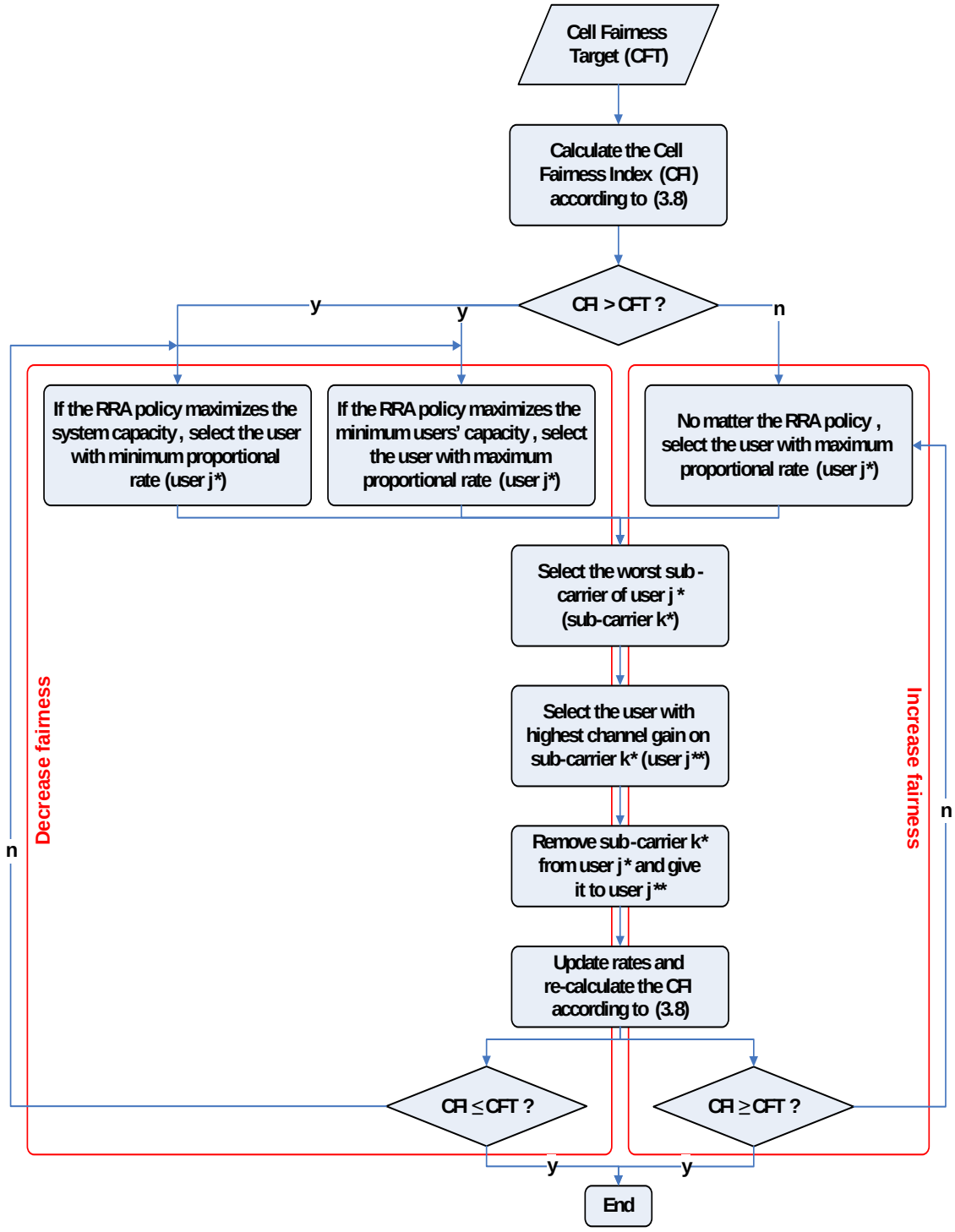


Figure 3.3.: General block diagram of the fairness-based Dynamic Sub-carrier Assignment (DSA) algorithm

1. Select a user j^* from the set of available users in such a way that fairness can be increased (or decreased) if a sub-carrier is removed from this user. This choice depends on the fairness/rate adaptive policy being used. As explained in section 3.4.4, the fairness/rate adaptive RRA policies proposed in this thesis aim either to maximize the system capacity, or to maximize the minimum users' capacity. If fairness needs to be increased, what we have to do is take resources from the richest user (maximum proportional rate) and give them to other users. If the objective is to decrease fairness, we have two possibilities: 1) If the RRA policy being considered aims to maximize the system capacity, we do not care about the users in bad channel conditions, and choose the user with minimum proportional rate to take resources from. 2) Otherwise, if the chosen RRA policy aims to maximize the minimum capacity among all users, we have to protect the resources of the worst user (minimum proportional rate); so the richest user with maximum proportional rate is chosen for resource removal.

2. From the subset of sub-carriers assigned to user j^*, select the one with the minimum channel quality with respect to this user (sub-carrier k^*).

3. Find the user j^{**} (different of user j^*) that has the maximum channel quality on sub-carrier k^*. This is the user that can benefit the most from the resource re-assignment.

4. Sub-carrier re-assignment: remove sub-carrier k^* from user j^* and give it to user j^{**}. The rates and subsets of assigned sub-carriers of users j^* and j^{**} must be updated.

5. Re-calculate the new value of CFI and repeat the process until a rough approximation of the CFT is achieved.

Some comments about the iterative DSA algorithm are listed below:

- Fairness decrease
  - In the fairness decrease procedure, we block sub-carrier k^* after moving it, i.e. we make sure that this sub-carrier will not be re-assigned anymore. This is done because now sub-carrier k^* belongs to the user with the highest channel gain on it, which is the best re-assignment that can be done in terms of fairness decrease.
  - If user j^* does not have any sub-carriers after the resource re-assignment or if they are all blocked, we remove this user from the set of available users.
- Fairness increase
  - During the fairness increase procedure, the sub-carriers have more freedom to move between the users. In order to avoid ping-pong effects, the sub-carrier k^* cannot return to its original owner (user j^*) in subsequent iterations of the algorithm. Due to this restriction, after some iterations, the sub-carrier k^* may not have any user eligible to receive it. In this case, this sub-carrier is removed from the set of available sub-carriers.

Algorithm 3.1 Fairness-Based Dynamic Sub-carrier Assignment (DSA)

Initialization

$\mathcal{M} \leftarrow \{1, 2, 3, \ldots, J\}$; $\mathcal{S} \leftarrow \{1, 2, 3, \ldots, K\}$; $\mathcal{B} \leftarrow \varnothing$ // Initialize users set, sub-carriers set and blocked sub-carriers subset
for all $j \in \mathcal{M}$ and $k \in \mathcal{S}$ **do**
  $\rho_{j,k} \leftarrow 0$ // Reset connection matrix
  $\mathcal{S}_j \leftarrow \varnothing$ // Reset sub-carriers subset of each user
  $\mathcal{Q}_k \leftarrow \varnothing$ // Reset blocked users subset of each sub-carrier
end for

Sub-carrier re-assignment to decrease or increase fairness

Calculate Φ_{cell} according to (3.8)
if $\Phi_{\text{cell}} > \Phi_{\text{cell}}^{\text{target}}$ **then** // Decrease fairness
  while $\Phi_{\text{cell}} > \Phi_{\text{cell}}^{\text{target}}$ **do**
    $j^* \leftarrow$ Select available user ($j \in \mathcal{M}$) according to chosen fairness/rate adaptive policy
    $k^* \leftarrow \arg\min_k \{\gamma_{j^*,k}\}, \forall k \in \mathcal{S}_{j^*}$ and $\forall k \notin \mathcal{B}$ // Find available sub-carrier assigned to user j^* with minimum channel quality
    $j^{**} \leftarrow \arg\max_j \{\gamma_{j,k^*}\}, \forall j \in \mathcal{M}$ // Find available user with maximum channel quality on sub-carrier k^*
    if $j^{**} \neq j^*$ **then**
      Remove sub-carrier k^* from user j^* and give it to j^{**}; update R_{j^*}, $R_{j^{**}}$, $\mathcal{S}_{j^*}$ and $\mathcal{S}_{j^{**}}$
    end if
    $\mathcal{B} \leftarrow \mathcal{B} + \{k^*\}$ // Update set of blocked sub-carriers
    if $(\mathcal{S}_{j^*} \cap \mathcal{B} == \mathcal{S}_{j^*})$ **or** $(\mathcal{S}_{j^*} == \varnothing)$ **then** // If user j^* does not have sub-carriers or if they are all blocked
      $\mathcal{M} \leftarrow \mathcal{M} - \{j^*\}$ // Remove user j^* from the set of available users
    end if
    Re-calculate Φ_{cell} according to (3.8)
  end while
else if $\Phi_{\text{cell}} < \Phi_{\text{cell}}^{\text{target}}$ **then** // Increase fairness
  while $\Phi_{\text{cell}} < \Phi_{\text{cell}}^{\text{target}}$ **do**
    $j^* \leftarrow$ Select available user ($j \in \mathcal{M}$) according to chosen fairness/rate adaptive policy
    $k^* \leftarrow \arg\min_k \{\gamma_{j^*,k}\}, \forall k \in \mathcal{S}_{j^*}$ and $\forall k \notin \mathcal{B}$ // Find available sub-carrier assigned to user j^* with minimum channel quality
    $\mathcal{Q}_{k^*} = \mathcal{Q}_{k^*} + \{j^*\}$ // Update subset of blocked users for sub-carrier k^*
    $j^{**} \leftarrow \arg\max_j \{\gamma_{j,k^*}\}, \forall j \in \mathcal{M}$ and $\forall j \notin \mathcal{Q}_{k^*}$ // Find available user with maximum channel quality on sub-carrier k^*
    if j^{**} exists **then**
      Remove sub-carrier k^* from user j^* and give it to j^{**}; update R_{j^*}, $R_{j^{**}}$, $\mathcal{S}_{j^*}$ and $\mathcal{S}_{j^{**}}$
    else
      $\mathcal{B} = \mathcal{B} + \{k^*\}$ // Update set of blocked sub-carriers
    end if
    Re-calculate Φ_{cell} according to (3.8)
  end while
end if

3.4.3. Fairness-Based Adaptive Power Allocation

After the execution of the fairness-based DSA algorithm described before, an approximate convergence of the CFI is achieved. With the proposed fairness-based APA algorithm, an accurate convergence with a pre-defined small error is possible. This is an iterative algorithm that is able to decrease or increase the fairness level by means of power re-allocations between sub-carriers of different users. The small slice of power that is re-allocated allows the accuracy in the fairness convergence. Fig. 3.4 presents the general block diagram and Algorithm 3.2 sketches the pseudo-code of the fairness-based APA algorithm.

The procedures to decrease or increase fairness are pretty much the same, as can be seen in lines 7-22 and 24-38 of Algorithm 3.2. The main difference is the selection of the user from whom the power will be removed, as will be explained in the following. The procedure of fairness adaptation of the proposed APA algorithm is composed of the following steps:

1. Calculate the CFI using (3.8). This fairness level is the output of the fairness-based DSA described in Algorithm 3.1. This CFI is compared with the CFT, and as a result the proposed APA algorithm will decrease or increase the fairness in the system in order to achieve an accurate approximation of the CFT.

2. The slice of power Δp that will be transferred among sub-carriers of different users is calculated by:

$$\Delta p = \max\left\{\frac{|\Phi_{\text{cell}} - \Phi_{\text{cell}}^{\text{target}}|}{1-(1/J)}, \frac{1}{J}\right\} \cdot \frac{P_{\text{total}}}{K}. \tag{3.12}$$

   This is a dynamic power step that is re-calculated at each iteration of the algorithm. This power slice depends on the difference between the current CFI and the desired CFT. According to the expression of the CFI given by (3.8), the maximum difference between the CFI and CFT can be $1-(1/J)$, where J is the number of MTs. When this happens, the power step Δp will have the maximum value of P_{total}/K. This allows a faster convergence in the initial iterations. When the difference between CFI and CFT decreases, Δp also decreases, which provides a finer tuning of the fairness adaptation when the CFI is getting closer and closer to the CFT. According to (3.12), the minimum value that Δp can reach is $P_{\text{total}}/(K \cdot J)$.

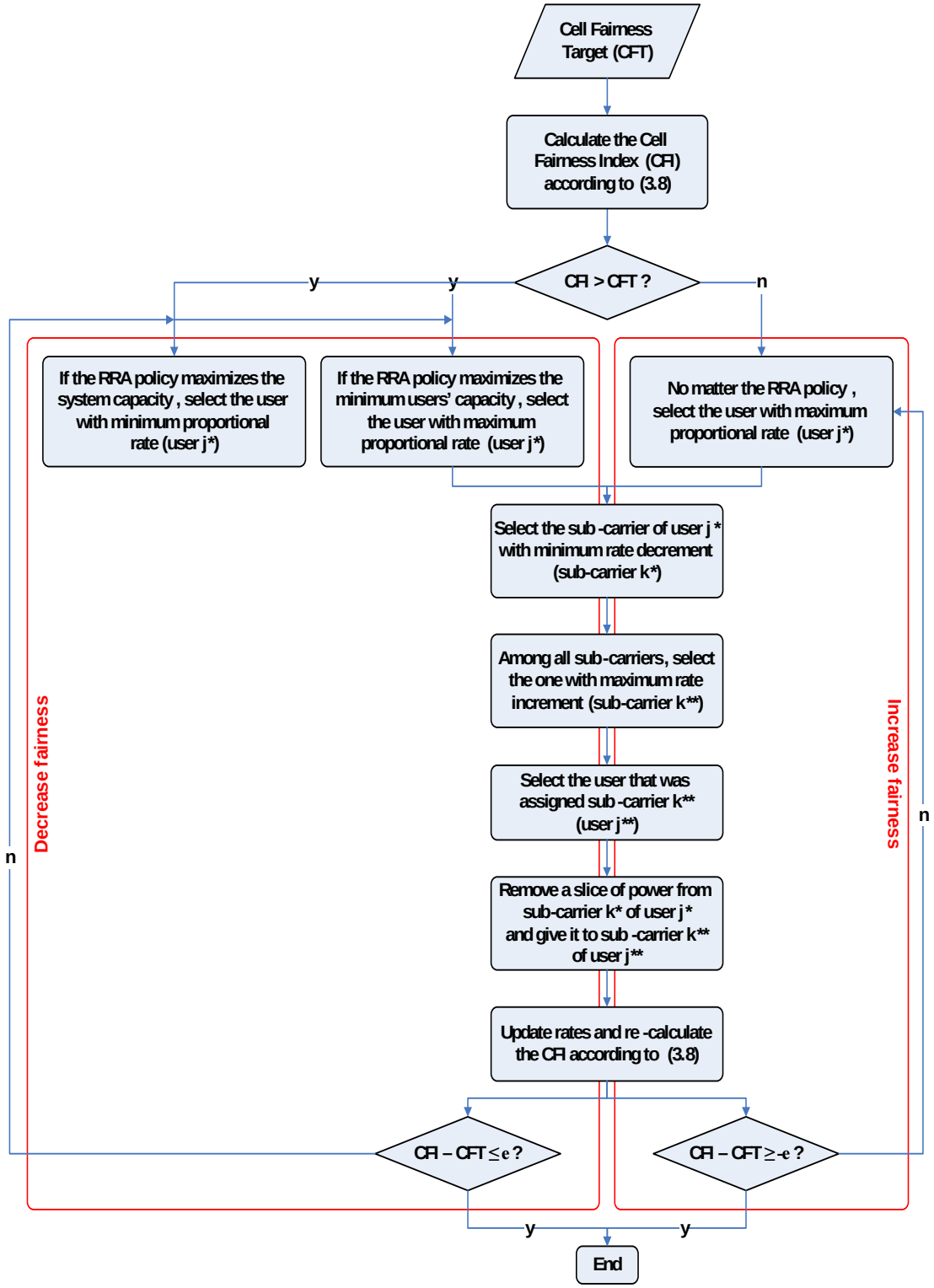


Figure 3.4.: General block diagram of the fairness-based Adaptive Power Allocation (APA) algorithm

3. Select an available user j^* from whom a slice of power will be removed. This selection depends on two factors: 1) the fairness/rate adaptive policy being considered (see more details in section 3.4.4); and 2) the direction of fairness adaptation (decrease or increase). If we want to decrease fairness, we have two possibilities. On one hand, if the RRA policy aims to maximize the overall system capacity, we take resources from the users in bad channel conditions (minimum proportional rate), and give them to the users that can have the highest profit from its usage. This process can potentially starve some users until a point where some of them will not have any resources. These users must be left out for the subsequent iterations of the algorithm for further fairness decrease. On the other hand, if the objective of the RRA policy is to maximize the capacity of the worst users, the iterative APA algorithm has to avoid degrading the QoS of the users in bad channel conditions. In this way, the users with maximum proportional rate must be selected for resource removal. However, in the case of fairness increase, no matter which RRA policy is being considered, what has to be done is remove power slices from the sub-carriers of the richest users with maximum proportional rates and transfer them to the other users, so that the QoS in the system can be shared more equally.

4. After selecting the user j^*, we have to decide the sub-carrier k^* belonging to that user from which power will be removed. This is done by calculating a rate decrement for all the sub-carriers belonging to user j^* considering that a slice of power Δp would be removed from them. The chosen sub-carrier k^* is the one that presents the smallest rate decrement.

5. For all sub-carriers of the other users (different of j^*), a rate increment is calculated considering that a slice of power Δp would be added to them. The sub-carrier k^{**} with the highest rate increment is selected.

6. Find the user j^{**} that was assigned sub-carrier k^{**}.

7. Perform the power transfer: remove an amount of power Δp from sub-carrier k^* of user j^* and add it to sub-carrier k^{**} of user j^{**}.

8. This power re-allocation changes the rates experienced by the users and has a direct impact on fairness decrease or increase. The CFI must be re-calculated by (3.8) and the process is repeated until the target CFT is met with sufficient accuracy (absolute error lower than a pre-defined value ϵ).

Algorithm 3.2 Fairness-Based Adaptive Power Allocation (APA)

Initialization

$\mathcal{M} \leftarrow \{1, 2, 3, \ldots, J\}$; $\mathcal{S} \leftarrow \{1, 2, 3, \ldots, K\}$; $\mathcal{U} \leftarrow \varnothing$; $\epsilon \leftarrow 10^{-3}$ // Initialize users set, sub-carriers set, blocked users subset and error tolerance of fairness index

for all $k \in \mathcal{S}$ **do**

  $p_k \leftarrow P_{\text{total}}/K$ // Sub-carriers' powers are initialized with equal values

end for

Power re-allocation to decrease or increase fairness

Calculate Φ_{cell} according to (3.8)

if $\Phi_{\text{cell}} > \Phi_{\text{cell}}^{\text{target}}$ **then** // Decrease fairness

  while $\Phi_{\text{cell}} - \Phi_{\text{cell}}^{\text{target}} > \epsilon$ **do**

    Calculate Δp according to (3.12) // Adaptive power step

    $\mathcal{U} \leftarrow \arg_j\{R_j == 0\}$ // Update blocked users subset

    $j^* \leftarrow$ Select available user ($j \notin \mathcal{U}$) according to chosen fairness/rate adaptive policy

    for all $k \in \mathcal{S}_{j^*}$ **do**

      $\Delta r_k^{\text{dec}} \leftarrow \frac{B}{K} \cdot \log_2\left(1 + p_k \cdot \gamma_{j^*,k}\right) - \frac{B}{K} \cdot \log_2\left(1 + (p_k - \Delta p) \cdot \gamma_{j^*,k}\right)$

    end for

    $k^* \leftarrow \arg\min_k\{\Delta r_k^{\text{dec}}\}, \forall k \in \mathcal{S}_{j^*}$ // Find sub-carrier of user j^* with min rate decrement

    for all $j \neq j^*$ and $k \in \mathcal{S}_j$ **do**

      $\Delta r_k^{\text{inc}} \leftarrow \frac{B}{K} \cdot \log_2\left(1 + (p_k + \Delta p) \cdot \gamma_{j,k}\right) - \frac{B}{K} \cdot \log_2\left(1 + p_k \cdot \gamma_{j,k}\right)$

    end for

    $k^{**} \leftarrow \arg\max_k\{\Delta r_k^{\text{inc}}\}, \forall k \neq k^*$ // Find sub-carrier with maximum rate increment

    $j^{**} \leftarrow \arg_j\{\rho_{j,k^{**}} == 1\}, \forall j \in \mathcal{M}$ // Find user who has been assigned sub-carrier k^{**}

    Remove Δp from sub-carrier k^* and give it to sub-carrier k^{**}; update R_{j^*} and $R_{j^{**}}$

    Re-calculate Φ_{cell} according to (3.8)

  end while

else if $\Phi_{\text{cell}} < \Phi_{\text{cell}}^{\text{target}}$ **then** // Increase fairness

  while $\Phi_{\text{cell}} - \Phi_{\text{cell}}^{\text{target}} < -\epsilon$ **do**

    Calculate Δp according to (3.12) // Adaptive power step

    $j^* \leftarrow$ Select available user ($j \in \mathcal{M}$) according to chosen fairness/rate adaptive policy

    for all $k \in \mathcal{S}_{j^*}$ **do**

      $\Delta r_k^{\text{dec}} \leftarrow \frac{B}{K} \cdot \log_2\left(1 + p_k \cdot \gamma_{j^*,k}\right) - \frac{B}{K} \cdot \log_2\left(1 + (p_k - \Delta p) \cdot \gamma_{j^*,k}\right)$

    end for

    $k^* \leftarrow \arg\min_k\{\Delta r_k^{\text{dec}}\}, \forall k \in \mathcal{S}_{j^*}$ // Find sub-carrier of user j^* with min rate decrement

    for all $j \neq j^*$ and $k \in \mathcal{S}_j$ **do**

      $\Delta r_k^{\text{inc}} \leftarrow \frac{B}{K} \cdot \log_2\left(1 + (p_k + \Delta p) \cdot \gamma_{j,k}\right) - \frac{B}{K} \cdot \log_2\left(1 + p_k \cdot \gamma_{j,k}\right)$

    end for

    $k^{**} \leftarrow \arg\max_k\{\Delta r_k^{\text{inc}}\}, \forall k \neq k^*$ // Find sub-carrier with maximum rate increment

    $j^{**} \leftarrow \arg_j\{\rho_{j,k^{**}} == 1\}, \forall j \in \mathcal{M}$ // Find user who has been assigned sub-carrier k^{**}

    Remove Δp from sub-carrier k^* and give it to sub-carrier k^{**}; update R_{j^*} and $R_{j^{**}}$

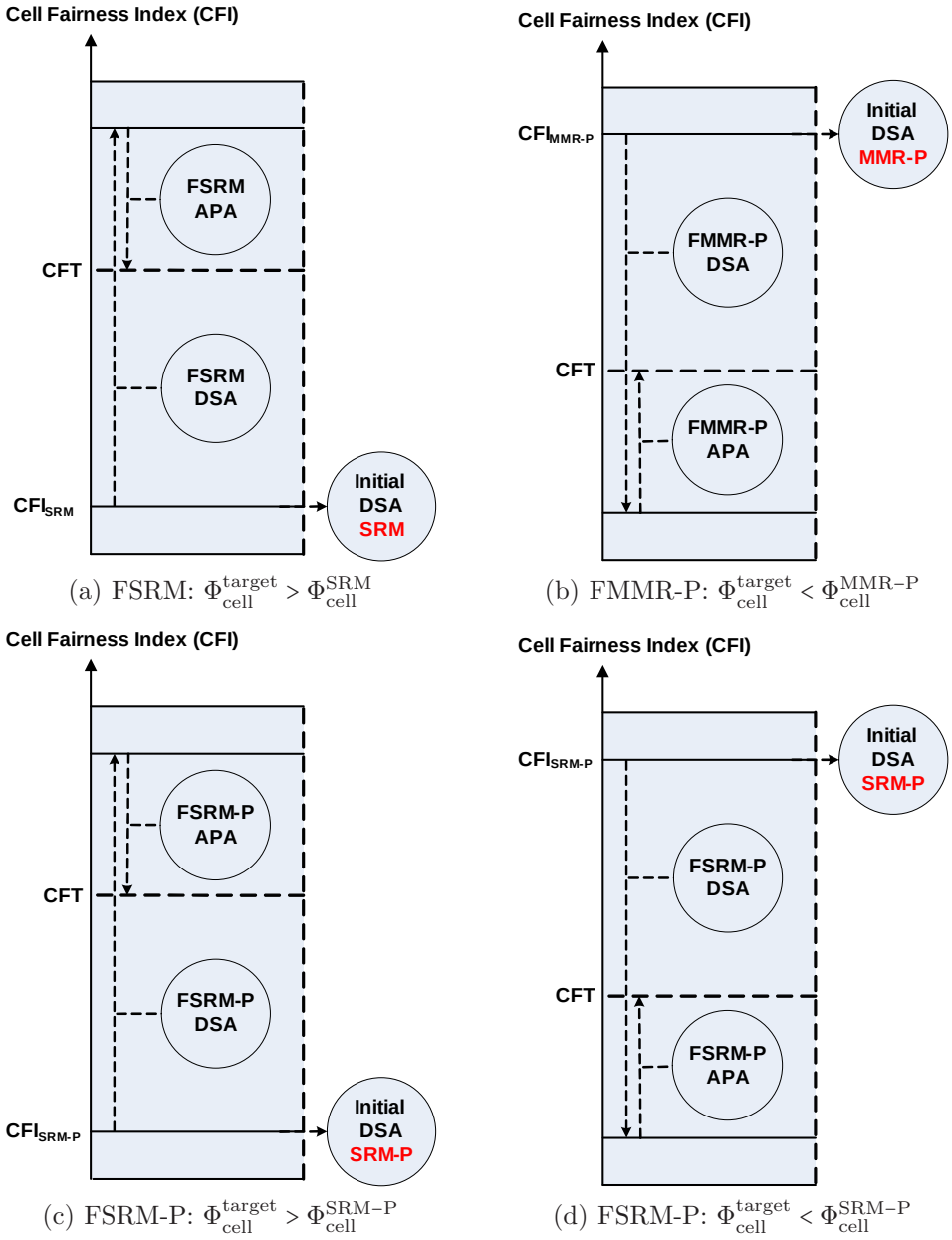
    Re-calculate Φ_{cell} according to (3.8)

  end while

end if

3.4.4. Particularization of the Fairness/Rate Adaptive Policies

This section shows the particular uses that the fairness/rate adaptive policies make of the three RRA algorithms described in the previous sections, namely initial DSA, fairness-based DSA and fairness-based APA. The particular schemes are presented in Fig. 3.5.



(a) FSRM: $\Phi_{\text{cell}}^{\text{target}} > \Phi_{\text{cell}}^{\text{SRM}}$ (b) FMMR-P: $\Phi_{\text{cell}}^{\text{target}} < \Phi_{\text{cell}}^{\text{MMR-P}}$

(c) FSRM-P: $\Phi_{\text{cell}}^{\text{target}} > \Phi_{\text{cell}}^{\text{SRM-P}}$ (d) FSRM-P: $\Phi_{\text{cell}}^{\text{target}} < \Phi_{\text{cell}}^{\text{SRM-P}}$

Figure 3.5.: Particularization of the fairness/rate adaptive policies with regard to the initial DSA, and fairness-based DSA and APA algorithms

Fairness-Based Sum Rate Maximization

Fig. 3.5(a) shows the RRA strategy adopted by the FSRM technique. The objective function (3.9) of the FSRM policy is the maximization of the system capacity. Thus, we propose to perform an initial sub-carrier assignment that yields maximum efficiency in the resource usage. This can be accomplished by the classic rate adaptive SRM policy (see section B.2 in appendix B for more details).

In terms of fairness, the SRM DSA assignment results in low levels of CFI, as indicated in Fig. 3.5(a). Therefore, it is most likely that the desired CFT value $\Phi_{\text{cell}}^{\text{target}}$ is higher than the initial CFI provided by SRM, which is represented by the variable $\Phi_{\text{cell}}^{\text{SRM}}$ in the figure. Based on that, the fairness-based DSA algorithm of the FSRM technique must increase the fairness until a value close to $\Phi_{\text{cell}}^{\text{target}}$.

In order to increase fairness, the FSRM DSA algorithm takes a small amount of resources (worst sub-carriers) from the richest user, i.e. the user with maximum proportional rate, and gives them to another user who can make the best use of it. Looking back to Fig. 3.3 and Algorithm 3.1 (see line 24), FSRM uses the following criterion to select the user from whom a sub-carrier will be removed:

$$j^* = \arg\max_j \{R_j/\lambda_j\}, \forall j \in \mathcal{M}, \tag{3.13}$$

where R_j and λ_j are the transmission rate and proportional rate requirement of the jth user, respectively, and $\mathcal{M}$ is the set of available users.

The FSRM DSA algorithm stops the sub-carrier re-assignment procedure when the CFI Φ_{cell} is just above $\Phi_{\text{cell}}^{\text{target}}$. This assures a rough approximation around the target value. After that, the FSRM APA algorithm re-allocates slices of power between users in order to decrease fairness until the difference between Φ_{cell} and $\Phi_{\text{cell}}^{\text{target}}$ is sufficiently small (see lines 7-22 of Algorithm 3.2). Aiming to decrease fairness, the FSRM APA algorithm takes a small amount of resources (power from the best sub-carriers) from the poorest user, i.e. the user with minimum proportional rate, and gives it to another user who can make the best use of it. The selection of the user that will be degraded follows the following criterion:

$$j^* = \arg\min_j \{R_j/\lambda_j\}, \forall j \notin \mathcal{U}, \tag{3.14}$$

where $\mathcal{U}$ is the subset of blocked users. The user selection formulated by (3.14) must be used in line 10 of Algorithm 3.2. After the power re-allocation procedure, the fairness requirement is met with sufficient accuracy.

As can be noticed, the way the resources are re-allocated in the FSRM policy guarantees that a desired CFT is met while maximum capacity is achieved.

Fairness-Based Sum Rate Maximization with Proportional Rate Constraints

The objective of FSRM-P is to maximize the system capacity while guaranteing a hard fairness constraint and maintaining the users' rates as proportional as possible. Figs. 3.5(c) and 3.5(d) show a general scheme of how fairness is adapted using the FSRM-P technique. We propose to use the DSA assignment of the classic rate adaptive SRM-P policy as the initial sub-carrier assignment (see section B.3 on appendix B for algorithmic details). After this initial resource allocation, we have the CFI Φ_{cell} equal to $\Phi_{\text{cell}}^{\text{SRM-P}}$, which is a value lower than one (generally 0.75). This is due to the fact that the SRM-P sub-carrier assignment procedure roughly meets the rate proportions, and so this algorithm alone is not able to provide maximum fairness (see [87] and appendix B for more details). The CFT $\Phi_{\text{cell}}^{\text{target}}$ can be higher or lower than $\Phi_{\text{cell}}^{\text{SRM-P}}$, which calls for different approaches that are illustrated in Figs. 3.5(c) and 3.5(d), respectively.

If $\Phi_{\text{cell}}^{\text{target}} > \Phi_{\text{cell}}^{\text{SRM-P}}$ (Fig. 3.5(c)), we have a situation similar to the one tackled by the FSRM policy: fairness must be increased from an initial point in order to be closer to the desired target value. Therefore, proper actions must be taken by the FSRM-P DSA algorithm, which is to remove the worst sub-carrier of the best user (maximum proportional rate) according to (3.13) and give it to the user with the highest channel quality to that sub-carrier. This process is repeated until Φ_{cell} is roughly above $\Phi_{\text{cell}}^{\text{target}}$. After that, a power re-allocation procedure is needed in order to decrease fairness a little and exactly meet the target. In this case, the FSRM-P APA algorithm works in the same way as the FSRM APA algorithm previously described.

On the other hand, if $\Phi_{\text{cell}}^{\text{target}} < \Phi_{\text{cell}}^{\text{SRM-P}}$ (Fig. 3.5(d)), fairness must be firstly decreased by the proper sub-carrier re-assignment procedure (lines 9-21 of Algorithm 3.1). Fairness is decreased by taking a small amount of resources (worst sub-carrier) from the poorest user (minimum proportional rate) and next giving it to another user who can make the best use of it. The process stops when Φ_{cell} is slightly below $\Phi_{\text{cell}}^{\text{target}}$. At this stage, fairness should be slightly increased by the FSRM-P APA algorithm (lines 24-38 of Algorithm 3.2), so that the fairness requirement is met with sufficient accuracy. In order to increase fairness, the user selected for power removal is the one with maximum proportional rate (see expression (3.13)).

In both directions, i.e. fairness decrease or increase, FSRM-P can meet the fairness requirement and maximize system capacity compromising the least the rate proportionalities.

Fairness-Based Max-Min Rate with Proportional Rate Constraints

The FMMR-P technique aims to protect the QoS of the worst users and at the same time achieve any desired fairness level while keeping the users' rates as proportional as possible. The initial sub-carrier assignment is given by the MMR-P DSA algorithm, which is an extension of the classic rate adaptive MMR policy [64] (a detailed explanation is given in section B.4 of appendix B). The initial MMR-P DSA assignment has the property of providing very high fairness levels. Therefore, we can expect that the target value of the cell fairness index will be lower than $\Phi_{\text{cell}}^{\text{MMR-P}}$, as shown in Fig. 3.5(b).

The FMMR-P DSA algorithm is executed in order to decrease fairness and reach the closest point below $\Phi_{\text{cell}}^{\text{target}}$ (see lines 9-21 of Algorithm 3.1). This is accomplished by removing the worst sub-carrier of the user with maximum proportional rate and assigning it to the user that has the highest channel gain to it. This means that the fairness is being decreased in an efficient way as far as resources are concerned. On the other hand, this resource re-assignment prevent the users with minimum capacity from decreasing their rates, as required by the FMMR-P objective function.

The next phase is performed by the FMMR-P APA algorithm, which has to take some actions in order to provide a slight fairness increase. Again, the 'richest' user is selected (user with maximum proportional rate), and some amount of power is removed from his worst sub-carrier and transferred to a sub-carrier of another user that can experience the highest rate increment. At each iteration, the fairness index is re-calculated and the process is repeated until $\Phi_{\text{cell}}^{\text{target}}$ is met with a small error tolerance.

Table 3.1 presents a summary of the features of the three proposed fairness/rate adaptive techniques with regard to the fairness adaptation schemes.

Table 3.1.: Features of the fairness adaptation schemes of the fairness/rate adaptive techniques

Policies	**RRA Algorithms**		
	Initial DSA	**DSA**	**APA**
FSRM	Low initial fairness $\Phi_{\text{cell}}^{\text{SRM}}$	Fairness increase until $\Phi_{\text{cell}} > \Phi_{\text{cell}}^{\text{target}}$	Fairness decrease until $\Phi_{\text{cell}} \approx \Phi_{\text{cell}}^{\text{target}}$
FSRM-P	Intermediate initial[a] fairness $\Phi_{\text{cell}}^{\text{SRM-P}}$	if $\Phi_{\text{cell}}^{\text{target}} > \Phi_{\text{cell}}^{\text{SRM-P}}$, fairness increase until $\Phi_{\text{cell}} > \Phi_{\text{cell}}^{\text{target}}$. Otherwise, fairness decrease until $\Phi_{\text{cell}} < \Phi_{\text{cell}}^{\text{target}}$.	if $\Phi_{\text{cell}}^{\text{target}} > \Phi_{\text{cell}}^{\text{SRM-P}}$, fairness decrease until $\Phi_{\text{cell}} \approx \Phi_{\text{cell}}^{\text{target}}$. Otherwise, fairness increase until $\Phi_{\text{cell}} \approx \Phi_{\text{cell}}^{\text{target}}$.
FMMR-P	High initial fairness $\Phi_{\text{cell}}^{\text{MMR-P}}$	Fairness decrease until $\Phi_{\text{cell}} < \Phi_{\text{cell}}^{\text{target}}$	Fairness increase until $\Phi_{\text{cell}} \approx \Phi_{\text{cell}}^{\text{target}}$

[a] The initial CFI $\Phi_{\text{cell}}^{\text{SRM-P}}$ provided by the classic SRM-P DSA algorithm is around 0.75 (see section B.3 of appendix B for more details).

3.5. Simulation Results

The performance evaluation of the classic rate adaptive and the proposed fairness/rate adaptive techniques is presented in this section. Table 3.2 shows the parameters considered in the system-

level simulations, where the main characteristics of an OFDMA-based macrocellular network were modeled. More details about the simulator structure and models can be found on appendix A.

Table 3.2.: Simulation parameters for the evaluation of the fairness/rate adaptive techniques

Parameter	Value
Number of cells	1
Maximum BS transmission power	1 W
Cell radius	500 m
MT speed	static
Carrier frequency	2 GHz
Number of sub-carriers	192
Effective sub-carrier bandwidth	14 kHz
Path loss	using (A.1)
Log-normal shadowing standard dev.	8 dB
Small-scale fading	Typical Urban (TU)
AWGN power per sub-carrier	-123.24 dBm
BER requirement	10^{-6}
Link adaptation	using (A.9)
Transmission Time Interval (TTI)	0.5 ms
NRT traffic model	Full buffer
User satisfaction requirement	512 kbps
Proportional rate requirements[a] (λ)	$1/J$
Target CFI ($\Phi_{\text{cell}}^{\text{target}}$)	Variable
Number of independent snapshots	10000

[a] In the simulations we considered that all users had the same proportional rate requirements λ, which is given by $1/J$, where J is the number of users.

The fairness/rate adaptive techniques are evaluated based on the following performance metrics and analyses:

- Preliminary evaluation of the classic rate adaptive strategies: Cumulative Distribution Functions (CDFs) and bar plots of users' rates;
- Convergence of the fairness/rate adaptive techniques;
- Rate distribution and rate proportionality analyses: CDFs and bar plots of users' rates;
- Fairness analysis: cell fairness index based on instantaneous rates versus the number of users;
- Efficiency analysis: total cell throughput versus the number of users and efficiency-fairness plane;

- User satisfaction analysis: satisfaction versus the number of users and satisfaction-fairness plane. A MT is considered satisfied if its transmission rate is higher than a threshold (see the value considered in the simulations in Table 3.2);

- Central Processing Unit (CPU) time: CPU time-fairness plane;

In the evaluation of the proposed fairness/rate adaptive strategies, two approaches were considered: joint and Equal Power Allocation (EPA). In the former, the initial DSA, the fairness-based DSA and the fairness-based APA algorithms are executed (see section 3.4). The latter approach means that after the initial DSA and the fairness-based DSA, an equal power allocation is executed, i.e. the total transmission power is equally divided among the sub-carriers.

Although the formulation of the fairness/rate adaptive policies is general and contemplate any proportional rate requirements (constants λ), the simulation results presented in this thesis consider equal requirements among the users. We consider this standard scenario as a proof-of-concept that facilitates the understanding of the intrinsic mechanisms behind the adaptive algorithms.

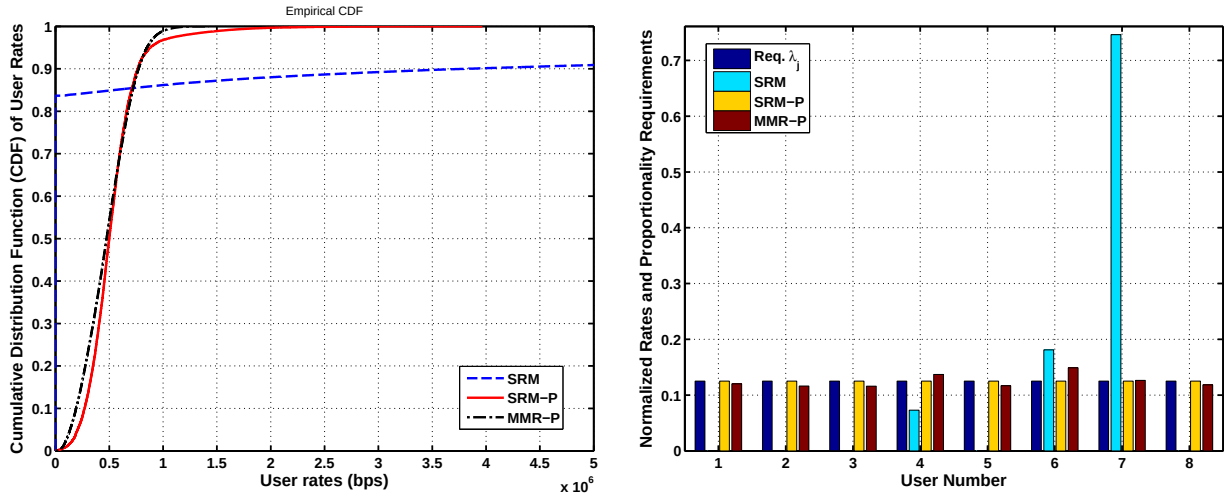
3.5.1. Preliminary Analysis of the Classic Rate Adaptive Techniques

It was shown in section 2.8 of chapter 2 and reinforced in section 3.2 that there is an intrinsic trade-off between resource efficiency and user fairness in wireless networks. Trying to maximize one of them causes the detriment of the other. We also deduced that classic rate adaptive techniques found in the literature have fixed performance regarding the balance between these two conflicting factors. This logical deduction is demonstrated to be true with the simulation results presented in this section.

Fig. 3.6(a) presents the CDF of the users' transmission rates for the classic rate adaptive strategies considering a scenario with 16 NRT users in the macrocell. It can be observed that the SRM technique, which uses a pure opportunistic policy that allocates the resources only to the best users, is the one that presents the highest rates. However, this benefit comes at the expense of a very unfair distribution of the QoS among the users, since approximately 83% of them do not have the opportunity to transmit due to the lack of resources.

At the other extreme we have the SRM-P and MMR-P techniques. Their CDFs are much steeper than the SRM curve, which means that the transmission rates of the users are more equalized, and therefore the fairness in the system is higher. However, the transmission rates of the users are also lower, which characterizes a capacity loss. Notice that the users' rates provided by SRM-P are slight higher than the ones achieved with MMR-P. This difference is due to the implementation of the DSA and APA algorithms of each of these techniques. SRM-P takes extra actions that allow a better utilization of the resources (see appendix B for more details).

The QoS performance of the classic policies and its impact on the fairness can also be visualized in Fig. 3.6(b), where an example of bar plots of the transmission rates considering a scenario with 8 users is presented. This figure also shows in the blue bars the proportional rate requirements,



(a) Cumulative Distribution Function (CDF) of the transmission rates considering a scenario with 16 users

(b) Example of bar plots of the transmission rates considering a scenario with 8 users

Figure 3.6.: Preliminary analysis of the classic rate adaptive policies

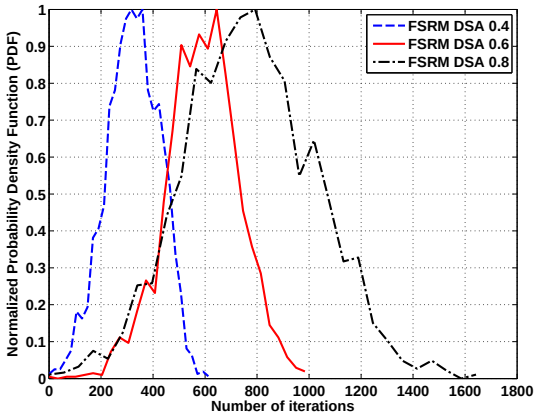
represented by the constants λ. It can be clearly seen that the opportunistic SRM privileges some users, while the others are forgotten. It does not take into account the proportional rate requirements. On the other hand, the SRM-P and MMR-P policies, which have the proportional rate constraints on their respective optimization problems, are able to meet these requirements. This provides a high fairness at the expense of lower system capacity.

From this preliminary analysis, it can be concluded that the classic rate adaptive strategies show very specific patterns regarding the competition between resource efficiency and user fairness. This motivated us to propose general fairness/rate adaptive techniques that could provide any desired fairness level. In this way, the trade-off between efficiency and fairness could be controlled by the network operator. Details about the performance evaluation of these generalized strategies are presented in the following sections.

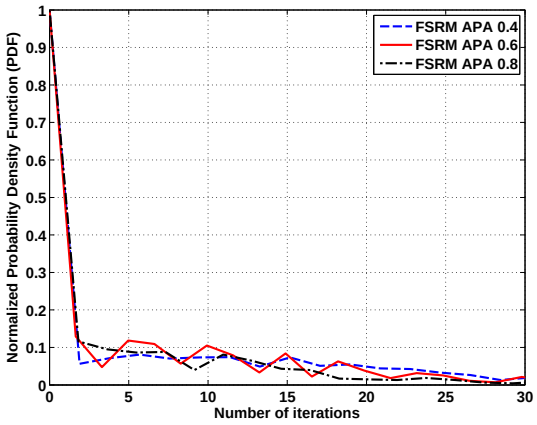
3.5.2. Convergence Analysis of the Fairness/Rate Adaptive Techniques

As said in section 3.4, the fairness/rate adaptive techniques proposed in this thesis are composed by three RRA algorithms: initial DSA based on a classic rate adaptive policy, fairness-based DSA and fairness-based APA. The last two are iterative algorithms that perform sub-carrier re-assignments and power re-allocations in order to decrease or increase the system fairness in accordance with the desired CFT. In order to evaluate the convergence of the fairness/rate adaptive strategies, we present Fig. 3.7. This figure shows the normalized Probability Density Function (PDF) of the number of iterations required by the fairness-based DSA and APA algorithms in order to converge the instantaneous CFI Φ_{cell} to the CFT $\Phi_{\text{cell}}^{\text{target}}$. Each fairness/rate adaptive technique is evaluated

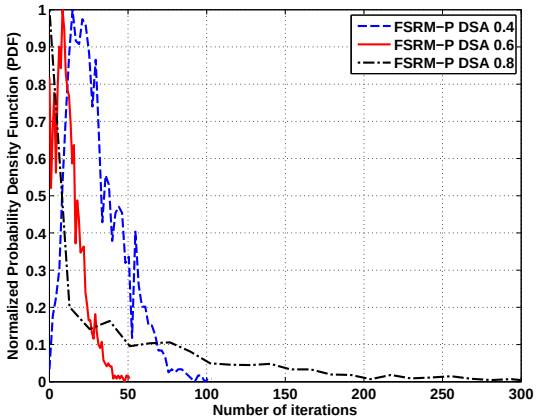
with several CFTs.



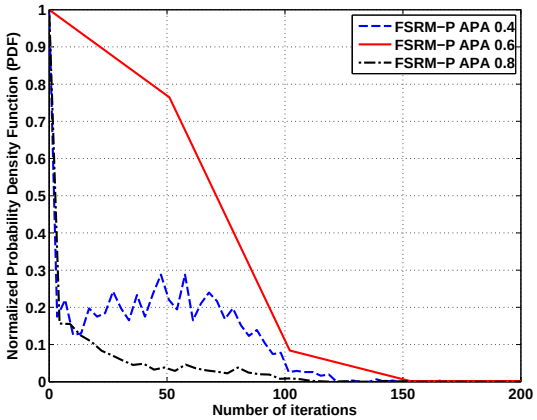
(a) FSRM DSA (fairness increase)



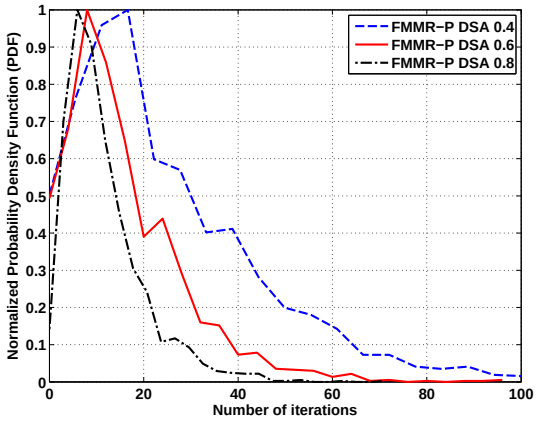
(b) FSRM APA (fairness decrease)



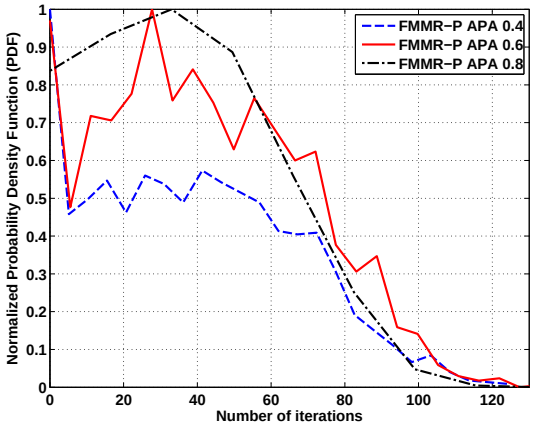
(c) FSRM-P DSA (fairness decrease, except CFT=0.8)



(d) FSRM-P APA (fairness increase, except CFT=0.8)



(e) FMMR-P DSA (fairness decrease)



(f) FMMR-P APA (fairness increase)

Figure 3.7.: Normalized Probability Density Function (PDF) of the number of required iterations for the convergence of the fairness/rate adaptive techniques

In order to better understand Fig. 3.7, it is important to take into account the fairness adaptation schemes corresponding to each of the policies, as summarized in Fig. 3.5 and Table 3.1. For the CFT values considered in this analysis, $\Phi_{\text{cell}}^{\text{target}} = [0.4, 0.6, 0.8]$, we have in general that FSRM applies a fairness increase (DSA) followed by a fairness decrease (APA), and that the FMMR-P applies a fairness decrease (DSA) followed by a fairness increase (APA). Regarding FSRM-P, since the initial sub-carrier assignment provides a CFI around 0.75, we have that the fairness adaptation scheme will be similar to FSRM for $\Phi_{\text{cell}}^{\text{target}} = 0.8$, and similar to FMMR-P for $\Phi_{\text{cell}}^{\text{target}} = [0.4, 0.6]$.

Firstly, let us analyze the convergence of the fairness-based DSA algorithms (Figs. 3.7(a), 3.7(c) and 3.7(e)). As expected, the largest the difference between the initial CFI Φ_{cell} and the CFT $\Phi_{\text{cell}}^{\text{target}}$, the higher the number of required iterations. This fact can be exemplified by Fig. 3.7(a), where we have that the initial CFI $\Phi_{\text{cell}}^{\text{SRM}}$ given by the classic SRM DSA algorithm is close to the minimum value $1/J$, where J is the number of users. Notice that the number of iterations required to achieve $\Phi_{\text{cell}}^{\text{target}} = 0.8$ is higher than the cases of $\Phi_{\text{cell}}^{\text{target}} = 0.6$ or 0.4. Another aspect is that it is easier for the algorithm to converge when fairness has to be decreased. This can be noticed comparing Figs. 3.7(a) and 3.7(e), for example. For the same CFTs, the FMMR-P DSA algorithm converges faster than the FSRM DSA algorithm. This fact is also observed with the FSRM-P DSA algorithm (see Fig. 3.7(c)). More iterations are required to converge to $\Phi_{\text{cell}}^{\text{target}} = 0.8$ (fairness increase) than the cases of $\Phi_{\text{cell}}^{\text{target}} = 0.6$ or 0.4 (fairness decrease).

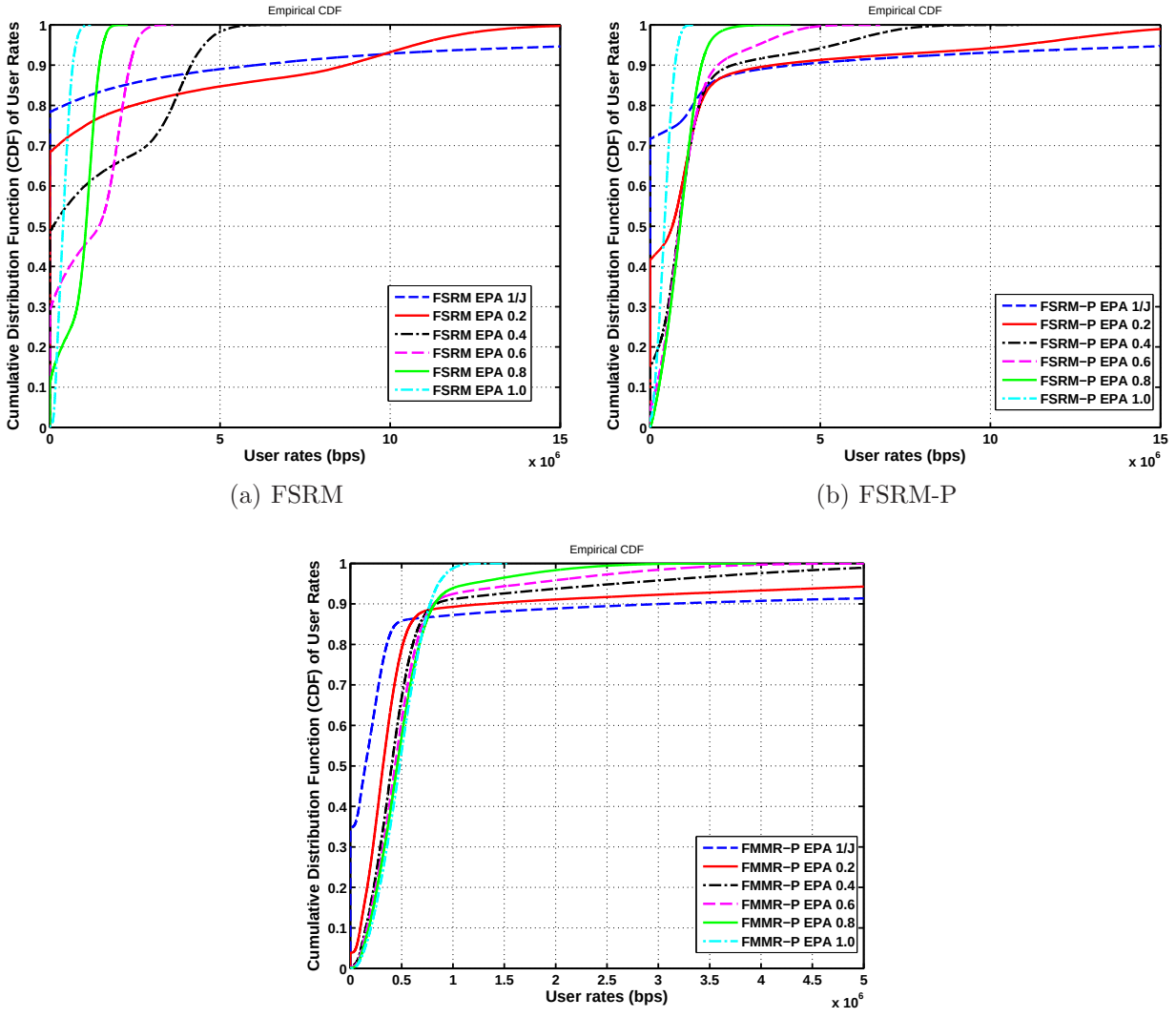
Regarding the convergence of the fairness-based APA algorithms, we see again that the convergence is slower when fairness has to be increased. This can be noticed by comparing Figs. 3.7(b) (fairness decrease) and 3.7(f) (fairness increase), for example. Furthermore, we can see that differently from the DSA case, the difference between CFI and CFT is not the main issue, since we already have a rough approximation of the CFT at the beginning of the APA algorithm. It seems that the main aspect is how high is the CFT when fairness has to be increased. Notice in Fig. 3.7(f) that the higher $\Phi_{\text{cell}}^{\text{target}}$, the more iterations are needed for the convergence of the APA algorithm.

When the initial CFI and the CFT are in opposite extremes of the efficiency-fairness plane, we have the worst case for the convergence of the DSA algorithm. In this case, it is wise to define a limit for the number of iterations of the algorithm. In the simulations performed for this thesis, we considered a maximum number of 2500 iterations. The worst-case scenario for the APA algorithm occurs when the CFT is located in any of the extremes of the efficiency-fairness plane. Again, a maximum of 2500 iterations was allowed in the simulations.

3.5.3. Rate Distribution and Rate Proportionality Analyses

In Fig. 3.8, the rate distribution provided by the fairness/rate adaptive techniques using the EPA approach is assessed. We show the CDFs of the users' rates in a scenario with 16 NRT users considering several CFTs. As the CFT varies, different rate distribution patterns are achieved by the RRA strategies. In general, we can see that high CFTs are characterized by steep curves (high fairness), while low CFTs are characterized by long-tale curves with many starving users (low

fairness). Despite this general trend, it can be seen that each technique treats the users in different ways.



(a) FSRM

(b) FSRM-P

(c) FMMR-P

Figure 3.8.: Cumulative Distribution Function (CDF) of the transmission rates for the fairness/rate adaptive policies considering a scenario with 16 users

The difference in the rate distribution pattern is more remarkable between FSRM and FMMR-P. Assuming that both have to guarantee the same fairness level, the former tries to maximize capacity and do not worry about rate proportionality or the QoS of the worst users, while the latter maximizes the capacity of the users with bad channel conditions trying to keep the users' rates as proportional as possible. This behavior is reflected in Figs. 3.8(a) and 3.8(c). For the same CFTs, the maximum transmission rates achieved by FSRM are higher than FMMR-P. On the other hand, a higher percentage of users are able to transmit with the FMMR-P technique.

FSRM-P (Fig. 3.8(b)) is a kind of hybrid between FSRM and FMMR-P. It has the objective of maximizing capacity, just like FSRM, but at the same time it tries to maintain the users' rates as proportional as possible, as FMMR-P does. Assuming the same CFTs, the percentage of users with very high rates and the percentage of users without resources (starving users) is in-between the values presented by FSRM and FMMR-P.

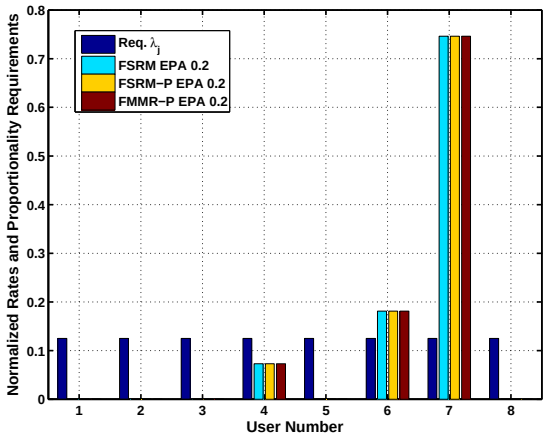
Now, let us analyze how the fairness/rate adaptive techniques behave in terms of rate proportionality considering several CFTs. Fig. 3.9 presents examples of bar plots of the rates of 8 users considering the EPA approach. In this figure, the proportional rate requirements λ are also presented (all requirements are equal). As expected, when a small CFT is considered (Fig. 3.9(a)), few users with best channel conditions (users 4, 6 and 7) are privileged while the others have no resources. In the other extreme (CFT equal to 1 in Fig. 3.9(e)), we notice that the users' rates are equalized and all RRA strategies are able to approximately meet the proportional rate requirements.

It is interesting to see how each technique re-assigns resources in order to adapt fairness to different target values. FSRM starts with a low initial CFI and has to increase fairness until the desired CFT is met. For a small CFT = 0.2 (Fig. 3.9(a)), the sub-carriers are concentrated in the hands of users 4, 6 and 7. Once higher CFTs are considered (from Fig. 3.9(b) to Fig. 3.9(e)), the resources are taken from these best users and re-assigned to other users. Notice that the users 3 and 5, which are the ones with worst channel conditions, are the last ones to be contemplated.

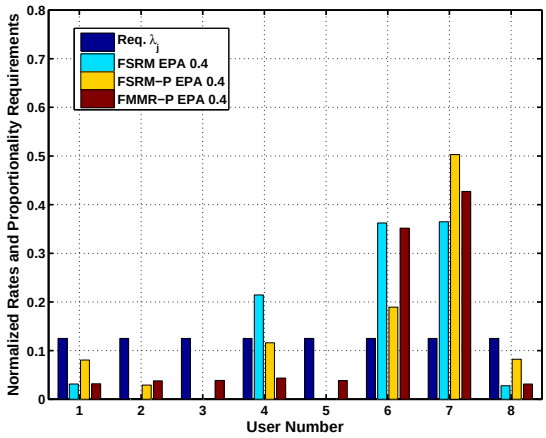
On the other hand, FSRM-P and FMMR-P start with a high initial CFI and have to decrease fairness in order to meet lower CFT values. In their fairness decrease procedure (from Fig. 3.9(e) to Fig. 3.9(a)), we can see their main difference: FMMR-P is able to protect as much as possible the minimum capacity among the users, while FSRM-P neglects some bad users (for example users 3 and 5) in favor of users with best channel conditions (users 4, 6 and 7) in order to increase the overall system capacity. However, notice that FSRM-P neglects fewer users than FSRM, since the former is aware about the proportional rate requirements and try to disregard them as little as possible.

3.5.4. Fairness Analysis

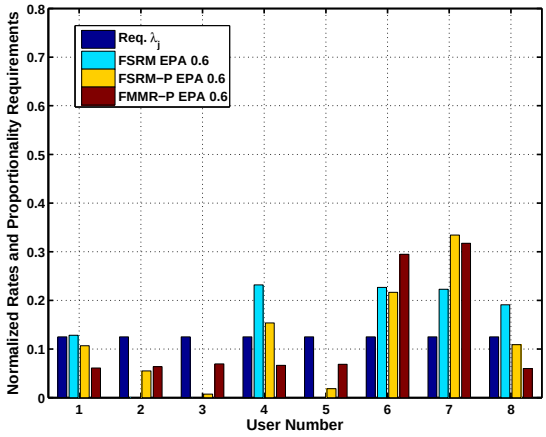
Fig. 3.10 depicts the mean CFI averaged over all snapshots as a function of the number of users for all classic rate adaptive and fairness/rate adaptive techniques considered in this thesis. In Figs. 3.10(a), 3.10(b) and 3.10(c), FSRM, FSRM-P and FMMR-P are compared in terms of fairness with the classic strategies, respectively. At this point we consider that the fairness/rate adaptive techniques use the EPA approach.



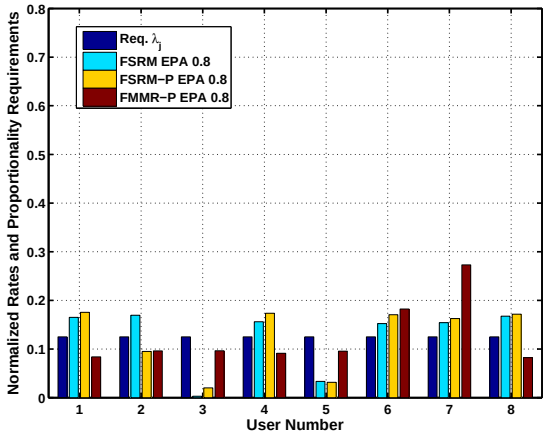
(a) CFT= 0.2



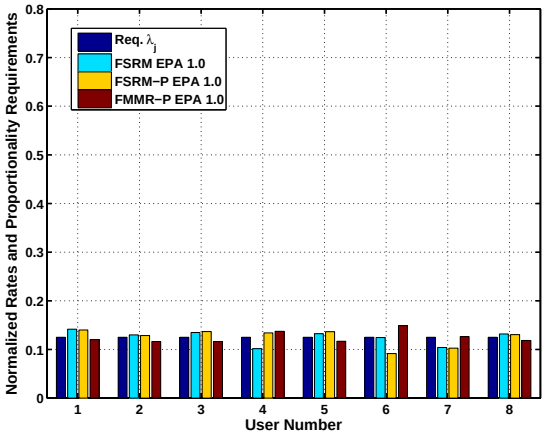
(b) CFT= 0.4



(c) CFT= 0.6



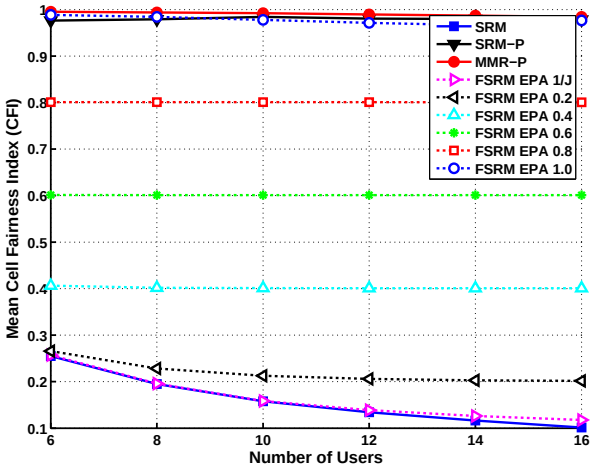
(d) CFT= 0.8



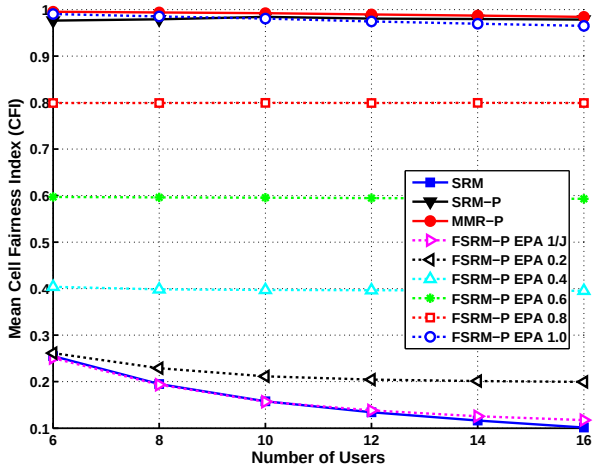
(e) CFT= 1.0

Figure 3.9.: Examples of bar plots of the transmission rates for the fairness/rate adaptive techniques considering a scenario with various cell fairness targets and 8 users

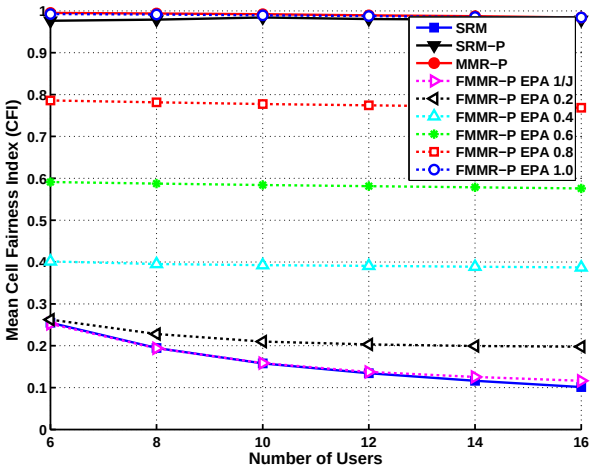
Regarding the classic rate adaptive policies, these graphics reinforce what was concluded in section 3.5.1: SRM presents the lowest values of CFI, while SRM-P and MMR-P are able to distribute the resources among the users in a fair manner. Notice that the higher the number of users, the lower the fairness provided by SRM. This is due to the multi-user diversity which is fully exploited by the opportunistic resource allocation of the SRM technique.



(a) FSRM



(b) FSRM-P



(c) FMMR-P

Figure 3.10.: Mean cell fairness index as a function of the number of users for the classic (solid lines) and fairness/rate adaptive techniques (dashed lines)

Considering the fairness/rate adaptive techniques, all of them are successful at guaranteeing the fairness targets. Notice that for lower system loads, the RRA strategies are not able to exactly meet very low CFTs. This happens due to two interrelated factors: 1) the multi-user diversity is not sufficient with a low number of users; and 2) the performance of the fairness/rate adaptive techniques are lower-bounded by the classic SRM policy. As explained in section 3.4, the proposed

fairness-based policies were meant to provide a controllable trade-off between resource efficiency and user fairness. In the fairness adaptation procedures, the two conflicting objectives that generate the trade-off rule the resource re-allocation. By one side, when fairness must be decreased, we re-allocate resources as efficiently as possible. In the extreme case of CFT = $1/J$, all sub-carriers will be re-allocated to their best owners. Notice that this extreme case exactly corresponds to the way that the classic opportunistic SRM assigns the sub-carriers. Therefore, the performance of the fairness/rate adaptive techniques configured with very low CFTs converges to the performance of the classic SRM. A similar behavior happens for extremely high values of CFT. The performance of the fairness-based RRA strategies converges to the performance of the classic MMR-P, which is the one that presents the highest values of CFI.

The fairness performance of the joint and EPA approaches for all fairness/rate adaptive techniques is compared in Table 3.3. The CFT = 0.6 is used as an example. It is observed that the fairness-based APA algorithm, which is used in the joint approach, is able to meet the fairness target with very high accuracy. It can also be seen that the fairness-based DSA algorithm alone is also able to provide very good fairness control, with values of CFI very close to the CFT. FMMR-P using EPA provides the most inaccurate results, but even in this case the difference is not significant. Similar results were obtained for the other CFTs.

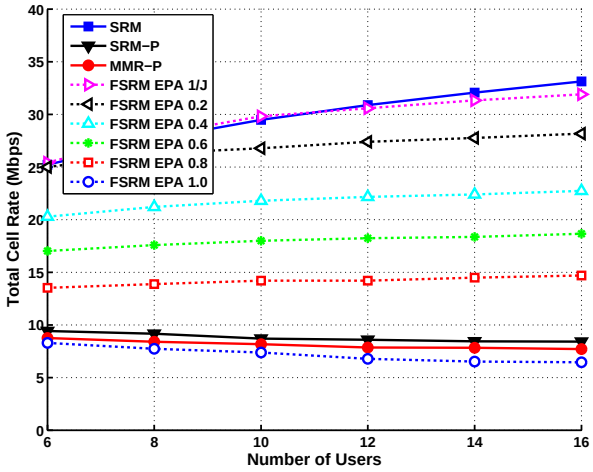
Table 3.3.: Mean cell fairness index as a function of the number of users for the fairness/rate adaptive techniques

RRA Policies	Number of users					
	J=6	J=8	J=10	J=12	J=14	J=16
FSRM EPA 0.6	0.602	0.601	0.601	0.601	0.601	0.601
FSRM Joint 0.6	0.601	0.600	0.600	0.600	0.600	0.600
FSRM-P EPA 0.6	0.597	0.596	0.596	0.595	0.594	0.593
FSRM-P Joint 0.6	0.599	0.599	0.599	0.599	0.599	0.599
FMMR-P EPA 0.6	0.591	0.587	0.584	0.581	0.578	0.576
FMMR-P Joint 0.6	0.600	0.600	0.600	0.600	0.600	0.600

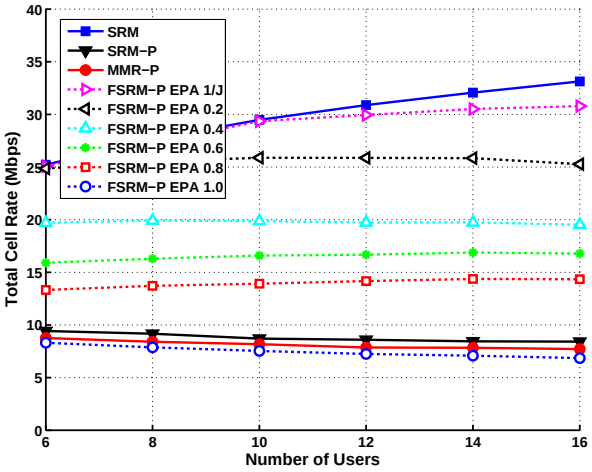
3.5.5. Efficiency Analysis

As a consequence of the trade-off, we have that the total cell throughput, which is the efficiency indicator that we use in this analysis, is inversely proportional to the CFT. The higher the CFT, the lower the total cell throughput, as can be seen in Fig. 3.11. Each graphic compares the performance of one of the proposed fairness/rate adaptive strategies with the classic rate adaptive techniques. The efficiency performance of the classic strategies confirms our expectations: SRM provides much better results in terms of system capacity than SRM-P and MMR-P. SRM-P also shows slightly better results than MMR-P due to its DSA algorithm that seeks the maximization of the capacity whenever possible.

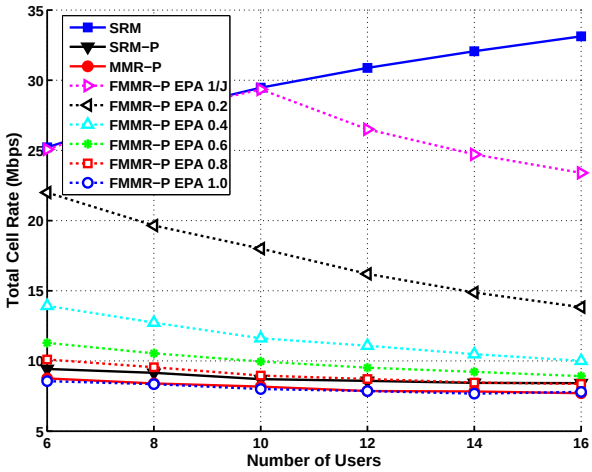
Regarding the fairness/rate adaptive techniques, one can see the inverse proportion between ca-



(a) FSRM



(b) FSRM-P



(c) FMMR-P

Figure 3.11.: Total cell rate as a function of the number of users for the classic (solid lines) and fairness/rate adaptive techniques (dashed lines)

pacity and CFT. Furthermore, comparing Figs. 3.11(a) and 3.11(b), it can be observed that FSRM provides higher capacity than FSRM-P for the same CFTs. This difference is more remarkable for high system loads, where FSRM can take more advantage from the multi-user diversity. Since this diversity is not exploited by RRA strategies that have to guarantee high degrees of fairness, the system capacity provided by such techniques is almost constant with the number of users, as can be seen in Fig. 3.11.

It is interesting to notice in Fig. 3.11(c) that FMMR-P provides the lowest system capacity among the fairness/rate adaptive techniques. Since it has to protect the QoS of the worst users, the capacity gain when lower CFTs are considered is very low. Considering the minimum CFT = $1/J$, the FMMR-P DSA algorithm is able to achieve the same performance of the classic SRM

technique only for a low number of users. On one hand, this is possible because the sub-carrier re-assignment procedure is simpler when we have few users in the system. On the other hand, more users implies that there are more possible resource allocations that yield the same fairness index, and FMMR-P is not able to pick the most efficient allocation. Moreover, we have the fact that the initial DSA assignment of FMMR-P is far from the lowest values of CFT. Thus, more iterations are necessary for the convergence and the number of iterations required by FMMR-P to choose the most efficient resource allocation is usually higher than the maximum allowed.

We compare again the performance of RRA techniques with regard to the joint and EPA approaches in Table 3.4 for the case of CFT = 0.6. Now the comparison is in terms of the total cell throughput. One more time, we cannot see a significant difference in performance, neither in the case of CFT = 0.6, nor in the cases of the other CFTs considered in the simulations. It means that the small fairness adaptation performed by the APA algorithm has not an important impact on system capacity.

Table 3.4.: Total cell throughput in Mbps as a function of the number of users for the fairness/rate adaptive techniques

RRA Policies	**Number of users**					
	J=6	**J=8**	**J=10**	**J=12**	**J=14**	**J=16**
FSRM EPA 0.6	17.03	17.58	17.99	18.23	18.36	18.66
FSRM Joint 0.6	17.27	17.83	18.14	18.31	18.54	18.65
FSRM-P EPA 0.6	15.91	16.28	16.58	16.66	16.87	16.77
FSRM-P Joint 0.6	17.06	17.12	17.07	16.83	16.94	16.78
FMMR-P EPA 0.6	11.29	10.55	9.96	9.52	9.22	8.92
FMMR-P Joint 0.6	9.97	9.50	9.21	8.74	8.55	8.42

The best way to evaluate the trade-off between resource efficiency and user fairness is plotting the 2D efficiency-fairness plane. The chosen efficiency and fairness indicators are the total cell data rate (capacity) and cell fairness index, respectively. Fig. 3.12 summarizes the most relevant aspects discussed so far. It compares the performance of the classic rate adaptive techniques (SRM, MMR-P and SRM-P), which are indicated as single markers, and the generalized fairness/rate adaptive strategies (FSRM, FMMR-P and FSRM-P), which are indicated as solid and dashed lines. In order to plot the efficiency-fairness plane, the number of users must be fixed, which in this case is 16.

The classic rate adaptive techniques are represented as single points in the efficiency-fairness plane because they represent static policies, i.e. each policy provides only one trade-off operation point. SRM provides maximum capacity at the expense of very poor fairness among users, while MMR-P and SRM-P are very fair in the rate distribution (CFI close to one) but as a consequence they achieve much lower system capacity.

On the other hand, the fairness/rate adaptive techniques, which are able to achieve a desired cell fairness target thank to a new fairness constraint in their respective optimization problems,

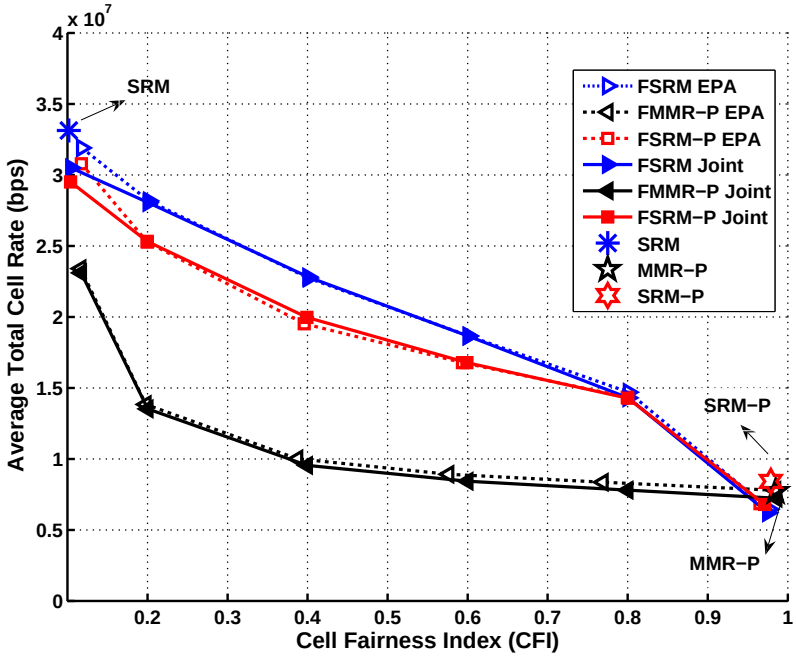


Figure 3.12.: Efficiency-Fairness plane for the classic and fairness/rate adaptive techniques

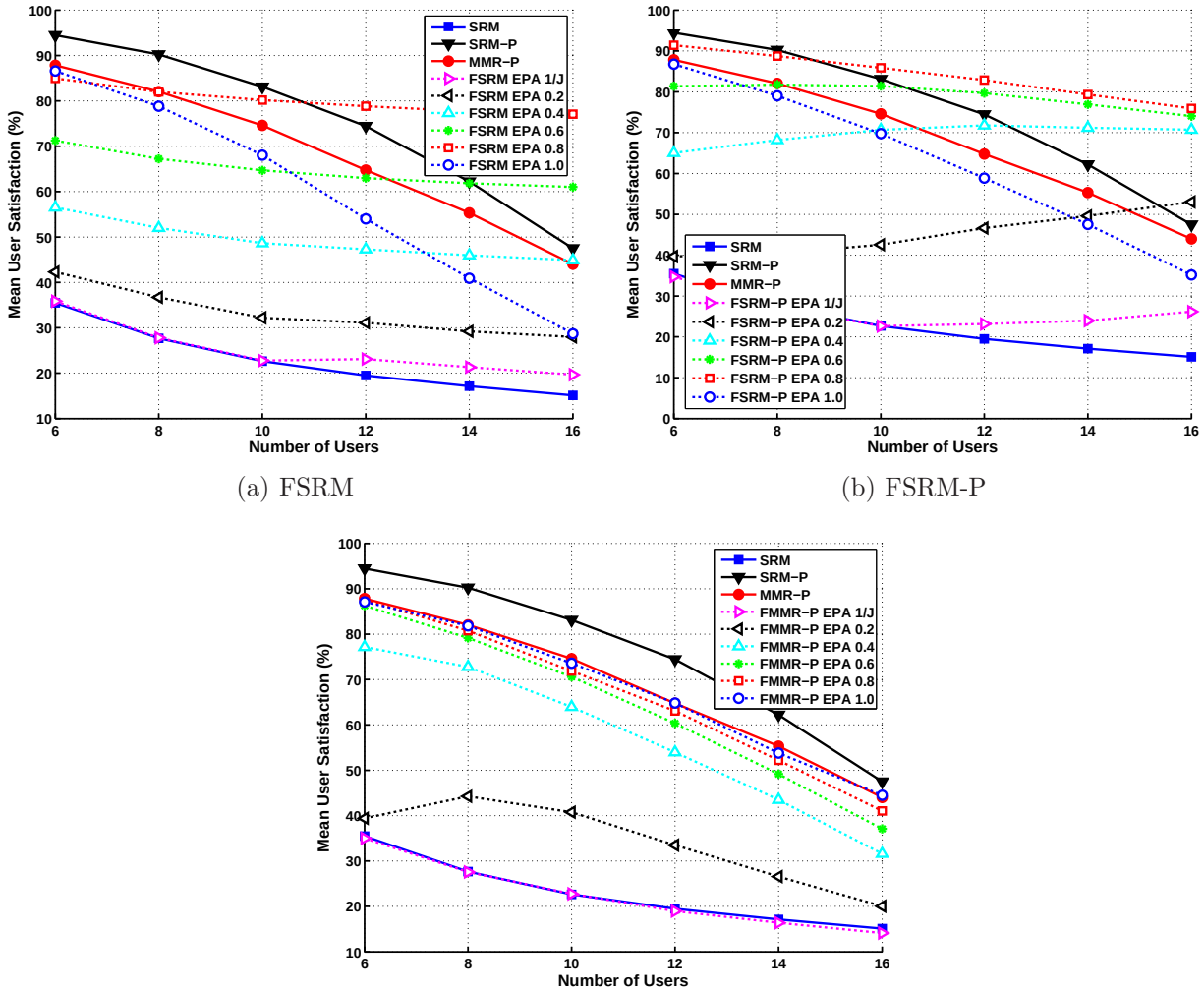
are able to cover the whole path between extreme points in the efficiency-fairness plane (classic rate adaptive points), drawing a complete efficiency-fairness curve. The fairness targets set in the simulations were $[1/J, 0.2, 0.4, 0.6, 0.8, 1.0]$, where J is total number of NRT users in the cell. One can observe that the performance of the proposed fairness/rate adaptive strategies converge to the results of the classic rate adaptive techniques in both extremes of the CFI, which are $1/J$ and 1. The FSRM variants (joint and EPA) show the best performance because they provide the largest efficiency-fairness regions, i.e. for the same CFIs, they present equal or higher system capacity. The FMMR-P variants present the smallest efficiency-fairness regions due to the fact that this RRA strategy protects the minimum user capacity as required by its objective function, which does not lead to an efficient resource usage. Although FMMR-P and FSRM-P techniques start at close initial points in the high fairness range of the efficiency-fairness plane, they follow very different paths towards the low fairness range. The higher capacity achieved by FSRM-P can be explained by the fact that it attempts to maximize the total cell rate no matter the QoS of the worst users (minimum user capacity).

Comparing the EPA and joint variants, one can see that the former achieves a rough approximation of the CFTs, while the latter is able to achieve a more exact approximation. In the fairness-based DSA algorithm, discrete sub-carriers are moved from one user to another in order to adapt fairness. However, in the fairness-based APA algorithm executed only in the joint approach, sufficiently small slices of power are re-allocated between sub-carriers in order to adapt the cell fairness index. Therefore, the power management allows a finer control of the fairness in the

system. Finally, it can be concluded from Fig. 3.12 that there is no significant difference in system capacity comparing the joint and EPA approaches.

3.5.6. Satisfaction Analysis

In this section, all RRA techniques are evaluated in terms of user satisfaction. This analysis is very important to see if the benefit of the control of the trade-off between resource efficiency and user fairness is not counterbalanced by a QoS degradation of the users. Fig. 3.13 depicts the mean user satisfaction as a function of the number of users for the classic rate adaptive strategies and the EPA variant of the fairness/rate adaptive techniques.



(a) FSRM

(b) FSRM-P

(c) FMMR-P

Figure 3.13.: Mean user satisfaction as a function of the number of users for the classic (solid lines) and fairness/rate adaptive techniques (dashed lines)

Assessing the classic policies, it can be observed that SRM, which maximizes capacity, achieves

low and almost constant user satisfaction for high system loads because it always gives priority to few users with best channel conditions. On the other hand, the techniques that privilege user fairness, namely SRM-P and MMR-P, present higher satisfaction for low system loads but the performance decreases fast when the number of users increases. SRM-P is the classic technique that shows the highest user satisfaction because it finds a good balance between prioritizing the best users to maximize capacity and avoiding the complete starvation of the worst users due to the proportional rate requirements.

Now, let us analyze the fairness/rate adaptive strategies. Looking at Figs. 3.13(a) and 3.13(b), which shows the results of FSRM and FSRM-P respectively, one can notice that as the CFT increases, the user satisfaction also increases until a point where further fairness increase yields a satisfaction decrease. This suggests that there is an optimum CFT where the user satisfaction is maximum. This is the case of the FSRM-P technique configured with CFT = 0.8, which provides the highest satisfaction for almost the whole range of system loads considered in the simulations (see Fig. 3.13(b)). Regarding FMMR-P, its satisfaction performance is constrained within the range delimited by the classic techniques SRM (lower bound) and MMR-P (upper bound). In this case, intermediate values of CFT are not translated into satisfaction results better than the classic strategies MMR-P or SRM-P.

Table 3.5 shows the satisfaction results of the joint and EPA variants of the fairness-based techniques configured with CFT = 0.6. In general, the results are very similar. Only FMMR-P shows a noticeable difference, where the EPA approach provides slight better results than the joint approach.

Table 3.5.: Percentage of user satisfaction as a function of the number of users for the fairness/rate adaptive techniques

RRA Policies	Number of users					
	J=6	**J=8**	**J=10**	**J=12**	**J=14**	**J=16**
FSRM EPA 0.6	71.28	67.26	64.70	62.98	61.84	61.03
FSRM Joint 0.6	71.18	67.14	64.75	62.86	61.86	60.93
FSRM-P EPA 0.6	81.38	81.77	81.44	79.69	76.94	74.05
FSRM-P Joint 0.6	81.79	82.19	81.71	79.95	77.03	74.15
FMMR-P EPA 0.6	86.28	79.18	70.58	60.34	49.16	37.08
FMMR-P Joint 0.6	82.93	74.70	66.44	54.36	43.64	33.13

Fig. 3.14 depicts the user satisfaction as a function of the cell fairness index, forming a satisfaction-fairness plane. A generic user j is considered satisfied if his data rate is equal or higher than his satisfaction requirement. Since we considered a simulation scenario where all proportional rate requirements λ are equal, we have that all satisfaction requirements are given by the value depicted in Table 3.2 (512 kbps). Again, it can be seen that the performance of the fairness/rate adaptive strategies converge to the results of the classic rate adaptive techniques, which demonstrates that the former are successful generalizations of the latter.

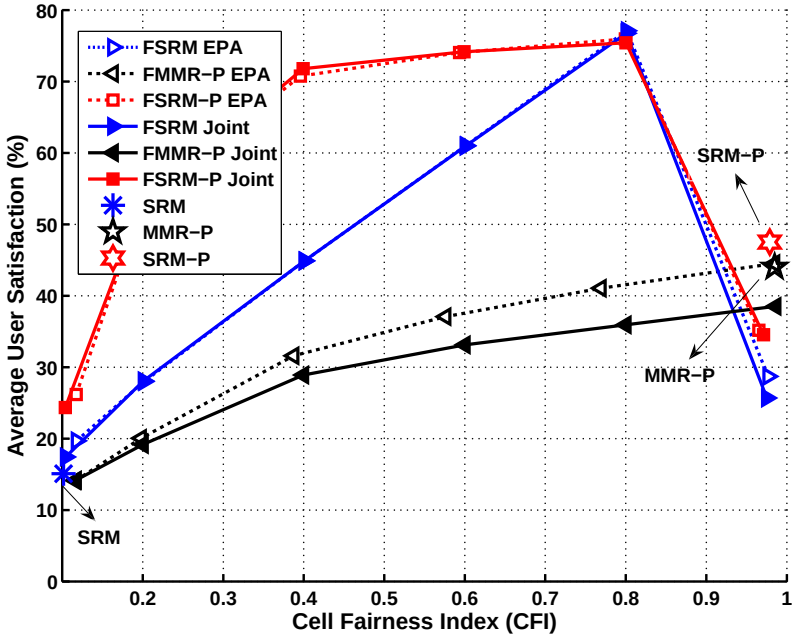


Figure 3.14.: Satisfaction-Fairness plane for the classic and fairness/rate adaptive techniques

In the low fairness range, all RRA techniques show low satisfaction levels, which can be expected because at these fairness levels the resources are shared among few users with good channel conditions. These lucky users are satisfied while the others are not. On the other hand, all RRA strategies present intermediate satisfaction levels in the high fairness range with the considered system load. When the fairness is high, the users have almost the same proportional rates, and depending on the value of the satisfaction requirement or the number of users competing for resources, the majority of the users as a group can be either satisfied or unsatisfied. If the network assumes a very large satisfaction requirement or the system is overloaded, the satisfaction levels of the RRA techniques will fall very drastically in the high fairness range.

It is interesting to see that intermediate fairness levels can provide higher satisfaction values, at least for FSRM-P and FSRM, whose maximum satisfaction was achieved for a CFI equal to 0.8. Also notice that FSRM-P surpasses all other techniques for almost the whole range of cell fairness indexes. Finally, it is important to highlight that the joint variants of all fairness-based strategies show satisfaction percentages approximately the same or lower than the EPA variants, which suggests that the action of exactly meeting the fairness targets is not advantageous in terms of user satisfaction.

3.5.7. CPU Time Analysis

The algorithmic complexity of the studied RRA techniques is represented by means of the CPU time needed to execute the respective algorithms. The CPU times were measured using the same

machine. Fig. 3.15 shows the average CPU time in a logarithmic scale against the cell fairness index, which characterizes a CPU time-fairness plane. Among the classic rate adaptive strategies, SRM is the fastest to run, while SRM-P and MMR-P shows similar complexity.

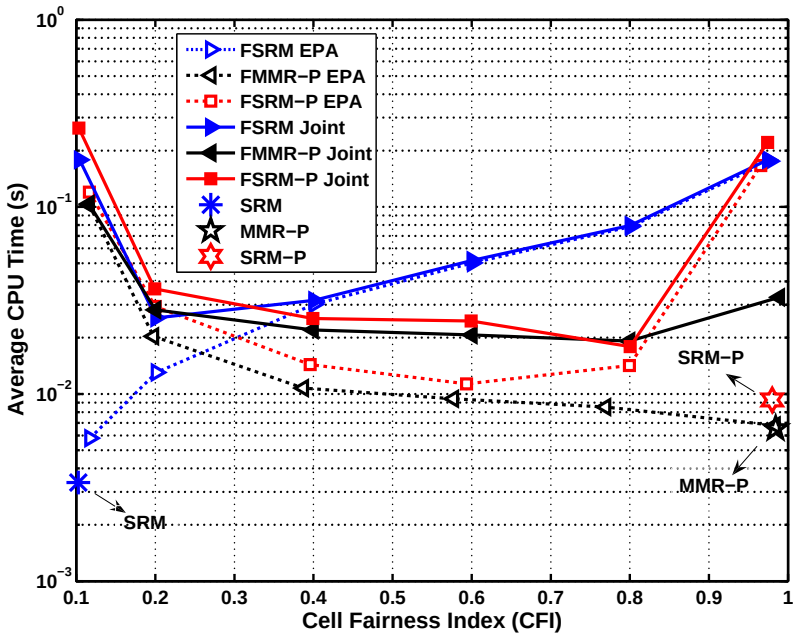


Figure 3.15.: CPU time-Fairness plane for the classic and fairness/rate adaptive techniques

Regarding the fairness-based techniques, one can clearly observe that the EPA approach takes less CPU time than the joint approach because the latter use an iterative power re-allocation procedure. The distance between the initial CFI given by the initial DSA assignment and the CFT has a significant impact on the computational burden. Considering only the EPA approach, CFTs close to the initial CFI present the minimum complexity: low fairness range for FSRM, high fairness range for FMMR-P and intermediate fairness range around 0.75 for FSRM-P. As a rule of thumb, we have that the required computational time of the techniques is very high when the CFT is located in one of the extremes of the CPU time-fairness plane, and this CFT is not close to the initial CFI.

Another aspect is that in the extreme CFTs, the fairness-based APA algorithms demand a lot of iterations to converge the CFI to the desired CFT with high accuracy.

Finally, as suggested by the convergence analysis presented in section 3.5.2, the action of decreasing the fairness requires less effort than increasing it. This is the reason why FMMR-P generally demands less computational resources than FSRM.

In summary, FSRM EPA is less computationally demanding for the low fairness range, while FMMR-P EPA requires less CPU time for intermediate and high fairness. FSRM-P EPA also has a complexity slightly higher than FMMR-P EPA for intermediate values of fairness. In their respective best fairness ranges, the EPA variant of the proposed fairness/rate adaptive techniques

present the required CPU time in the same order of magnitude of the classic rate adaptive strategies.

3.6. Conclusions

The fundamental problem of the trade-off between resource efficiency and user fairness in OFDMA networks using opportunistic or fairness-oriented radio resource allocation was addressed in this chapter. Three novel fairness/rate adaptive RRA policies are proposed, which consider in their respective optimization problems a new fairness constraint based on the concept of a cell fairness index. The policies are called Fairness-Based Sum Rate Maximization (FSRM), Fairness-Based Sum Rate Maximization with Proportional Rate Constraints (FSRM-P) and Fairness-Based Max-Min Rate with Proportional Rate Constraints (FMMR-P). They are generalizations of the classic rate adaptive Sum Rate Maximization (SRM), Sum Rate Maximization with Proportional Rate Constraints (SRM-P) and Max-Min Rate with Proportional Rate Constraints (MMR-P) policies, respectively. We also proposed heuristic-based RRA techniques that are able to implement the aforementioned policies in practice. They are comprised of three independent RRA algorithms: initial DSA, fairness-based DSA and fairness-based APA. Using these proposed fairness-based RRA techniques, the network operator is able to control the trade-off between efficiency and fairness.

The main conclusions taken from the performance evaluation of the RRA techniques are summarized as follows:

- Well-known classic rate adaptive policies found in the literature and studied in this thesis have fixed performance regarding the balance between resource efficiency and user fairness. The SRM technique, which uses a pure opportunistic policy that allocates the resources only to the best users, is the one that maximizes capacity but is the unfairest in the resource distribution. On the other hand, we have that the SRM-P and MMR-P techniques are able to guarantee the proportional rate requirements providing a high fairness, but this comes at the expense of a lower system capacity. This static trade-off provided by the classic strategies was the main motivation to propose generalized fairness/rate adaptive techniques that offer to the network operator the possibility to have a strict short-term control of the fairness in the system, and consequently manage the aforementioned trade-off according to its interests.

- The fairness-based techniques rely on two iterative RRA algorithms that perform sub-carrier re-assignments and power re-allocations in order to decrease or increase the system fairness in accordance with the desired CFT. We have seen that the largest the difference between the initial CFI Φ_{cell} and the CFT $\Phi_{\text{cell}}^{\text{target}}$, the higher the number of required iterations. We observed that it is easier for the algorithms to converge when fairness has to be decreased. This could be explained by the fact that all fairness/rate adaptive strategies share the common characteristic that the resource re-allocation is always directed to the users who could take the most benefit from the usage of the new resource. This aspect generally benefits the users with good channel conditions and can accelerate the fairness decrease process.

- Analyzing the rate distribution and rate proportionality, one can clearly see the different guidelines of the fairness/rate adaptive policies. FSRM tries to maximize capacity and do not worry about rate proportionality or the QoS of the worst users. FMMR-P maximizes the capacity of the users with bad channel conditions trying to keep the users' rates as proportional as possible. The FSRM-P technique is a kind of hybrid between FSRM and FMMR-P. It has the objective of maximizing capacity, just like FSRM, but at the same time it tries to maintain the users' rates as proportional as possible, as FMMR-P does.

- All fairness/rate adaptive techniques are successful at guaranteeing the fairness targets. Their performances are constrained by the performance of the classic SRM and MMR-P techniques. The reason is because our proposed algorithms play with the competition of two paradigms: efficiency-oriented (resource-centric) and fairness-oriented (user-centric). SRM is the maximum exponent of the former paradigm, while MMR-P is the best representant of the latter.

- The fairness/rate adaptive strategies, which are able to achieve a desired cell fairness target thank to a new fairness constraint in their respective optimization problems, are able to cover the whole path between the extreme points in the efficiency-fairness plane, drawing a complete efficiency-fairness curve. One can observe that the performance of the proposed fairness-based techniques converge to the results of the classic rate adaptive strategies in both extremes of the CFI, which are $1/J$ and 1. The FSRM variants (joint and EPA) show the best performance because they provide the largest efficiency-fairness regions, i.e. for the same CFIs, they present equal or higher system capacity. FSRM-P shows a performance slightly worse than FSRM. The FMMR-P variants present the smallest efficiency-fairness regions due to the fact that this policy protects the minimum user capacity as required by its objective function, which does not lead to an efficient resource usage.

- Analyzing the satisfaction-fairness plane, one can see that intermediate fairness levels can provide higher satisfaction values, at least for the FSRM-P and FSRM policies, whose maximum satisfaction was achieved for a CFI equal to 0.8. It can also be noticed that FSRM-P surpasses all other techniques for almost the whole range of cell fairness indexes.

- Two approaches of the fairness/rate adaptive techniques were evaluated in the simulations: EPA and joint. The former, where only the fairness-based DSA algorithm is used to adapt fairness, achieves rough approximation of the CFT. The joint approach, where both DSA and APA algorithms are used to adapt fairness, always provides very accurate results regarding the fairness convergence. It was seen in the simulation results that the rough fairness approximation of the EPA approach is sufficiently accurate for the cases of FSRM and FSRM-P. Furthermore, observing the efficiency-fairness and satisfaction-fairness planes, we can see that there is not significant difference in performance between both approaches, except for the user satisfaction provided by the FMMR-P technique, where the EPA approach shows even better

results than the joint. Since the joint approach is more computationally demanding than EPA, we can conclude that in order to achieve a controllable trade-off between resource efficiency and user fairness, it is sufficient to consider a DSA algorithm able to adapt the fairness and next apply equal power allocation among the sub-carriers. Therefore, we have that adherence to the proportionality constraints does not need to be strictly enforced in practical systems. A soft fairness guarantee is acceptable as long as the capacity and/or satisfaction are maximized and the algorithm complexity is low.

- FSRM EPA is less computationally demanding for the low fairness range, while FMMR-P EPA requires less CPU time for intermediate and high fairness. FSRM-P EPA also has a complexity slightly higher than FMMR-P EPA for intermediate fairness values.

- Taking into account important network metrics, such as system capacity, user fairness, user satisfaction and algorithm complexity, the following methodology can be suggested to the network operator: use FSRM for the low fairness range ($1/J \leq \Phi_{\text{cell}} \leq 0.2$); FSRM-P for the intermediate fairness range ($0.2 < \Phi_{\text{cell}} \leq 0.8$); and FMMR-P for the high fairness range ($0.8 < \Phi_{\text{cell}} \leq 1$); where J is the number of users in the cell. If the operator preferred only one proposed fairness/rate adaptive technique to cover the whole fairness index range, we would suggest FSRM-P, since it provides very accurate fairness guarantee with high user satisfaction and intermediate system capacity and computational complexity.

Chapter 4

Utility-Based Resource Allocation for Macrocell Networks

4.1. Introduction

In this chapter, the problem of the management of the trade-off between resource efficiency and user fairness in the downlink of a cellular system is revisited. This problem was introduced in section 2.8 of chapter 2 and evaluated in chapter 3 considering a macrocell network with Non-Real Time (NRT) services and using rate adaptive resource allocation based on instantaneous data rates. Now, this Radio Resource Management (RRM) problem for NRT services is addressed using a different approach: resource allocation based on Utility Theory and average data rates (throughput). Furthermore, in this chapter we extend the study and propose a utility-based resource allocation based on delay, which is suitable for a scenario with Real Time (RT) services.

The consideration of average data rates instead of instantaneous rates in the formulation of Radio Resource Allocation (RRA) problems has some advantages. The time window used for rate averaging relax the constraint of fairness requirements. Rather than guaranteing the fairness instantaneously at each time slot, the use of a time window allows a smoother and more flexible fairness control.

Utility Theory is a powerful tool that can be used to design RRA algorithms able to achieve different levels of fairness in the resource allocation process [66]. This theory can be used in communication networks to quantify the benefit of usage of certain resources, e.g. bandwidth, power; or evaluate the degree to which a network satisfies service requirement of users' applications, for example in terms of throughput and delay. In particular, this thesis considers the flexibility of utility functions to propose adaptive RRA frameworks able to provide strict fairness control for both NRT and RT services. In this way, the network operator can have a good idea of the corresponding system capacity for every fairness level and is able to balance resource efficiency and user fairness in a controllable way. The concept of fairness assumed in this work is based on Quality

of Service (QoS) parameters. Throughput and delay were chosen as the fairness indicators in the scenarios with NRT and RT services, respectively.

The chapter is organized as follows. Section 4.2 presents the state-of-the-art on the management of the trade-off between resource efficiency and user fairness using Utility Theory and highlights the novel contributions of this thesis regarding this subject. Utility-based resource allocation algorithms suitable for NRT and RT services are described in section 4.3, while section 4.4 shows particular parametric RRA frameworks that can be designed to work as well-known classic RRA policies or dynamically adjusted according to the network operator's objectives. The impact of these utility-based RRA frameworks on the considered trade-off is evaluated by means of extensive system-level simulations in section 4.5. Finally, the conclusions are summarized in section 4.6.

4.2. Management of the Trade-Off between Resource Efficiency and User Fairness Using Utility Theory

Among the works that have proposed RRA algorithms to cope with this trade-off in a NRT scenario, three main approaches can be highlighted: optimization-based rate adaptive resource allocation [94–97], cross-layer Packet Scheduling (PSC) [98–105] and utility theory-based resource allocation [106–114].

In the case of the rate adaptive-based papers, the notion of fairness criteria was determined by rate proportionality constraints [94, 95], maximization of the sum of the logarithm of user rates [96], or both of them [97]. The logarithm function was used in [96, 97] because it was proved in [106] that this function is intimately associated with the concept of proportional fairness. This RRA approach was analyzed in details in chapter 3.

Most of the works that proposed PSC algorithms to effect a compromise between efficiency and fairness among NRT flows, for example [98–101], are based on the Proportional Fairness (PF) PSC algorithm proposed in [115] for High Data Rate (HDR) Code Division Multiple Access (CDMA) systems. However, there are some works like [102, 103] that used different approaches. The former introduced a PSC algorithm with a fairness controlling parameter that accounts for any intermediate policy between the instantaneous throughput fairness and the opportunistic policies, while the latter evaluated a scheduling algorithm whose priority function is a linear combination between instantaneous channel capacity and the average throughput. As a generalization of the PF criterion, we can highlight the weighted α-proportional fairness PSC algorithm, which is also known as the alpha-rule, that was initially proposed by [104] and later used in [105]. The idea behind this algorithm is to embody a number of fairness concepts, such as rate maximization, proportional fairness and max-min fairness, by varying the values of the parameter α and the weight parameter.

A more general class of RRA algorithms is based on utility fairness. Utility fairness is defined with a utility function that composes the optimization problem, where the objective is to find a feasible resource allocation that maximizes the utility function specific to the fairness concept

used. Some examples of utility functions can be found in [106–108]. There is a general family of utility functions that were presented and/or evaluated in [109–111] that includes the weighted α-proportional fairness algorithm as a special case. Some works followed a similar approach, but using different utility functions, e.g. [112–114].

The aforementioned trade-off is much less studied in a scenario with RT services. To the best of our knowledge, [116] was the only work that investigated in more details the trade-off between efficiency and delay-based fairness. The authors concluded that channel-aware opportunistic schedulers cause big rate and delay variabilities, which can lead to unfair situations frequently. On the other hand, other works studied the compromise between efficiency and QoS guarantee (delay bounds). These works can be classified in two main approaches: cross-layer PSC [117–121] and utility theory [122–124].

The opportunistic PSC algorithms suitable for RT services found in the literature have priority functions that always use an efficiency indicator, such as the instantaneous transmission rate (rate maximization policy) or the ratio between the instantaneous transmission rate and throughput (proportional fairness policy), and a QoS indicator based on delay. In this way, these PSC algorithms implicitly add to the problem some kind of delay-based fairness, because the flows that presented higher delays would have more priority to use the resources. Reference [118] went further and proposed a delay-based indicator that used the exponential function to equalize the weighted delays of all the queues when their differences were large. This policy was called the exponential rule and was proved to be throughput-optimal, i.e., it makes the queues stable if it is feasible to do so with any other scheduling rule. The algorithm proposed by [117], which was called Modified Largest Weighted Delay First (M-LWDF), was also proved to be throughput-optimal.

The utility-based PSC algorithms adopted a similar but more general procedure. The difference is that the QoS indicator used in the priority functions is now a marginal utility function based on delay. For example, [122] and [123] used z-shaped utility functions while [124] used particularly designed utility functions suitable to the services investigated on the paper. Notice that the utility-based approach is more general than classic PSC priority functions because the utility functions can be freely designed to provide the desired trade-off between efficiency and delay-based performance.

From the explained before, it can be concluded that utility theory is a flexible and powerful tool that can be used to develop RRA algorithms that are able to provide a controllable trade-off between efficiency in the resource usage and QoS-based fairness among NRT or RT flows. Based on that, the present work proposes two utility-based RRA frameworks suitable for NRT and RT services, respectively, that can balance efficiency and fairness in wireless systems according to the network operator's interest.

In the following, the most relevant and novel contributions of the thesis regarding this subject are presented:

1. ***RRA framework for NRT services***: the utility fairness concept is used to propose a general RRA policy suitable for NRT services. This RRA policy can be formulated as a

throughput-based utility optimization problem, which can be solved by a Dynamic Subcarrier Assignment (DSA) and Adaptive Power Allocation (APA) algorithms (see section 4.3.1). Previous works have used the utility fairness in a NRT scenario to propose only PSC algorithms [109–111] (alpha-rule). However, this thesis generalizes the use of the alpha-rule to also propose novel utility-based APA algorithms based on multi-level waterfilling. Based on this general RRA policy, a parametric RRA framework is proposed, which we call utility-based alpha-rule. Adjusting only one parameter in its parametric structure, this RRA framework can be designed to work as any of well known classic RRA policies, such as Max-Rate (MR) [61, 112], PF [106], Max-Min Fairness (MMF) [98], or any hybrid among them (see section 4.4.1). This is a powerful tool for the cellular operators, who have the possibility of dynamically choosing which RRA strategy is more convenient for their interests at a given instant.

2. ***Management of the trade-off between efficiency and fairness for NRT services***: some previous works that dealt with NRT services proposed and evaluated parametric solutions that can provide different levels of compromise between resource efficiency and fairness by varying a controlling parameter [108, 109, 114]. However, they only evaluated static trade-offs, i.e. the controlling parameter was not adapted during the network operation. This thesis goes beyond and presents a novel criterion to adapt the fairness controlling parameter α of the utility-based alpha-rule framework: a feedback control loop that meets a system fairness target. This is accomplished by the novel Adaptive Throughput-Based Fairness (ATF) policy (see section 4.4.1). The system throughput-based fairness is calculated using a fairness index based on the general fairness function proposed by [86]. This new idea states that the trade-off between efficiency and fairness can be managed by adaptively controlling a system fairness index. That is, the network operator sets a system fairness target and the ATF policy dynamically adapts the utility function of the alpha-rule framework in order to operate at the desired trade-off point. In this way, a mid-term (average) fairness control can be achieved.

3. ***RRA framework for RT services***: as far as we are concerned, this is the first work to use the utility fairness concept to propose a general RRA policy suitable for RT services. This RRA policy can be formulated as a delay-based utility optimization problem, which can be solved by a DSA and APA algorithms (see section 4.3.2). Based on this utility-based RRA policy, we propose a parametric RRA framework called utility-based beta-rule that can be designed to work as any of well-known classic RRA policies by adjusting only the fairness controlling parameter β in its parametric structure (see section 4.4.2). In this way, it comprises the PF [106], M-LWDF [117], First-In-First-Out (FIFO) [118], or any hybrid among these policies. This flexible RRA framework can be dynamically configured depending on the network conditions and the network operator's objectives.

4. ***Management of the trade-off between efficiency and fairness for RT services***: to

the best of our knowledge, among the works that studied this trade-off for RT services, only [116] proposed a parametric PSC algorithm that could balance in different ways the efficiency and delay-based fairness. However, the authors in [116] only assessed static trade-offs, i.e. evaluated the algorithm for specific values of the controlling parameter. Not only does the present work propose a new parametric framework (utility-based beta-rule) but it introduces a novel criterion to manage the trade-off by controlling the system delay-based fairness index as well. This mid-term (average) fairness control is performed by the novel Adaptive Delay-Based Fairness (ADF) policy (see section 4.4.2). The network operator has the simple task to decide which level of delay-based fairness it wants in the system. Given this fairness target, the proposed ADF policy will dynamically adapt the utility function used in the DSA and APA algorithms of the beta-rule framework in order to keep the system fairness around this planned value. This fairness operation point will directly define the capacity operation point. This is a simple and efficient way to manage the trade-off between efficiency in resource usage and delay-based fairness among users.

The concept of managing the trade-off by controlling the system fairness index is general and can be applied to any wireless system in which the users compete for centralized network resources. As a proof-of-concept we consider the Orthogonal Frequency Division Multiple Access (OFDMA) system, which is the chosen multiple access scheme for Fourth Generation (4G) broadband cellular networks, such as 3GPP Long Term Evolution-Advanced (LTE-Advanced) and IEEE Mobile Worldwide Interoperability for Microwave Access (WiMAX) Release 2.

4.3. Utility-Based Resource Allocation for OFDMA Systems

Utility Theory is a powerful tool that can be used to design RRA algorithms able to achieve different levels of fairness in the resource allocation process [106, 112].

The utility-based RRA frameworks for OFDMA systems designed in this work run in a distributed manner in each macrocell. The general utility-based optimization problem considered in this work is formulated below.

$$\max_{\mathcal{S}_j,\mathbf{p}} \sum_{j=1}^{J} U_j(\mathbf{x}) \tag{4.1}$$

$$\text{subject to } \bigcup_{j=1}^{J} \mathcal{S}_j \subseteq \mathcal{S}, \tag{4.2}$$

$$\mathcal{S}_i \bigcap \mathcal{S}_j = \varnothing, \quad i \neq j \quad \forall i,j = 1:J, \tag{4.3}$$

$$\sum_{k=1}^{K} p_k \leq P_{\text{total}}, \tag{4.4}$$

$$p_k \geq 0, \quad \forall k = 1:K, \tag{4.5}$$

where J is the total number of Mobile Terminals (MTs) in the macrocell, K is the total number

of sub-carriers in the system, $\mathcal{S}$ is the set of all sub-carriers in the system, $\mathcal{S}_j$ is the subset of sub-carriers assigned to the jth MT, $\mathbf{p}$ is the vector of powers for all sub-carriers, p_k is the power of the kth sub-carrier, P_{total} is total transmit power of the Macrocell Base Station (MBS), and $U_j(\mathbf{x})$ is a concave utility function based on a generic variable $\mathbf{x}$ that can represent a resource usage or QoS metric.

Constraints (4.2) and (4.3) state that the subsets of sub-carriers assigned to different users must be disjoint, i.e. the same sub-carrier cannot be shared by two or more users, and that the union of all these subsets must be contained in the total set of sub-carriers available in the system. On the other hand, constraints (4.4) and (4.5) require that the powers allocated to the sub-carriers are positive and the total sum of the powers over all sub-carriers must not surpass the total transmit power of the MBS.

Depending of the utility function used in (4.1), the optimum solution for the joint optimization problem (4.1)-(4.5) is very difficult to be found. The majority of the sub-optimum solutions proposed in the literature are based on the problem-splitting technique, which splits the problem in two stages: first, DSA with fixed power allocation, and next, APA with fixed sub-carrier assignment. In the present work, we also use this technique, as explained in the following.

Depending on the utility function and the variable $\mathbf{x}$, several RRA policies can be designed. In this study, we are interested at formulating general RRA policies suitable for NRT or RT services. In the former, we consider the variable $\mathbf{x}$ to be the users' throughput (average data rates), while in the latter the variable $\mathbf{x}$ represents the users' Head-Of-Line (HOL) delay. These two utility-based RRA policies are described in details in sections 4.3.1 and 4.3.2.

4.3.1. Non-Real Time Services

This section formulates the DSA and APA algorithms that use Utility Theory in order to find an efficient trade-off between system resource efficiency and throughput-based fairness among MTs that use NRT services. The considered optimization problem is the maximization of the total utility with respect to the throughput:

$$\max_{\mathcal{S}_j, \mathbf{p}} \sum_{j=1}^{J} U_j\left(T_j[n]\right). \tag{4.6}$$

$U_j(T_j[n])$ is a concave and increasing utility function based on the current throughput $T_j[n]$ of the jth MT. Notice that for this specific problem, the objective function (4.6) substitutes the general objective function (4.1). Constraints (4.2)-(4.5) remain valid for this problem.

The throughput (average data rate) of the jth MT is calculated using an exponential smoothing filtering, as indicated below.

$$T_j[n] = (1 - f_{\text{thru}}) \cdot T_j[n-1] + f_{\text{thru}} \cdot R_j[n], \tag{4.7}$$

where $R_j[n]$ is the instantaneous data rate of the jth MT calculated by (A.10) in appendix A and f_{thru} is a filtering constant used in the throughput averaging.

Evaluating the objective function in (4.6) and the throughput expression in (4.7), the derivative of $U_j(T_j)$ with respect to the transmission rate R_j is given by:

$$\frac{\partial U_j}{\partial R_j} = \frac{\partial U_j}{\partial T_j} \cdot \frac{\partial T_j}{\partial R_j} = f_{\text{thru}} \cdot \left.\frac{\partial U_j}{\partial T_j}\right|_{T_j=(1-f_{\text{thru}})\cdot T_j[n-1]+f_{\text{thru}}\cdot R_j[n]},$$

where f_{thru} is the aforementioned filtering constant in the throughput calculation. In the case that f_{thru} is sufficiently small, the expression above can be further simplified, as indicated below [113].

$$\frac{\partial U_j(T_j[n])}{\partial R_j[n]} \approx f_{\text{thru}} \cdot \left.\frac{\partial U_j}{\partial T_j}\right|_{T_j=T_j[n-1]},$$

where the previous resource allocation totally determines the current values of the marginal utilities. Using the one-order Taylor formula, the following expression can be derived [113]:

$$\sum_{j=1}^{J} U_j(T_j[n]) - \sum_{j=1}^{J} U_j(T_j[n-1]) \approx \sum_{j=1}^{J} \left.\frac{\partial U_j}{\partial T_j}\right|_{T_j=T_j[n-1]} \cdot (f_{\text{thru}} \cdot R_j[n] - f_{\text{thru}} \cdot T_j[n-1]). \tag{4.8}$$

Let us consider our simplified optimization problem given by (4.8). The maximization of (4.8) leads to the maximization of the original objective function (4.6). Since f_{thru} is a constant and $T_j[n-1]$ is fixed at the current Transmission Time Interval (TTI) n, the objective function of our simplified optimization problem becomes linear, as can be seen in the following.

$$\max_{\mathcal{S}_j,\mathbf{p}} \sum_{j=1}^{J} U_j'(T_j[n-1]) \cdot R_j[n] \tag{4.9}$$

where $U_j'(T_j[n-1]) = \left.\frac{\partial U_j}{\partial T_j}\right|_{T_j=T_j[n-1]}$ is the marginal utility of the jth MT with respect to its throughput in the previous TTI. The optimization problem (4.9) is a weighted sum rate maximization problem [125], whose weights are adaptively controlled by the marginal utilities. Notice that we started with an optimization formulation based on throughput (average data rate) given by (4.6), made some logical assumptions and mathematical simplifications, and ended up with a linear optimization formulation based on instantaneous rates given by (4.9). Therefore, we can affirm that the instantaneous optimization maximizing (4.9) leads to a long-term optimization that maximizes (4.6). For didactic reasons, we represent the marginal utility corresponding to the jth MT as a weight w_j^{nrt} to be later used on the RRA algorithms suitable for NRT services, as indicated

below:

$$w_j^{\mathrm{nrt}} = U_j^{'}\left(T_j\left[n-1\right]\right). \tag{4.10}$$

Dynamic Sub-carrier Assignment

The linear objective function greatly simplifies the corresponding algorithms. The DSA problem, which is the optimization problem (4.1)-(4.5) with equal power allocation, has a closed form solution when the objective function is given by (4.9) [113, 126]. The MT $m(k,n)$ is chosen to transmit on the kth sub-carrier in the nth TTI if it satisfies the condition given by (4.11):

$$m(k,n) = \arg\max_j \left\{w_j^{\mathrm{nrt}} \cdot c_{j,k}\left[n\right]\right\}, \tag{4.11}$$

where w_j^{nrt} is the utility-based weight suitable for NRT services calculated by (4.10) and $c_{j,k}\left[n\right]$ denotes the instantaneous achievable transmission efficiency of the kth sub-carrier with respect to the jth MT (Shannon capacity given by (A.9) on appendix A) assuming equal power allocation per sub-carrier. One can notice that in this first stage of the problem-splitting technique, the DSA sub-problem took into account the constraints (4.2) and (4.3).

Fig. 4.1 explains how the utility-based DSA algorithm proposed above works. Consider a scenario in which two NRT users i and j compete for 7 sub-carriers, where the former has better channel conditions than the latter. The channel qualities $\gamma_{i,k}^*$ and $\gamma_{j,k}^*$ plotted in the figure are utility-scaled versions of their original channel qualities $\gamma_{i,k}$ and $\gamma_{j,k}$, respectively, i.e. $\gamma_{i,k}^* = w_i^{\mathrm{nrt}} \cdot \gamma_{i,k}$ and $\gamma_{j,k}^* = w_j^{\mathrm{nrt}} \cdot \gamma_{j,k}$. According to (4.11), sub-carriers $k = 1,\ldots,3$ are assigned to user i and sub-carriers $k = 4,\ldots,7$ are assigned to user j. Notice that if the utility-based weights w_i^{nrt} and w_j^{nrt} were not used, all sub-carriers would have been assigned to user i, who originally had better channel conditions ($\gamma_{i,k} > \gamma_{j,k}$). Thus, the utility-based weights provided a fairer resource allocation.

Adaptive Power Allocation

Assuming that the DSA was already done, the optimal power allocation of the weighted sum rate maximization problem (4.9) has the solution in the form of a utility-based multi-level waterfilling [112, 125]:

$$p_k^*\left[n\right] = \left[\mu \cdot w_j^{\mathrm{nrt}} - \frac{\Gamma}{\gamma_{j,k}\left[n\right]}\right]^+, \quad \forall k \in \mathcal{S}_j \tag{4.12}$$

where $\left[x\right]^+ \triangleq \max\left(0,x\right)$, $p_k^*\left[n\right]$ is the current optimal power allocated to the kth sub-carrier belonging to the jth MT ($k \in \mathcal{S}_j$), $\Gamma/\gamma_{j,k}\left[n\right]$ is the inverse of the effective Channel-to-Noise Ratio (CNR), i.e. channel quality, of the kth sub-carrier assigned to the jth MT at the nth TTI, and μ is a non-negative variable that represents the water-level of the waterfilling problem. The constant Γ is the so-called Signal-to-Noise Ratio (SNR) gap, which indicates the difference between the theoretical limit and the SNR needed to achieve a certain data transmission rate for a practical system [112]. Notice that in this second stage of the problem-splitting technique, the APA sub-problem finds the

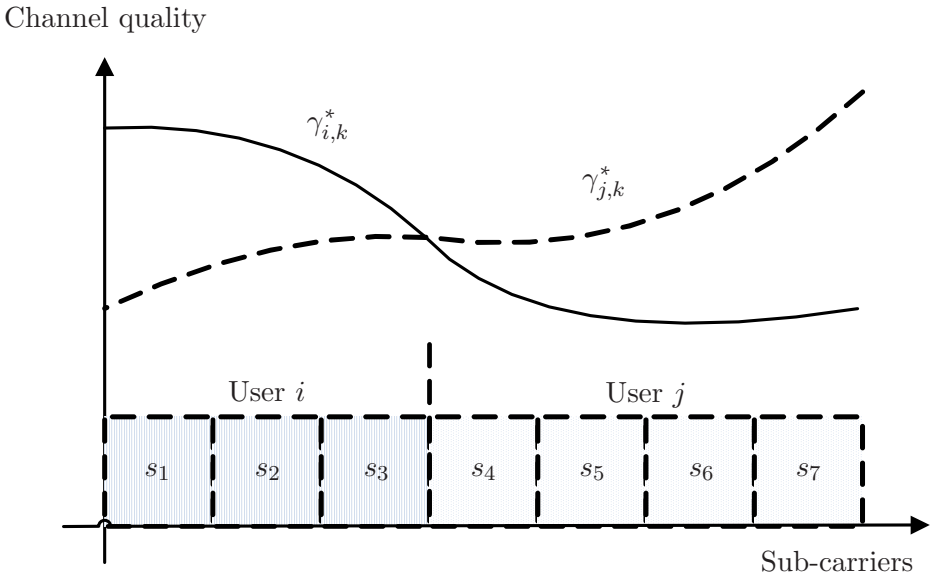


Figure 4.1.: Utility-based Dynamic Sub-carrier Assignment (DSA)

optimum power solution taking into account the constraints (4.4) and (4.5). For example, the value of the water-level μ used in (4.12) is calculated in accordance with constraint (4.4).

Fig. 4.2 illustrates different examples of waterfilling power allocations. We consider a total of seven sub-carriers, where the first three were assigned to the ith MT and the last four were assigned to the jth MT. Let us assume more general waterfilling expressions such as $p_k = [\mu a_k - b_k]^+$, where a_k's and b_k's are arbitrary positive numbers [127]. In the classic waterfilling problem, we have for all $k = 1, \ldots, K$ and $j = 1, \ldots, J$ that $a_k = 1$ and $b_k = \left(\Gamma/\gamma_{j,k}\right)$. In this case, it does not matter for the power allocation how the sub-carriers were assigned. The solution has the visual interpretation of pouring water over a surface given by the inverse of the sub-carriers effective CNRs, as can be seen in Fig. 4.2(a). The water-level μ is calculated taking into account the power constraint $\sum_k p_k = P_{\text{total}}$. Notice that in this example, sub-carrier 7 does not have sufficient quality and so does not receive any power.

Now, let us consider a utility-based waterfilling power allocation, where we have $a_k = w_i^{\text{nrt}}$ for $k = 1, \ldots, 3$, and $a_k = w_j^{\text{nrt}}$ for $k = 4, \ldots, 7$. This problem has also a visual interpretation with a weighted power constraint given by $\sum_k p_k a_k = P_{\text{total}}$, where the scaled marginal utilities (a_k's) are weights that can be visually interpreted as the width of each of the sub-carriers, as can be observed in Fig. 4.2(b). Notice that the weights are dependent on the marginal utilities (see expression (4.10)).

This general expression of the utility-based waterfilling allocation can be also regarded as a multi-level waterfilling, as can be seen in Fig. 4.2(c). In this new visualization, the water-levels are calculated for each user i and j and are given by $\mu_i = \mu w_i^{\text{nrt}}$ and $\mu_j = \mu w_j^{\text{nrt}}$, respectively. Comparing Figs. 4.2(a) and 4.2(c), one can notice that the marginal utilities added to the problem a new kind

of QoS-based prioritization among users, which did not exist in the classic waterfilling allocation. In the utility-based APA considered in this work, the users that have higher marginal utilities will have more power available to their sub-carriers.

4.3.2. Real Time Services

In this section we formulate DSA and APA algorithms that use Utility Theory in order to find an efficient trade-off between system resource efficiency and delay-based fairness among MTs that use RT services. The considered optimization problem is the maximization of the total utility with respect to the HOL packet delays of the users, as indicated in the following.

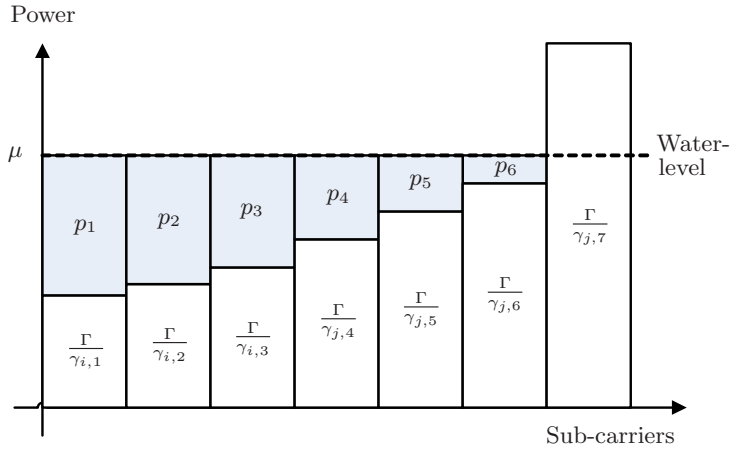
$$\max_{\mathcal{S}_j,\mathbf{p}} \sum_{j=1}^{J} U_j\left(d_j^{\text{hol}}[n]\right), \tag{4.13}$$

where $U_j\left(d_j^{\text{hol}}[n]\right)$ is a concave and decreasing utility function based on the current HOL delay $d_j^{\text{hol}}[n]$ of the jth MT. Notice that for this specific problem, the objective function (4.13) substitutes the general objective function (4.1). Constraints (4.2)-(4.5) are also applied for this problem.

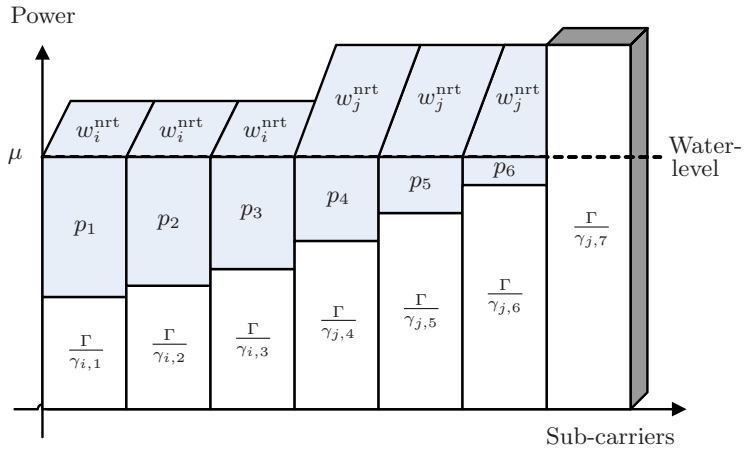
The HOL delay is the time the oldest packet in the user buffer has to wait to gain access to the wireless channel. Considering a generic MT j, it can be calculated approximately by the following recursive equation:

$$d_j^{\text{hol}}[n+1] = d_j^{\text{hol}}[n] + \frac{b_j^{\text{hol}}[n] - R_j[n]\cdot t_{\text{tti}}}{T_j[n-1]} \tag{4.14}$$

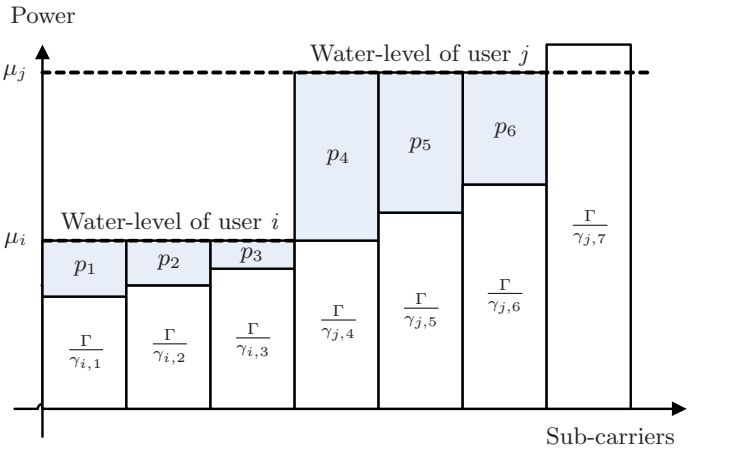
where $b_j^{\text{hol}}[n]$ is the current number of bits in the HOL packet, t_{tti} is the duration of the TTI in seconds, $T_j[n-1]$ is the average data rate (throughput) given by (4.7) up to the previous transmission interval and $R_j[n]$ is the instantaneous achievable transmission rate given by (A.10) on appendix A. If the jth MT has not been served by any sub-carrier in the nth TTI, $R_j[n]$ is equal to zero and the HOL delay is incremented. This delay increment is calculated assuming that the remaining bits of the HOL packet will be transmitted using a rate equal to the throughput experienced so far by the MT. For sake of simplicity, we assume that the packets' interarrival time is equal to the TTI duration. It is well known that in real networks, the packet interarrival time depends on the type of application. For example, for the Voice over IP (VoIP) service, the arrival time is equal to the voice frame duration, i.e. each VoIP packet arrives in the user buffer every 20 ms. However, this assumption does not invalidate the mathematical and conceptual RRA framework, and make the optimization model much more tractable. Taking this into account, one can see in (4.14) that if the instantaneous transmission rate is such that all remaining bits of the HOL packet are transmitted in the current TTI, the HOL delay remains constant because the previous packet in the buffer will be the HOL packet now. Finally, the HOL delay is decremented when the instantaneous achievable transmission rate is high enough to transmit the remaining bits of the HOL packet and some bits of the preceding packets in the queue.



(a) Classic waterfilling



(b) Utility-based general waterfilling



(c) Utility-based multi-level waterfilling

Figure 4.2.: Waterfilling power allocation

Since the joint optimization problem (4.1)-(4.5) is very difficult to be solved when considering (4.13) as the objective function, we adopt the same strategy used in section 4.3.1 for the NRT scenario, and split the problem in two stages, DSA and APA.

Firstly, we will evaluate the optimization problem more carefully and make some simplifications. Assessing the objective function in (4.13) and the HOL delay expression in (4.14), we can see that the derivative of $U_j\left(d_j^{\text{hol}}\right)$ with respect to the transmission rate R_j can be expressed as:

$$\frac{\partial U_j}{\partial R_j} = \frac{\partial U_j}{\partial d_j^{\text{hol}}} \cdot \frac{\partial d_j^{\text{hol}}}{\partial R_j} = \frac{\partial U_j}{\partial d_j^{\text{hol}}} \cdot \left(-\frac{t_{\text{tti}}}{T_j\left[n-1\right]}\right)$$

Using the result above and assuming that the TTI duration is sufficiently small, the Lagrange theorem of the mean can be used, which says that [66, 123]:

$$\begin{aligned}
&\sum_{j=1}^{J} U_j\left(d_j^{\text{hol}}\left[n+1\right]\right) - \sum_{j=1}^{J} U_j\left(d_j^{\text{hol}}\left[n\right]\right) \\
&\approx \sum_{j=1}^{J} \left.\frac{\partial U_j}{\partial R_j}\right|_{R_j=R_j[n-1]} \cdot \left(R_j\left[n\right] - R_j\left[n-1\right]\right) \\
&= \sum_{j=1}^{J} -\left.\frac{\partial U_j}{\partial d_j^{\text{hol}}}\right|_{d_j^{\text{hol}}=d_j^{\text{hol}}[n]} \cdot \frac{t_{\text{tti}}}{T_j\left[n-1\right]} \cdot \left(R_j\left[n\right] - R_j\left[n-1\right]\right) \\
&= \sum_{j=1}^{J} \left|\left.\frac{\partial U_j}{\partial d_j^{\text{hol}}}\right|\right|_{d_j^{\text{hol}}=d_j^{\text{hol}}[n]} \cdot \frac{t_{\text{tti}}}{T_j\left[n-1\right]} \cdot \left(R_j\left[n\right] - R_j\left[n-1\right]\right)
\end{aligned} \tag{4.15}$$

The absolute value operator was used in (4.15) because the utility function was assumed to be concave and decreasing, which yields negative marginal utilities and cancels the negative sign in (4.15). Notice that the maximization of (4.15) leads to the maximization of (4.13). Taking into account (4.15), we have that t_{tti} is a constant and $R_j\left[n-1\right]$ is known and fixed at the nth TTI. So a simplified optimization objective function can be expressed as:

$$\max_{\mathcal{S}_j,\mathbf{p}} \sum_{j=1}^{J} \frac{\left|U_j'\left(d_j^{\text{hol}}\left[n\right]\right)\right|}{T_j\left[n-1\right]} \cdot R_j\left[n\right] \tag{4.16}$$

where $U_j'\left(d_j^{\text{hol}}\left[n\right]\right) = \left.\frac{\partial U_j\left(d_j^{\text{hol}}\right)}{\partial d_j^{\text{hol}}}\right|_{d_j^{\text{hol}}=d_j^{\text{hol}}[n]}$ is the marginal utility of the jth MT with respect to its current HOL delay, and $d_j^{\text{hol}}\left[n\right]$ can be obtained from the recursive equation (4.14). As in the case of the simplified optimization problem for NRT services presented in (4.9), the problem (4.16) is a weighted sum rate maximization [125], where the weights are given by the ratio between the absolute value of the marginal utility with respect to the current HOL delay and the throughput calculated up to the previous TTI.

For didactic reasons that we will see later on, let us define a user-specific weight w_j^{rt} given by

$$w_j^{\mathrm{rt}} = \frac{\left|U_j'\left(d_j^{\mathrm{hol}}[n]\right)\right|}{T_j[n-1]}. \tag{4.17}$$

This utility-based weight plays an important role on the DSA and APA algorithms suitable for RT services, as explained in the following.

Dynamic Sub-carrier Assignment

One can also notice that the optimization objective function in (4.16) is a linear function of $R_j[n]$. According to [66, 123], we can rely on this fact to state that the DSA problem with equal power allocation for RT services employs the following reasoning: the MT $m(k,n)$ is chosen to transmit on the kth sub-carrier at the nth TTI if it satisfies the condition given by (4.18) below:

$$m(k,n) = \arg\max_j \left\{w_j^{\mathrm{rt}} \cdot c_{j,k}[n]\right\}, \tag{4.18}$$

where the weight w_j^{rt} is given by (4.17) and $c_{j,k}[n]$ denotes the instantaneous achievable transmission efficiency of the kth sub-carrier with respect to the jth MT assuming equal power allocation per sub-carrier.

The mechanism of the utility-based DSA algorithm explained in Fig. 4.1 is also valid for the RT scenario. The difference is that the utility-based weight to be considered is w_j^{rt}, which depends on the marginal utility with respect to the HOL delay. The users that have higher weights w_j^{rt} will have priority in the sub-carrier assignment procedure.

Adaptive Power Allocation

Applying the same reasoning used in section 4.3.1 for the NRT scenario, we can conclude that the optimal power allocation for the weighted sum rate maximization problem (4.16) with fixed sub-carrier allocation has a solution in the form of a utility-based multi-level waterfilling [112, 125]:

$$p_k^*[n] = \left[\mu \cdot w_j^{\mathrm{rt}} - \frac{\Gamma}{\gamma_{j,k}[n]}\right]^+, \quad \forall k \in \mathcal{S}_j \tag{4.19}$$

We can look back at Figs. 4.2(b) and 4.2(c) and make a visual interpretation of the APA solution (4.19). Fig. 4.2(b) represents a general utility-based waterfilling problem where the widths of each sub-carrier are weights given by w_i^{rt} for all $k \in \mathcal{S}_i$ belonging to the ith MT and w_j^{rt} for all $k \in \mathcal{S}_j$ belonging to the jth MT. This can also be interpreted as a multi-level waterfilling problem where the water-levels are calculated for each user i and j and are given by $\mu_i = \mu \cdot w_i^{\mathrm{rt}}$ and $\mu_j = \mu \cdot w_j^{\mathrm{rt}}$, respectively, as can be observed in Fig. 4.2(c). In the utility-based APA considered in (4.19), the delay-based marginal utilities and the throughput added to the problem a new kind of QoS-based

prioritization among users, which did not exist in the classic waterfilling allocation (see Fig. 4.2(a)). Therefore, the users that have higher ratios between their delay-based marginal utilities and their throughput will have more power available to their sub-carriers.

4.4. Adaptive Resource Allocation Frameworks

In this section, specific families of utility functions suitable for NRT or RT services are used in the general optimization problems described in section 4.3. As a result, two adaptive RRA frameworks emerge, which we call utility-based alpha-rule and beta-rule. Sections 4.4.1 and 4.4.2 present detailed explanations of these frameworks.

4.4.1. Utility-Based Alpha-Rule for Non-Real Time Services

General Framework

We consider a family of utility functions based on throughput of the form presented in (4.20) below [110].

$$U_j\left(T_j\left[n\right]\right) = \frac{T_j\left[n\right]^{1-\alpha}}{1-\alpha} \tag{4.20}$$

where $\alpha \in [0, \infty)$ is a non-negative parameter that determines the degree of fairness.

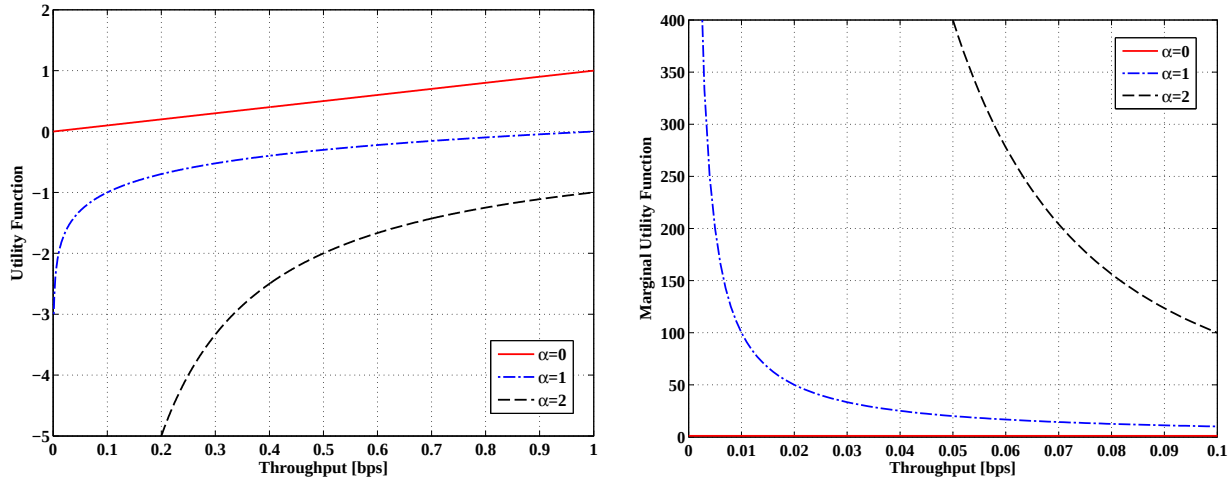
Figs. 4.3(a) and 4.3(b) depict, for different values of α, the utility and marginal utility functions, respectively. Fig. 4.3(a) shows a family of concave and increasing utility functions, which represent that the satisfaction of the MTs increases when their throughput increases. As commented before, the weights w_j^{nrt}, which are dependent on the marginal utilities, play an important role in the DSA and APA algorithms, as explained in section 4.3.1. The higher the weight, the higher the priority of the MT to get a sub-carrier and the higher the amount of power reserved to the sub-carriers assigned to him. Fig. 4.3(b) also shows that MTs experiencing poor QoS (low throughput) will have higher priority in the resource allocation process. And such priority is higher when α increases. Therefore, one can conclude that when α increases, the MTs with poorest QoS are benefited, and so the fairness in the system becomes stricter.

Taking into account (4.20), the expression of the weight w_j^{nrt} becomes

$$w_j^{\mathrm{nrt}} = \frac{1}{T_j\left[n-1\right]^{\alpha}}. \tag{4.21}$$

The corresponding DSA and APA algorithms for the NRT scenario, which are given by (4.11) and (4.12), must use the particular expression of w_j^{nrt} presented in (4.21).

Depending on the value of the fairness controlling parameter α, the alpha-rule framework presented above can be designed to work as different RRA policies, achieving different performances in terms of resource efficiency and throughput-based fairness. The main characteristics of the alpha-rule framework and the four particular RRA policies contemplated by this framework are presented



(a) Utility functions (b) Marginal utility functions

Figure 4.3.: Family of utility functions used in the resource allocation framework for NRT services (utility-based alpha-rule)

in Table 4.1. The first three RRA policies are well-known classic policies, namely Max-Rate (MR), Max-Min Fairness (MMF) and Proportional Fairness (PF), which are described later on. This work also proposes a novel adaptive policy called Adaptive Throughput-Based Fairness (ATF), which is described in details at the end of this section.

Table 4.1.: Features of the utility-based alpha-rule framework - $U_j\left(T_j\left[n\right]\right) = \frac{T_j[n]^{1-\alpha}}{1-\alpha}$

Policies	Parameter α	Weight w_j^{nrt}	Characteristics
MR	0	1	High resource efficiency and low throughput-based fairness
PF	1	$\frac{1}{T_j\left[n-1\right]}$	Static trade-off between resource efficiency and throughput-based fairness
MMF	$\alpha \to \infty$	$\lim_{\alpha\to\infty} \frac{1}{T_j\left[n-1\right]^{\alpha}}$	Low resource efficiency and high throughput-based fairness
ATF	adaptive	$\frac{1}{T_j\left[n-1\right]^{\alpha}}$	Dynamic trade-off between resource efficiency and throughput-based fairness

Max Rate (MR)

The MR policy considers a linear utility function $U_j\left(T_j\left[n\right]\right) = T_j\left[n\right]$, which yields a constant marginal utility $U_j^{'}\left(T_j\left[n\right]\right) = 1$ [61, 112]. Taking into account (4.10), this corresponds to setting $\alpha = 0$ in (4.21).

As the final result according to (4.11) and (4.12), each sub-carrier will be assigned to the MT that

has the highest channel gain on it, and the classic waterfilling power allocation will be implemented over all sub-carriers [61].

It is well known that the MR criterion maximizes the system capacity at the cost of unfairness among the MTs, because those with poor radio link quality will probably not have chance to transmit due to lack of sub-carriers and power.

Max-Min Fairness (MMF)

The utility function of the MMF criterion is the limit of the function in (4.20), when $\alpha \to \infty$ [110].

According to (4.10) and (4.11), the DSA priority function is dependent on the marginal utility $U_j^{'}(T_j[n])$ and the achievable instantaneous transmission efficiency $c_{j,k}[n]$. However, in the case of the MMF criterion and when considering MTs with lower data rates, the influence of the marginal utility when $\alpha \to \infty$ is so high that the influence of the channel quality becomes negligible.

Regarding the MMF APA algorithm, we can see from (4.12) and (4.21) that when α is a big number, small differences in the throughput among the MTs will cause big variations in the weights of the utility-based waterfilling formulation, i.e. the throughput will have a lot of influence in the calculation of the water-levels in the resulting MMF multi-level waterfilling power allocation (see Fig. 4.2(c)). Rather than giving importance to the efficiency in the resource usage, the MMF APA will allocate more power to the sub-carriers assigned to the users with worst QoS values.

The MMF criterion gives priority in the resource allocation (sub-carrier and power) to the MT that has experienced the worst throughput so far. In terms of throughput distribution, it is the fairest criterion possible, since all MTs will have approximately the same average data rate in the long-term. However, since this criterion maximizes the throughput of the worst MTs, it will provide low aggregate system capacity.

Proportional Fairness (PF)

In Utility Theory, the logarithmic utility function is associated with the PF criterion [106, 112]. In the general family of utility functions presented in (4.20), the logarithmic function can be achieved when $\alpha = 1$. The utility function is not well-defined at this point, so let us consider the case where $\alpha \to 1$. Notice that maximizing the sum of $U_j(T_j[n])$ given by (4.20) yields the same optimum as maximizing the sum of $V_j(T_j[n])$ given below [110]:

$$V_j(T_j[n]) = \frac{T_j[n]^{1-\alpha} - 1}{1-\alpha}$$

Applying L'Hospital's rule, it follows that [110]:

$$\lim_{\alpha \to 1} V_j(T_j[n]) = \lim_{\alpha \to 1} \frac{T_j[n]^{1-\alpha} - 1}{1-\alpha} = \log(T_j[n])$$

Therefore, by setting $\alpha = 1$ in (4.21), the priority function of the PF DSA algorithm given by

(4.11) becomes the well-known expression $P_{j,k}^{\mathrm{pf}} = c_{j,k}[n]/T_j[n-1]$, which was firstly proposed for PSC algorithms in HDR CDMA systems [115].

The PF APA solution is derived by setting $\alpha = 1$ in (4.21) and using it on (4.12). One can notice that, differently from MR and MMF, the PF criterion provides a good balance between user QoS (throughput) and channel quality in the power allocation decision.

Based on the particularized DSA and APA algorithms exposed above, one can conclude that a trade-off between resource efficiency and throughput-based fairness can be achieved by means of the PF criterion. The resources will be allocated efficiently taking into account the multi-user and frequency diversities, but at the same time, MTs experiencing poor QoS will have priority in receiving sub-carriers and power.

Adaptive Throughput-Based Fairness (ATF)

We propose in this thesis the ATF policy, which is an adaptive version of the utility-based alpha-rule. It aims to achieve an efficient trade-off between resource efficiency and throughput-based fairness planned by the network operator in a scenario with NRT services. This is done by means of the adaptation of the fairness controlling parameter α in the utility function presented in (4.20). The user priority in the resource allocation is very sensitive to the value of α, as can be seen in Fig. 4.3(b). So small values are sufficient to provide desired fairness degrees on the ATF DSA and APA algorithms.

The ATF policy is based on the definition of a User Fairness Index (UFI) ϕ_j^{nrt}, which is based on throughput and calculated for each NRT MT in the cell. The UFI changes with time and is defined as

$$\phi_j^{\mathrm{nrt}}[n] = \frac{T_j[n-1]}{T_j^{\mathrm{req}}}, \tag{4.22}$$

where T_j^{req} is the throughput requirement of the jth MT.

Next, a fairness index for the whole cell comprising all NRT flows is defined by

$$\Phi_{\mathrm{cell}}^{\mathrm{nrt}}[n] = \frac{\left(\sum_{j=1}^{J} \phi_j^{\mathrm{nrt}}[n]\right)^2}{J \cdot \sum_{j=1}^{J} \left(\phi_j^{\mathrm{nrt}}[n]\right)^2}, \tag{4.23}$$

where J is the number of MTs in the cell. This proposed Cell Fairness Index (CFI) is a particularization of the well-known Jain's fairness index proposed by Jain et al. in [86] and explained in details in section A.7.3 of appendix A. The general Jain's fairness function is independent of the allocation metric being used. In our case, the allocation metric is given by the UFI $\phi_j^{\mathrm{nrt}}[n]$.

Notice that $1/J \leq \Phi_{\mathrm{cell}}^{\mathrm{nrt}}[n] \leq 1$. A perfect fair allocation is achieved when $\Phi_{\mathrm{cell}}^{\mathrm{nrt}}[n] = 1$, which means that the throughput allocated to all MTs are equally proportional to their throughput requirements (all UFIs are equal). The unfairest allocation occurs when $\Phi_{\mathrm{cell}}^{\mathrm{nrt}}[n] = 1/J$, which means that all resources were allocated to only one MT. It is important to notice that the fairness

calculation procedure presented above is general in the sense that different classes of NRT users with different throughput requirements can be contemplated.

The objective of the ATF policy is to assure a strict throughput-based fairness distribution among the MTs, i.e. the instantaneous CFI $\Phi_{\text{cell}}^{\text{nrt}}[n]$ must be kept around a planned value $\Phi_{\text{target}}^{\text{nrt}}$. Therefore, the ATF policy adapts the parameter α in the utility-based alpha-rule framework in order to achieve the desired operation point. Aiming this objective, the new value of the parameter α is calculated using a feedback control loop of the form:

$$\alpha[n] = \alpha[n-1] - \eta_{\text{nrt}} \cdot \left(\Phi_{\text{filt}}^{\text{nrt}}[n] - \Phi_{\text{target}}^{\text{nrt}}\right) \tag{4.24}$$

where $\Phi_{\text{filt}}^{\text{nrt}}[n]$ is a filtered version of the CFI $\Phi_{\text{cell}}^{\text{nrt}}[n]$ using an exponential smoothing filtering, which is used to suppress short-run fluctuations and smooth time series with slowly varying trends; $\Phi_{\text{target}}^{\text{nrt}}$ is the Cell Fairness Target (CFT), i.e. the desired value for the CFI; and the parameter η_{nrt} is a step size that controls the adaptation speed of the parameter α.

The ATF technique is an iterative and sequential process. At each TTI, the steps indicated in Fig. 4.4 are executed. This process is executed indefinitely. After some iterations (TTIs), the ATF technique reaches a stable convergence of the fairness pattern defined by the target CFI. The simplicity of the ATF policy makes it a robust and reliable way to control the trade-off between resource efficiency and throughput-based fairness among NRT flows. By keeping the cell fairness around a planned target value, the network operator can have a stricter control of the network QoS and also have a good prediction about the performance in terms of system capacity.

The formulation of the utility-based alpha-rule framework presented in this section is the one formally proposed and evaluated in this thesis. However, we also proposed a modified version of the alpha-rule framework based on DSA algorithms with equal power allocation using a different range of the parameter α. We show that this modified alpha-rule framework can also be used to balance resource efficiency and user fairness. Furthermore, we present preliminary results that show that this modified framework can be applied to maximize the user satisfaction for different system loads. The modified alpha-rule framework and these two examples of applications are described in details in appendix C.

4.4.2. Utility-Based Beta-Rule for Real Time Services

General Framework

We consider a novel family of utility functions based on the HOL delay of the form presented below:

$$U_j\left(d_j^{\text{hol}}[n]\right) = -\frac{\left(d_j^{\text{hol}}[n]\right)^{1+\beta}}{1+\beta} \tag{4.25}$$

where $\beta \in [0, \infty)$ is a non-negative parameter that determines the degree of delay-based fairness.

Figs. 4.5(a) and 4.5(b) depict, for different values of β, the utility functions and the absolute

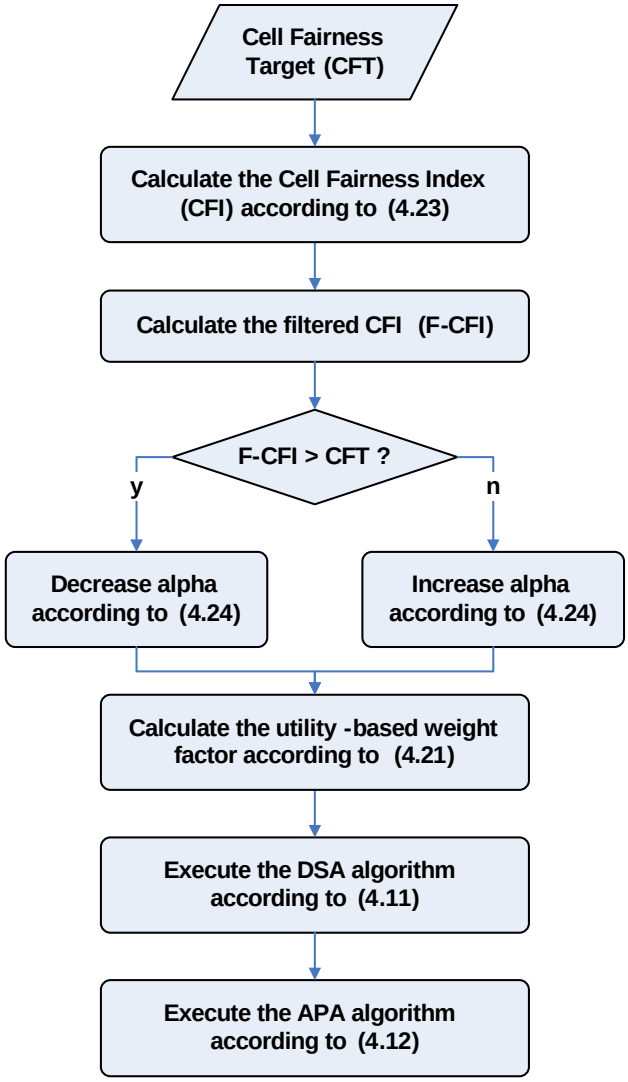
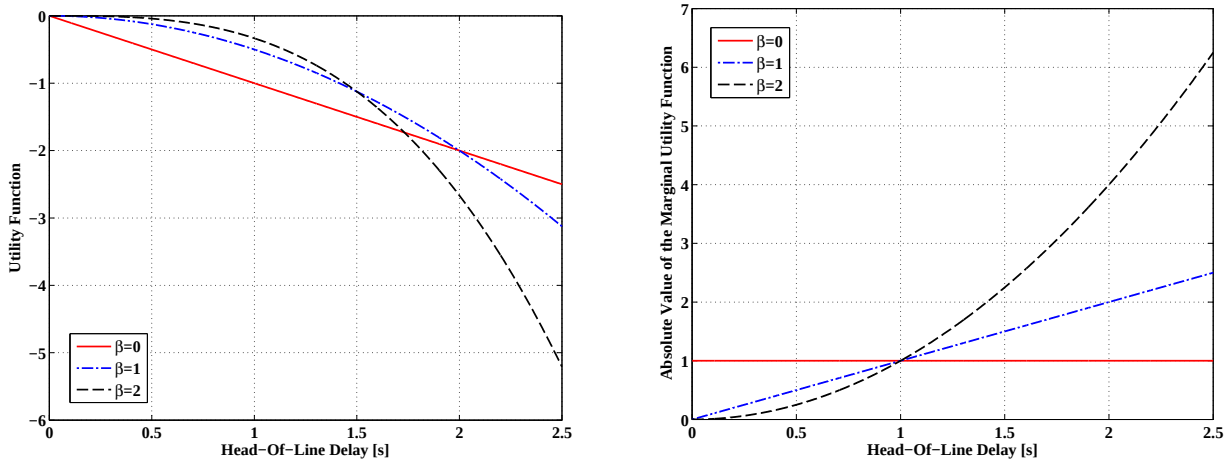


Figure 4.4.: Block diagram of the Adaptive Throughput-Based Fairness (ATF) technique

value of the marginal utility functions, respectively. It is clear that the longer the HOL delay a user experiences, the lower level of satisfaction the user has. Thus, we can assume that $U_j\left(d_j^{\text{hol}}\right)$ is a decreasing and strictly concave function, as shown in Fig. 4.5(a). This implies that the marginal utility, which is the derivative $\partial U_j/\partial d_j^{\text{hol}}$ is a negative and decreasing function. However, the absolute value of the marginal utility is used in our proposed RRA policy suitable for RT services, as a component of the weight w_j^{rt} given by (4.17). Looking at (4.17) and Fig. 4.5(b), one can clearly see that the higher the weight w_j^{rt}, i.e. the higher the HOL delay experienced by a given MT, the higher will be the priority of this MT to get a sub-carrier and the higher will be the amount of power reserved to the sub-carriers assigned to him. And such priority is higher when β increases. Therefore, one can conclude that when β increases, the MTs with poorest QoS (higher HOL delay) are benefited, and so the fairness in the system becomes stricter.



(a) Utility functions (b) Absolute value of the marginal utility functions

Figure 4.5.: Family of utility functions used in the resource allocation framework for RT services

The expression of the weight w_j^{rt} is particularized for the utility function (4.25) as follows:

$$w_j^{\mathrm{rt}} = \frac{\left(d_j^{\mathrm{hol}}[n]\right)^{\beta}}{T_j[n-1]}. \tag{4.26}$$

This particular weight must be used in the corresponding DSA and APA algorithms for the RT scenario, which are given by (4.18) and (4.19), respectively.

Different performances in terms of resource efficiency and delay-based fairness can be achieved depending on the value of the fairness controlling parameter β. Varying β, the beta-rule framework presented above can be designed to work as different classic RRA policies suitable for RT services, such as Proportional Fairness (PF), Modified Largest Weighted Delay First (M-LWDF) and First-In-First-Out (FIFO), which are presented later on. Moreover, this work also proposes a novel adaptive policy called Adaptive Delay-Based Fairness (ADF), which is described at the end of this section. Table 4.2 summarizes the main characteristics of the beta-rule framework and the four particular RRA policies contemplated by this framework.

Proportional Fairness (PF)

Regarding the scenario with RT services, we can achieve the PF policy using a linear utility function $U_j\left(d_j^{\mathrm{hol}}[n]\right) = -d_j^{\mathrm{hol}}[n]$ for all MTs, which provides a constant marginal utility. This is determined when the parameter β is set to zero in the expression of the weight w_j^{rt} given by (4.26).

Due to this fact, the HOL delay $d_j^{\mathrm{hol}}[n]$ is not taken into consideration in the DSA and APA expressions given by (4.18) and (4.19), respectively. As expected, the PF criterion as a particularization of the beta-rule framework with $\beta = 0$ is the same of the one contemplated by the alpha-rule

Table 4.2.: Features of the utility-based beta-rule framework - $U_j\left(d_j^{\text{hol}}[n]\right) = \frac{-\left(d_j^{\text{hol}}[n]\right)^{1+\beta}}{1+\beta}$

Policies	Parameter β	Weight $\mathbf{w_j^{rt}}$	Characteristics
PF	0	$\frac{1}{T_j[n-1]}$	High resource efficiency and low delay-based fairness
M-LWDF	1	$\frac{d_j^{\text{hol}}[n]}{T_j[n-1]}$	Static trade-off between resource efficiency and delay-based fairness
FIFO	$\beta \to \infty$	$\lim_{\beta\to\infty} \frac{d_j^{\text{hol}}[n]^{\beta}}{T_j[n-1]}$	Low resource efficiency and high delay-based fairness
ADF	adaptive	$\frac{d_j^{\text{hol}}[n]^{\beta}}{T_j[n-1]}$	Dynamic trade-off between resource efficiency and delay-based fairness

framework for NRT services (see section 4.4.1).

We expect that the PF criterion will not provide high delay-based fairness. On the other hand, PF is able to provide a good efficiency in the resource usage in a scenario with RT services, since the channel quality plays an important role on the DSA and APA expressions.

First In First Out (FIFO)

The FIFO policy is associated with very strict delay-based fairness given by large values of the β parameter in (4.25). This yields a very steep marginal utility function indicating that the RRA framework will give explicit priority to the MT with the highest HOL delay (see Fig. 4.5(b)), which transforms FIFO in the fairest policy in terms of delay. This happens because the FIFO policy minimizes the maximum HOL packet delays among all users in the system, which yields a low delay variability and consequently a high delay-based fairness.

Considering the impact of the weight w_j^{rt} when $\beta \to \infty$ into the formulation of the DSA and APA algorithms, one can see that the delay component $d_j^{\text{hol}}[n]$ becomes much more prominent than the transmission efficiency or throughput components ($c_{j,k}[n]$ and $T_j[n-1]$, respectively). In this way, the impact of the transmission efficiency and the throughput will be negligible in the sub-carrier assignment and power allocation decisions.

Based on that, one can notice that the FIFO DSA algorithm assumes a Time Division Multiple Access (TDMA) behavior, giving to the user with the highest HOL delay, i.e. the one that has the oldest packet in the MBS buffer, the right to transmit over all sub-carriers. Thus, the FIFO DSA algorithm does not use Channel State Information (CSI) of the sub-carriers, ignoring the frequency diversity offered by the OFDMA system. Furthermore, from (4.19) we can conclude that there will not be delay-based prioritization among users in the APA algorithm because all sub-carriers were allocated to the same user. The resulting power allocation will follow a classic waterfilling approach.

The aforementioned combined strategy of the DSA and APA algorithms based on the FIFO policy is not efficient in the resource usage, so it is expected to provide lower system throughput.

Modified Largest Weighted Delay First (M-LWDF)

The M-LWDF criterion was originally proposed in [117] to be used as a PSC algorithm. It states that its priority function is given by

$$P_{j,k}^{\mathrm{m-lwdf}} = d_j^{\mathrm{hol}}[n] \cdot \frac{c_{j,k}[n]}{T_j[n-1]} \tag{4.27}$$

The authors in [117] also consider a weight for each user in the priority function that is dependent on the maximum due delay time and the maximum allowed probability of the packet delay exceeding this due time, which provides a QoS differentiation among users. However, in the present work we assume that all MTs have the same characteristics and there is no need to use this kind of QoS differentiation.

We generalize the use of the M-LWDF criterion in our utility-based beta-rule framework by considering a utility function of the form $U_j\left(d_j^{\mathrm{hol}}[n]\right) = -\left(d_j^{\mathrm{hol}}[n]\right)^2/2$, which is achieved by setting $\beta = 1$ in (4.25). If we use $\beta = 1$ in the expression of the weight w_j^{rt} given by (4.26), and substitute that in the DSA formulation shown in (4.18), we achieve exactly the same criterion as the original proposal in [117].

The same reasoning must be followed to the M-LWDF APA case. Notice that in this case, we have a QoS differentiation among the MTs characterized by the different water-levels in the multi-level waterfilling problem, which are dependent on both the HOL delay and the throughput of each MT.

Comparing the DSA and APA algorithms using the PF, FIFO and M-LWDF policies, one can notice that M-LWDF is a trade-off between the other two. In this way it should provide intermediate delay-based fairness and resource usage efficiency.

Adaptive Delay-Based Fairness (ADF)

It was shown that a general RRA technique based on (4.25) is able to provide several degrees of delay-based fairness. The ADF policy explores this flexibility in order to achieve an efficient trade-off between resource efficiency and delay-based fairness planned by the network operator in a scenario with RT services. This is done by the ADF policy by means of the adaptation of the fairness controlling parameter β within the utility-based beta-rule framework shown in Table 4.2.

Notice that it is not necessary to use the whole range of β values in order to achieve suitable fairness degrees. In fact, the trade-off between resource efficiency and delay-based fairness is very sensitive to the value of β. It can be seen in Fig. 4.5 that $\beta = 2$ provides a much stricter criterion in terms of delay-based fairness than $\beta = 1$ or 0. Thus, we have that small values are sufficient to provide desired fairness degrees on the ADF DSA and APA algorithms.

The ADF policy is based on the definition of the UFI ϕ_j^{rt}, which depends on the filtered HOL delay and is calculated for each MT in the cell. The UFI changes with time and is defined as:

$$\phi_j^{\mathrm{rt}}[n] = \frac{d_j^{\mathrm{req}}}{d_j^{\mathrm{hol,filt}}[n]} \tag{4.28}$$

Normally, the delay requirement of the jth MT d_j^{req} is the same for all users of the same type and is equal to the delay budget of the RT service (maximum time that a packet can spend in the buffer before being discarded). Note that in (4.28), a filtered version of the HOL delay using a low-pass exponential filtering was considered. This was done in order to smooth the time series and allow a more stable control of the delay-based fairness. The filtering is done as follows:

$$d_j^{\mathrm{hol,filt}}[n] = \left(1 - f_{\mathrm{delay}}\right) \cdot d_j^{\mathrm{hol,filt}}[n-1] + f_{\mathrm{delay}} \cdot d_j^{\mathrm{hol}}[n], \tag{4.29}$$

where $d_j^{\mathrm{hol}}[n]$ is the instantaneous HOL delay of the jth MT calculated by (4.14) and f_{delay} is a filtering constant used in the HOL delay averaging.

Now, a fairness index $\Phi_{\mathrm{cell}}^{\mathrm{rt}}[n]$ for the whole cell comprising all RT flows must be calculated. The same expression 4.23 used for the ATF policy in the NRT scenario is used here. The difference is that the UFI to be used in the calculation of the CFI for the ADF policy must be the one given by (4.28).

Notice that $1/J \leq \Phi_{\mathrm{cell}}^{\mathrm{rt}}[n] \leq 1$, where J is the total number of MTs in the cell. A perfect fair allocation is achieved when $\Phi_{\mathrm{cell}}^{\mathrm{rt}}[n] = 1$, which means that the HOL packet delays of all MTs are equally proportional to their delay requirements (all user fairness indexes are equal). The worst allocation occurs when $\Phi_{\mathrm{cell}}^{\mathrm{rt}}[n] = 1/J$, which means that all sub-carriers were allocated to only one MT, i.e. the HOL packet delay of one user is very low while the others are very high. It is relevant to emphasize that the fairness calculation procedure presented above is general in the sense that different classes of RT users with different delay requirements can be contemplated.

The objective of the ADF policy is guarantee that the instantaneous CFI $\Phi_{\mathrm{cell}}^{\mathrm{rt}}[n]$ converges to a planned value $\Phi_{\mathrm{target}}^{\mathrm{rt}}$ by adapting the parameter β in the utility-based beta-rule framework. The parameter β is adapted using a feedback control loop, as indicated below.

$$\beta[n] = \beta[n-1] - \eta_{\mathrm{rt}} \cdot \left(\Phi_{\mathrm{filt}}^{\mathrm{rt}}[n] - \Phi_{\mathrm{target}}^{\mathrm{rt}}\right) \tag{4.30}$$

where $\Phi_{\mathrm{filt}}^{\mathrm{rt}}[n]$ is a filtered version of the CFI $\Phi_{\mathrm{cell}}^{\mathrm{rt}}[n]$ using an exponential smoothing filtering; $\Phi_{\mathrm{target}}^{\mathrm{rt}}$ is the desired value for the CFI (CFT); and the parameter η_{rt} is a step size that controls the adaptation speed of the parameter β.

The ADF technique is an iterative and sequential process just like ATF. At each TTI, steps similar to the ones presented in Fig. 4.4 for the ATF technique, are executed. These steps are:

1. Calculate the CFI $\Phi_{\mathrm{cell}}^{\mathrm{rt}}[n]$ according to (4.23), but replacing $\phi_j^{\mathrm{nrt}}[n]$ by the UFI $\phi_j^{\mathrm{rt}}[n]$ given

by (4.28);

2. Calculate the filtered CFI $\Phi_{\text{filt}}^{\text{rt}}[n]$;
3. Calculate β using (4.30);
4. Calculate the utility-based weight factor w_j^{rt} according to (4.26);
5. Execute the DSA algorithm as indicated by (4.18);
6. Execute the APA algorithm according to (4.19).

The network management benefits of controlling the CFI are also possible for a scenario with RT services thank to the use of the adaptive ADF policy.

We also proposed a variant of the utility-based beta-rule that uses a different range of values for the parameter β and can be used by DSA algorithms with equal power allocation suitable for RT services. Details about this variant are given in appendix C.

4.5. Simulation Results

The performances of the utility-based alpha-rule and beta-rule frameworks are evaluated by means of system-level simulations in sections 4.5.1 and 4.5.2, respectively. The simulations took into account the main characteristics of an OFDMA system. In the following, some common aspects regarding both evaluations are discussed.

The general simulation parameters used in both evaluations are depicted in Table 4.3. More details about the simulation modeling and the system-level simulator used in this study can be found in appendix A.

The RRA frameworks are evaluated taking into account the following analyses and performance metrics:

- Preliminary evaluation of the classic RRA policies: mean user throughput or 90th percentile of packet delays versus the number of users;
- Convergence of the adaptive algorithms;
- Fairness analysis: cell fairness index based on throughput or delay versus the number of users;
- Efficiency analysis: total cell throughput versus the number of users and efficiency-fairness plane;
- User satisfaction analysis: satisfaction versus the number of users and satisfaction-fairness plane. A NRT user is considered satisfied if its session throughput is higher than a threshold (see the value considered in the simulations in Table 4.4). A RT user is considered satisfied if its Frame Erasure Rate (FER) is lower than a threshold. In our simulation model, we

Table 4.3.: General simulation parameters for the evaluation of the utility-based frameworks

Parameter	Value
Number of cells	1
Maximum BS transmission power	1 W
Cell radius	500 m
MT speed	static
Carrier frequency	2 GHz
Number of sub-carriers	192
Effective sub-carrier bandwidth	14 kHz
Path loss	using (A.1)
Log-normal shadowing standard dev.	8 dB
Small-scale fading	Typical Urban (TU)
AWGN power per sub-carrier	-123.24 dBm
BER requirement	10^{-6}
Link adaptation	using (A.9)
Transmission Time Interval (TTI)	0.5 ms
Simulation time span	5 s
Number of independent simulation runs	70

assume that a frame is lost if a packet arrives at the MT receiver later than the delay budget of the RT service. The values of the FER threshold and the delay budget considered in the simulations are presented in Table 4.8;

- Central Processing Unit (CPU) time: CPU time-fairness plane;
- Opportunistic allocation analysis: mean user throughput or 90th percentile of packet delays versus number of users;

In order to emphasize some aspects of the trade-off between resource efficiency and user fairness, the analysis of inner and outer users is done. This analysis consists in separating equally the users within a macrocell in two groups of same size, inner and outer users, based on path loss and shadowing. The former are the ones closest to the MBS (inner zone of the cell) that experience better channel conditions, while the latter are those that are far from the MBS (outer zone of the cell). Figs. A.2(b) and A.2(c) of appendix A present a visualization of this separation. Notice that this evaluation approach is also very useful for the study of the compromise between fairness and coverage (see chapter 1 for a brief explanation about this compromise).

In the evaluation of the RRA frameworks, two approaches were considered: joint and Equal Power Allocation (EPA). In the former, both DSA and APA algorithms use a given RRA policy, for example PF. The latter approach means that the chosen policy is used only on the DSA algorithm, while on the power allocation step the total transmission power is equally divided among the sub-carriers.

4.5.1. Performance Evaluation of the Alpha-Rule Framework

Table 4.4 shows the specific simulation parameters used in the performance evaluation of the utility-based alpha-rule framework.

Table 4.4.: Specific simulation parameters for the evaluation of the utility-based alpha-rule framework

Parameter	Value
NRT traffic model	Full buffer
Throughput filtering time constant (f_{thru})	1000
Minimum α value	0
Maximum α value	10
ATF control time window	0.5 ms
ATF fairness target ($\Phi^{\text{nrt}}_{\text{target}}$)	Variable
ATF step size (η_{nrt})	0.1
ATF filtering time constant	10
User throughput requirement (T_j^{req})	512 kbps

Preliminary Analysis of the Classic Policies

As explained in section 2.8 of chapter 2, the conflict between resource efficiency and user fairness appears in scenarios where there are users with different efficiency indicators with regard to the resources. Opportunistic resource allocation algorithms that try to explore this multi-user diversity automatically generate an uneven resource distribution among the users, which will cause fairness problems. This panorama can be seen in Fig. 4.6, which shows the mean user throughput as a function of the number of NRT users for various classic RRA policies considering inner and outer groups.

It can be observed that the MR policy, which is a pure channel-based opportunistic policy, benefits the users with better channel conditions (inner group) and prevent the users of the outer group from accessing the resources. As a consequence, the inner group has very high mean user throughput while the outer group presents the worst performance. The use of a non-opportunistic RRA policy, which is the case of MMF, leads to a similar performance between the two groups. The drawback of not using the resources efficiently is the low overall mean user throughput. The PF policy, which takes into account both channel quality and user QoS, provides a better balance between the two groups. The mean user throughput of the inner group provided by PF is worse than MR, but the performance of the outer group is also better than the one provided by MMF.

Fig. 4.6 clearly shows that the classic RRA policies evaluated in this work treat the inner and outer groups in very different ways, depending on how these policies reckon the importance of the opportunistic allocation and the user QoS. This preliminary analysis reaffirms the existence of a trade-off between resource efficiency and user fairness in cellular networks. This motivated

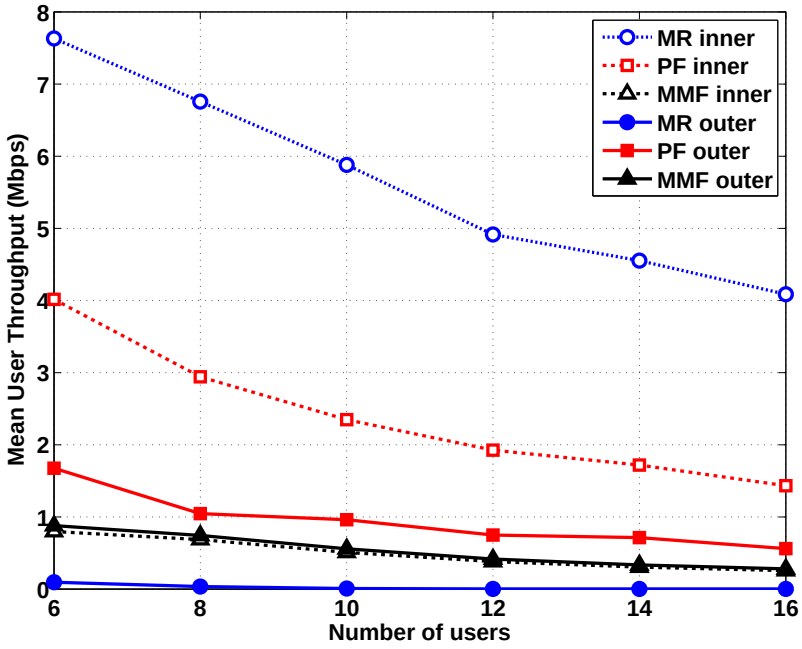


Figure 4.6.: Mean user throughput as a function of the number of users for various classic RRA policies considering inner and outer groups

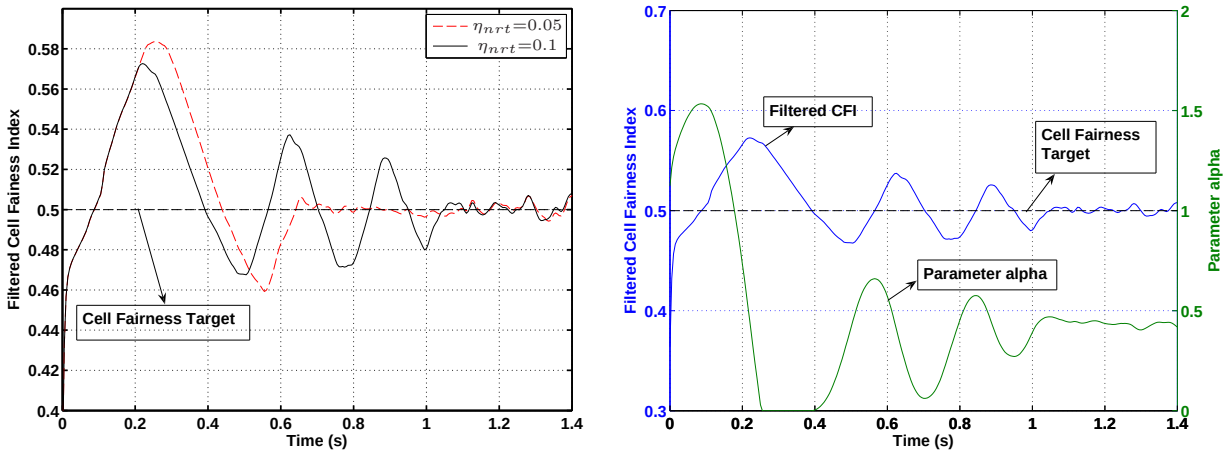
the proposal of the adaptive ATF policy, which is able to explore the flexibility of the utility-based alpha-rule framework and provide several planned degrees of fairness. In this way, the aforementioned trade-off can be managed.

Convergence Analysis of the ATF Policy

Before showing the performance evaluation of the ATF policy, let us first analyze its convergence. The ATF policy relies on the adaptation of the parameter α of the alpha-rule framework in order to achieve a desired target for the cell fairness index (see expression (4.24)). An example of such a convergence is presented in Fig. 4.7(a). This figure shows the convergence of the filtered CFI $\Phi^{\text{nrt}}_{\text{filt}}$ over several TTIs using two different values of the step size η_{nrt}. The target CFT $\Phi^{\text{nrt}}_{\text{target}}$ was set to 0.5. Notice that the convergence speed and oscillation directly depend on η_{nrt}. A higher step size ($\eta_{\text{nrt}} = 0.1$) allows $\Phi^{\text{nrt}}_{\text{filt}}$ to reach the target $\Phi^{\text{nrt}}_{\text{target}}$ earlier, but the oscillation around this value is higher. With the lower step size $\eta_{\text{nrt}} = 0.05$, $\Phi^{\text{nrt}}_{\text{filt}}$ cross the target CFT later, but oscillates with a small error.

Fig. 4.7(b) shows the adaptation of the parameter α in accordance to the variation of the filtered CFI $\Phi^{\text{nrt}}_{\text{filt}}$. Notice again that $\Phi^{\text{nrt}}_{\text{target}} = 0.5$. As stipulated by (4.24) and also illustrated in Fig. 4.4, when $\Phi^{\text{nrt}}_{\text{filt}} > \Phi^{\text{nrt}}_{\text{target}}$, there is an excess of fairness in the system, and so α must be decreased in the alpha-rule framework in order to decrease $\Phi^{\text{nrt}}_{\text{filt}}$. In the opposite way, if $\Phi^{\text{nrt}}_{\text{filt}} < \Phi^{\text{nrt}}_{\text{target}}$, α is increased by the ATF policy, and as a consequence $\Phi^{\text{nrt}}_{\text{filt}}$ is also increased. It is interesting to observe that the

points of crossing between the CFI $\Phi_{\text{filt}}^{\text{nrt}}$ and the CFT $\Phi_{\text{target}}^{\text{nrt}}$ correspond to minimum or maximum values of the parameter α, i.e. points in which α changes its tendency of increase or decrease.



(a) Example of convergence of the throughput-based filtered cell fairness index for various step sizes η_{nrt}

(b) Example of variation of the throughput-based filtered cell fairness index and adaptation of the parameter α

Figure 4.7.: Convergence analysis of the ATF policy

Fairness Analysis

The throughput-based CFI calculated by (4.23) averaged over all simulation snapshots is depicted in Fig. 4.8 for various system loads and considering the RRA policies with equal power allocation. The performance of the ATF policy is compared to the three classic RRA policies (MMF, PF and MR). In this simulation scenario, several CFTs were considered for the ATF policy, namely $\Phi_{\text{target}}^{\text{nrt}} = [1/J, 0.2, 0.4, 0.6, 0.8, 1.0]$. It can be observed that ATF is successful at achieving its main objective, which is to guarantee a strict fairness distribution among the NRT MTs. This is achieved due to the feedback control loop that dynamically adapts the parameter α of the alpha-rule framework.

Notice that the structure of the utility-based alpha-rule framework bounds the performance of the ATF policy between the performances of the MR and MMF policies. According to Table 4.1, the extremes values of the parameter α are 0 and ∞ (in practice a very large number), which correspond to MR and MMF policies, respectively. We considered in the simulations a range of values from 0 to 10 for the adaptation of the parameter α by the ATF policy. Notice that this upper limit of $\alpha = 10$ was sufficient for the ATF policy configured with $\Phi_{\text{target}}^{\text{nrt}} = 1.0$ to be very close to the performance of the MMF policy. On the other extreme, it is clear that MR works as a lower bound for ATF configured with $\Phi_{\text{target}}^{\text{nrt}} = 1/J$.

Regarding the classic RRA policies, as expected, MMF provided the highest fairness, very close to the maximum value of 1, while MR was the unfairest strategy with a high variance on the fairness

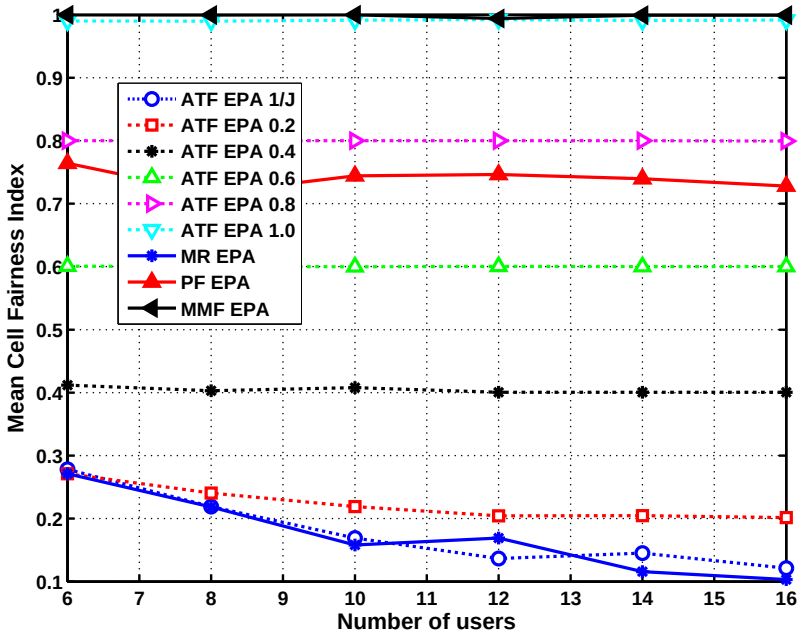


Figure 4.8.: Mean cell fairness index as a function of the number of users for the utility-based alpha-rule framework

distribution for high cell loads. PF presented a good intermediate fairness distribution.

It is interesting to notice the relationship between the cell fairness index shown in Fig. 4.8 and the throughput performance of the inner and outer groups presented in Fig. 4.6. This group analysis is very important to give us an insight about the distribution of fairness in the system. Focus on the difference between the inner and outer curves within the same algorithm in Fig. 4.6. For instance, if we consider MR, the difference between the throughput curves is the biggest. This means that MR assigns more priority to inner users than to outer users, producing a sensible decrease in fairness (see Fig. 4.8). On the other extreme, MMF is the algorithm with the smallest difference between inner and outer curves, which demonstrates that it is the fairest algorithm is terms of user throughput. Since PF is a trade-off, it presents an intermediate behavior. In summary, the lower the difference in performance between the inner and outer groups, the higher the CFI.

From this fairness analysis, it can be concluded that the advantage of the ATF policy compared with the classic RRA strategies is that the former can be designed to provide any required fairness distribution, while the latter are static and do not have the freedom to adapt themselves and guarantee a specific performance result.

Finally, let us compare the performance of the joint and EPA variants of the studied RRA policies presenting Table 4.5, which depicts the mean CFI as a function of the number of users for both variants. It can be observed that there is no significant difference in performance between the joint and EPA approaches regarding the cell fairness index. It means that the main contribution for the fairness result is due to the execution of the DSA algorithm. Although the Table presents only the

results of the ATF policy with a CFT equal to 0.6, it was observed in the simulation results that both variants of ATF achieved the same results for other CFTs.

Table 4.5.: Mean cell fairness index as a function of the number of users for the utility-based alpha-rule framework

RRA Policies	Number of users					
	J=6	J=8	J=10	J=12	J=14	J=16
ATF EPA 0.6	0.601	0.600	0.600	0.600	0.600	0.600
ATF Joint 0.6	0.601	0.600	0.600	0.600	0.600	0.600
MR EPA	0.271	0.218	0.158	0.169	0.116	0.103
MR Joint	0.282	0.210	0.178	0.150	0.129	0.122
PF EPA	0.764	0.717	0.744	0.747	0.740	0.728
PF Joint	0.741	0.743	0.745	0.756	0.749	0.745
MMF EPA	1.000	1.000	1.000	0.994	0.999	0.999
MMF Joint	1.000	1.000	1.000	1.000	1.000	1.000

Efficiency Analysis

In this section, we analyze how the RRA policies behave in terms of efficiency in the resource usage. We consider the total cell throughput (cell capacity) as the efficiency indicator, which is presented in Fig. 4.9 as a function of the number of users in the macrocell for various policies using the EPA approach.

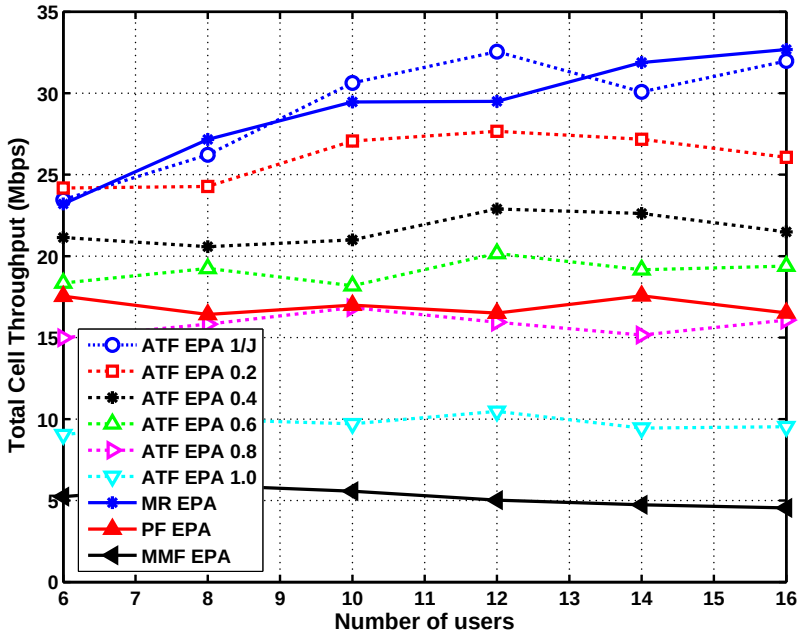


Figure 4.9.: Total cell throughput as a function of the number of users for the utility-based alpha-rule framework

As expected, MR was able to maximize the system capacity, while MMF presented the lowest cell throughput, since it is not able to exploit efficiently the available resources. PF is a trade-off between MR and MMF, so its performance lied between them. The ATF policy is able to achieve several cell throughput performances depending on the value of the chosen CFT. In this way, we realize that ATF is able to work as a hybrid policy between any classic RRA strategy contemplated in the framework.

Table 4.6 shows the comparison between the joint and EPA approaches. Again, it is observed that the approaches present similar performance in terms of system capacity, which means that the APA algorithm has a little impact on the system capacity in the scenario studied in this work.

Table 4.6.: Total cell throughput in Mbps as a function of the number of users for the utility-based alpha-rule framework

RRA Policies	Number of users					
	J=6	**J=8**	**J=10**	**J=12**	**J=14**	**J=16**
ATF EPA 0.6	18.355	19.248	18.181	20.159	19.166	19.390
ATF Joint 0.6	18.902	17.486	19.227	19.598	19.011	19.253
MR EPA	23.200	27.160	29.460	29.504	31.884	32.686
MR Joint	24.920	25.825	27.811	32.426	30.543	32.392
PF EPA	17.537	16.417	16.999	16.503	17.560	16.513
PF Joint	16.027	17.555	17.596	17.742	18.441	17.878
MMF EPA	5.244	5.933	5.575	5.027	4.737	4.552
MMF Joint	6.841	5.501	6.886	4.861	6.326	5.789

Looking at Figs. 4.8 and 4.9, one can clearly see the conflicting objectives of capacity and fairness maximization, and how MR and MMF are able to achieve one objective in detriment of the other. PF and ATF were able to achieve a static and a dynamic trade-off, respectively. A good way to explicitly evaluate the trade-off between resource efficiency and user fairness is to combine Figs. 4.8 and 4.9 and plot a 2D plane between total cell throughput (capacity) and the cell fairness index. We call this visualization as the efficiency-fairness plane and it is a novel contribution of this thesis, as explained in section 3.2 of chapter 3. This visualization tool was already used in this thesis to evaluate the aforementioned trade-off when using fairness/rate adaptive RRA techniques (see section 3.5 of chapter 3 for more details). Fig. 4.10 presents the plane built from the simulations of all variants of the studied RRA policies on a scenario with 16 active NRT flows.

In Fig. 4.10, the classic RRA policies are indicated as single markers, and the adaptive policy ATF is indicated as solid and dashed lines (joint and EPA variants, respectively). One can clearly see the static behavior of the classic policies on the efficiency-fairness plane. MMF is able to provide maximum throughput-based fairness at the expense of low system capacity, while MR is the most efficient on the resource usage but provides an unfair throughput distribution among users. The PF policy appears as a fixed trade-off between MMF and MR, with intermediate throughput-based fairness and system capacity.

The ATF policy, which controls the parameter α adaptively according to (4.24) in order to

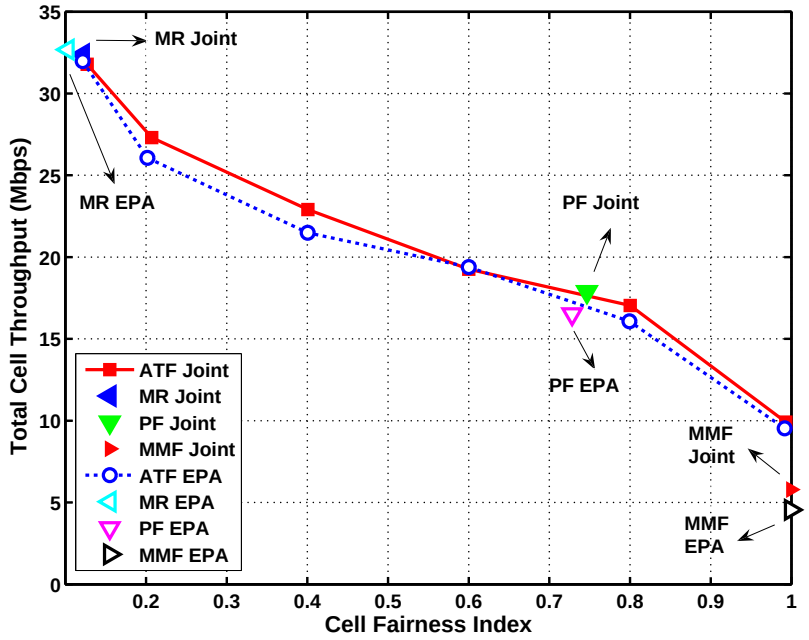


Figure 4.10.: Efficiency-Fairness plane for the utility-based alpha-rule framework

achieve a desired cell fairness target, is able to cover the whole path between the classic policies in the efficiency-fairness plane. Notice in the ATF curves that the fairness targets set in the simulations (0.2, 0.4, 0.6, 0.8 and 1.0) are always met. One can observe that the performance of the ATF policy for very low fairness range converges to the performance of the MR policy, as expected. In this way, it can be concluded that the ATF policy can adaptively adjust the utility-based alpha-rule framework presented in Table 4.1 in order to provide a dynamic trade-off between resource efficiency and throughput-based fairness.

It can also be seen in Fig. 4.10 that there is not a considerable advantage in using an adaptive power allocation for the problem and scenario considered in this work. The joint approach presents a small gain in cell throughput for the same CFIs compared with the EPA approach. For the case of ATF, this gain is due to the faster convergence of the parameter α when both DSA and APA algorithms are used. Furthermore, α stabilizes in lower values when APA is used, which yields higher cell throughput. This small gain comes at the expense of higher computational cost, as will be explained later on.

Satisfaction Analysis

The user satisfaction as a function of the number of NRT users for different RRA policies using EPA is depicted in Fig. 4.11. It is interesting to see that the policies that achieve a trade-off between resource efficiency and user fairness, namely PF and ATF, are the ones that present the highest user satisfaction. This indicates that it is not advantageous in terms of user satisfaction to

use RRA policies that are located in the extremes of the efficiency-fairness plane (maximization of system capacity or maximization of user fairness).

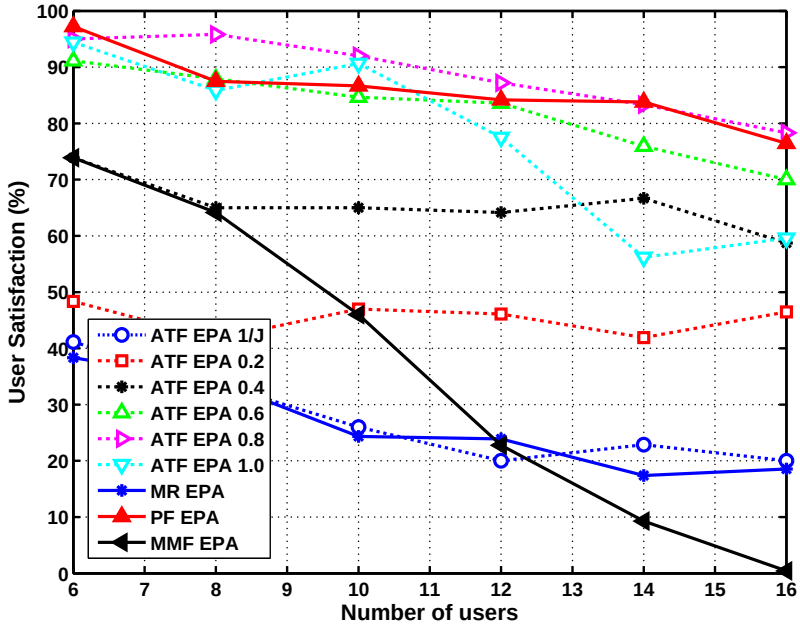


Figure 4.11.: User satisfaction as a function of the number of users for the utility-based alpha-rule framework

In general terms, the MR policy, which maximizes capacity, achieves low and almost constant user satisfaction for a number of users higher than 10 because it always give priority to few users with best channel conditions. On the other hand, the MMF policy, which privileges user fairness, presents higher satisfaction for low system loads but the performance decreases very fast when the number of users increases. Regarding the policies able to achieve a trade-off, PF shows a very good result in terms of user satisfaction because it takes into account both the channel quality and the QoS of the users (throughput). However, the flexibility of the adaptive ATF policy allows several performances regarding user satisfaction, and the ATF policy configured with a CFT equal to 0.8 provides satisfaction results even better than PF for all considered system loads.

Table 4.7 presents the comparison between the joint and EPA approaches in terms of the percentage of user satisfaction. In general, it can be noticed that applying the corresponding RRA criterion in both DSA and APA algorithms provides slight gains over the case where the utility-based DSA algorithm is followed by an equal power allocation step.

We use an approach similar to the one previously used in the efficiency analysis in order to create a 2D plot for the satisfaction analysis. We combine Figs. 4.8 and 4.11 and create a satisfaction-fairness plane for a fixed number of 16 NRT users, as depicted in Fig. 4.12. This performance result is a valuable tool for the analysis of the compromise between fairness and QoS (see chapter 1 for more details about this compromise). In the figure, the classic RRA policies are again indicated

Table 4.7.: Percentage of user satisfaction as a function of the number of users for the utility-based alpha-rule framework

RRA Policies	Number of users					
	J=6	J=8	J=10	J=12	J=14	J=16
ATF EPA 0.6	91.111	87.917	84.667	83.611	75.952	70.000
ATF Joint 0.6	90.000	86.667	84.667	81.389	75.238	72.292
MR EPA	38.333	34.167	24.333	23.889	17.381	18.542
MR Joint	38.889	34.583	28.000	22.500	20.714	19.583
PF EPA	97.222	87.500	86.667	84.167	83.810	76.458
PF Joint	98.333	92.083	91.333	87.778	86.191	82.292
MMF EPA	73.889	64.167	46.000	22.778	9.286	0.417
MMF Joint	89.444	62.917	69.000	22.500	27.857	8.333

as single markers, and the adaptive policy ATF is indicated as solid and dashed lines (joint and EPA variants, respectively).

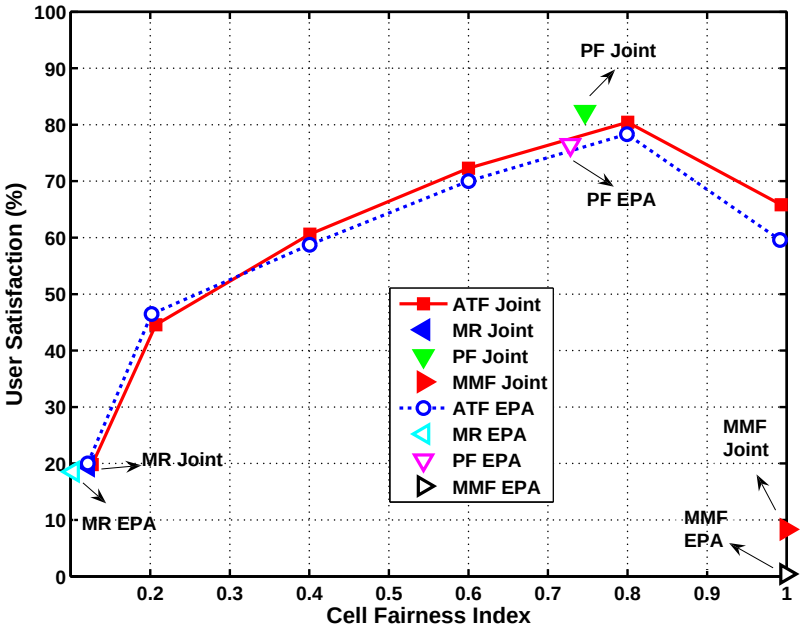


Figure 4.12.: Satisfaction-Fairness plane for the utility-based alpha-rule framework

The same conclusions taken from Fig. 4.11 still remain valid. The advantage of Fig. 4.12 is the possibility to analyze the performance of the ATF policy for the whole range of possible CFTs. Considering the lower limit of the fairness index, it can be observed that the ATF performance is very close to MR, while for the upper limit, ATF provides a much better satisfaction than MMF. Furthermore, it becomes clear again that policies that are able to balance resource efficiency and fairness provide better satisfaction results, as indicated by PF and ATF configured with intermediate CFT values. Finally, it can be concluded that the utility-based APA algorithm does not play a

very significant role regarding user satisfaction, since the gain of the joint approach in comparison with the EPA approach is small.

CPU Time Analysis

It was shown in Figs. 4.10 and 4.12 by means of the efficiency-fairness and satisfaction-fairness planes that the joint approach provided small gains in comparison with the EPA approach. The former executes both the utility-based DSA and APA algorithms, while the latter executes the utility-based DSA algorithm and a subsequent equal power allocation step.

As expected, this performance gain comes at the expense of a higher processing time, as indicated in the CPU time-fairness plane depicted in Fig. 4.13. This excess in the demanded processing power is higher for the ATF policy than the classic RRA policies. Therefore, it is concluded that the small gain achieved by the joint approach does not worth the higher computational burden, and so the APA algorithm can be simplified to an equal power allocation.

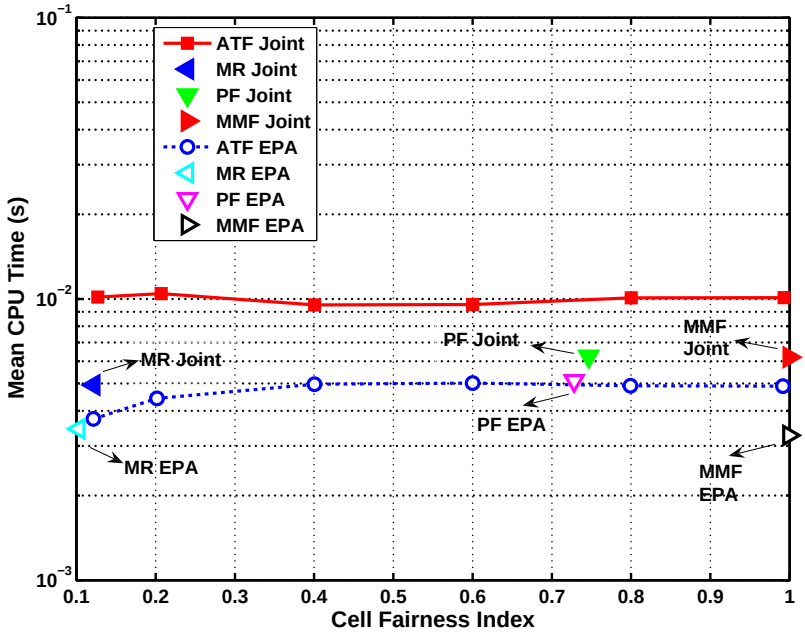


Figure 4.13.: CPU time-Fairness plane for the utility-based alpha-rule framework

It is important to emphasize that the ATF policy using EPA demands a processing time comparable with the classic policies, which demonstrates that the benefits achieved by ATF are not spoilt by an excessive computational burden.

Opportunistic Allocation Analysis of the ATF Policy

In this section, we consider again the performance of the separate inner and outer groups in order to analyze in more details how the ATF policy makes use of opportunistic allocation. Fig. 4.14

presents the mean user throughput using the ATF policy with various CFTs. It can be seen that the higher the CFT, the higher the overall fairness in the system, and the shorter the distance between the performances of the inner and outer groups. For the minimum CFT, ATF works as a pure opportunistic policy and explicit priority is given to the users of the inner group, which have better channel conditions. When the CFT increases, the ATF policy degrades the performance of the users with best channel quality and give more resources to the users of the outer group, improving their QoS. For the maximum CFT, both groups experience the same QoS in terms of mean user throughput. Notice that when the CFT increases from the minimum to the maximum value, the QoS degradation of the inner group is more prominent than the QoS improvement of the outer group. That is another reason why it is more advantageous to find a trade-off policy with an intermediate CFT value that is able to find a good balance between the inner and the outer groups.

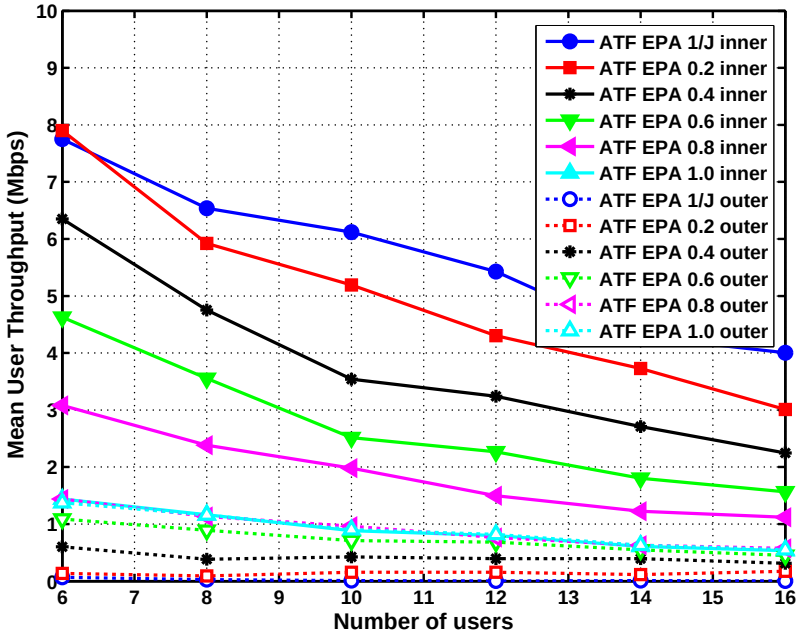


Figure 4.14.: Mean user throughput as a function of the number of users for the ATF policy considering inner and outer groups

4.5.2. Performance Evaluation of the Beta-Rule Framework

The specific simulation parameters used in the performance evaluation of the utility-based beta-rule framework are presented in Table 4.8.

Table 4.8.: Specific simulation parameters for the evaluation of the utility-based beta-rule framework

Parameter	Value
RT traffic model	Packets of 32 bytes with interarrival time of 2 ms
HOL delay filtering time constant (f_{delay})	100
Minimum β value	0
Maximum β value	10
ADF control time window	0.5 ms
ADF fairness target ($\Phi^{\text{rt}}_{\text{target}}$)	Variable
ADF step size (η_{rt})	0.1
ADF filtering time constant	10
FER threshold	2%
RT delay budget	100 ms

Preliminary Analysis of the Classic Policies

The classic RRA policies suitable for RT services that are particular cases of the utility-based beta-rule, namely PF, M-LWDF and FIFO, are assessed in Fig. 4.15. This figure presents the 90th percentile of packet delays separating the users in the inner and outer groups and considering the EPA approach. As expected, the 90th percentile of the delays of the users that belong to the outer group, i.e. those ones with worse channel conditions, is higher than the experienced by the inner group, except in the case of FIFO that provides the same results for inner and outer users as expected.

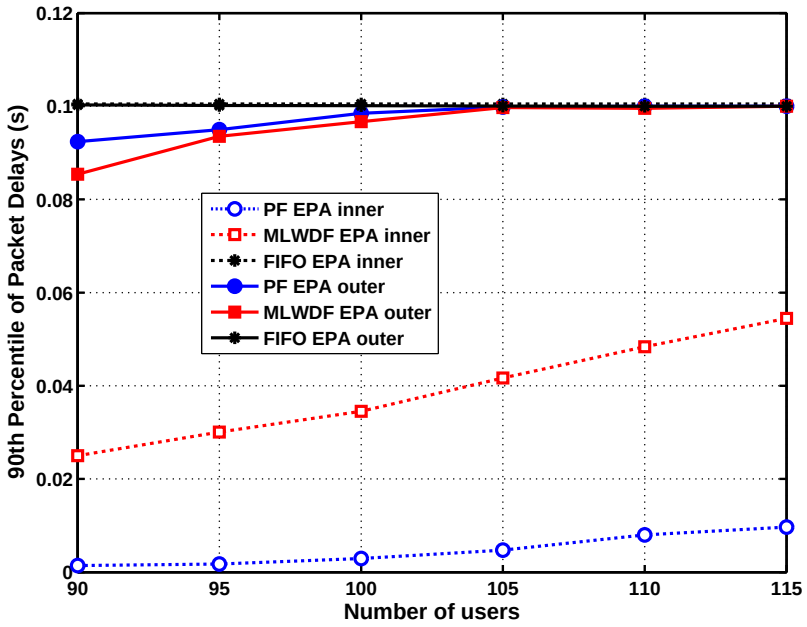


Figure 4.15.: 90th percentile of packet delays as a function of the number of users for various classic RRA policies considering inner and outer groups

Comparing the aforementioned classic policies, PF is the most opportunistic policy, since it takes into account only the channel quality and the throughput of the user. That is why PF shows the lowest packet delays for the inner group. The M-LWDF policy does not only consider the channel quality and the throughput of the user, but the HOL packet delay as well. The consequence of this criterion is twofold: the performance of the outer group is slightly better due to QoS prioritization and the performance distance between the inner and outer groups is shorter. Finally, FIFO is a non-opportunistic policy that gives priority to the users only based on the time that their packets are waiting in the buffers. Its inefficiency in the usage of the resources makes the performance of the groups to be very similar, but with unacceptable delays, as indicated in Fig. 4.15. Since the RT delay budget considered in the simulations is 100 ms, it can be concluded that the FIFO policy is not able to provide an acceptable QoS for the RT service for the system loads considered in Fig. 4.15.

It becomes evident that these classic policies have little control over the trade-off between resource efficiency and delay-based fairness. The proposed adaptive ADF policy, which fully explores the flexibility of the utility-based beta-rule framework, can fill this gap, as will be shown in the following sections.

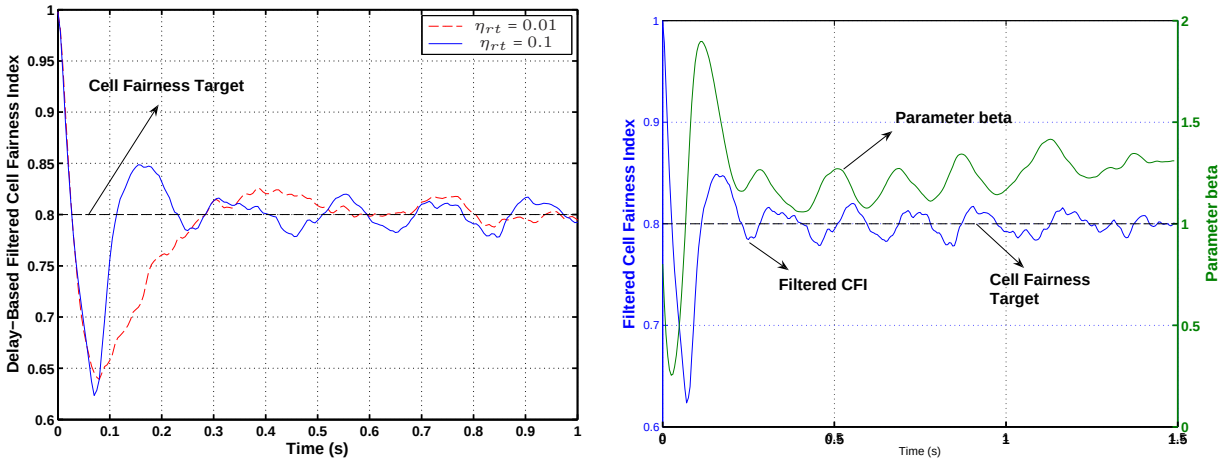
Convergence Analysis of the ADF Policy

The ADF policy was meant to provide to the network operator the possibility to control the fairness distribution by setting a target cell fairness index (CFT). In this section, we present an example of the convergence of the filtered CFI $\Phi^{\mathrm{rt}}_{\mathrm{filt}}$ when using the ADF technique, which is illustrated in Fig. 4.16(a) for a CFT equal to 0.8. In this figure, two different step sizes η_{rt}, which are used in the control loop of the parameter β (see expression (4.30)), are assessed. It can be noticed that a higher value of η_{rt} allows a faster convergence in comparison with a lower value, but the steady-state error is also higher.

Fig. 4.16(b) shows how the parameter β responds to the variation of the filtered CFI $\Phi^{\mathrm{rt}}_{\mathrm{filt}}$. When $\Phi^{\mathrm{rt}}_{\mathrm{filt}}$ is higher than 0.8, which is the chosen CFT, the ADF technique decreases β in order to decrease the fairness in the system to a value closer to the desired target. Otherwise, when the filtered cell fairness index is not sufficiently high, β must be increased by the control loop, which forces the resource allocation to be fairer. This control is done indefinitely, which allows the convergence of $\Phi^{\mathrm{rt}}_{\mathrm{filt}}$, and consequently a strict control of the delay-based fairness in the system.

Fairness Analysis

In Fig. 4.17 the fairness results considering the EPA approach are shown. We can see that the two extremes are clear: FIFO provides the highest fairness, while PF is the worst. This confirms the preliminary analysis done for the classic policies: the largest the performance distance between the inner and outer groups, the least the fairness in the system.



(a) Example of convergence of the delay-based filtered cell fairness index for various step sizes η_{rt}

(b) Example of variation of the delay-based filtered cell fairness index and adaptation of the parameter β

Figure 4.16.: Convergence analysis of the ADF Policy

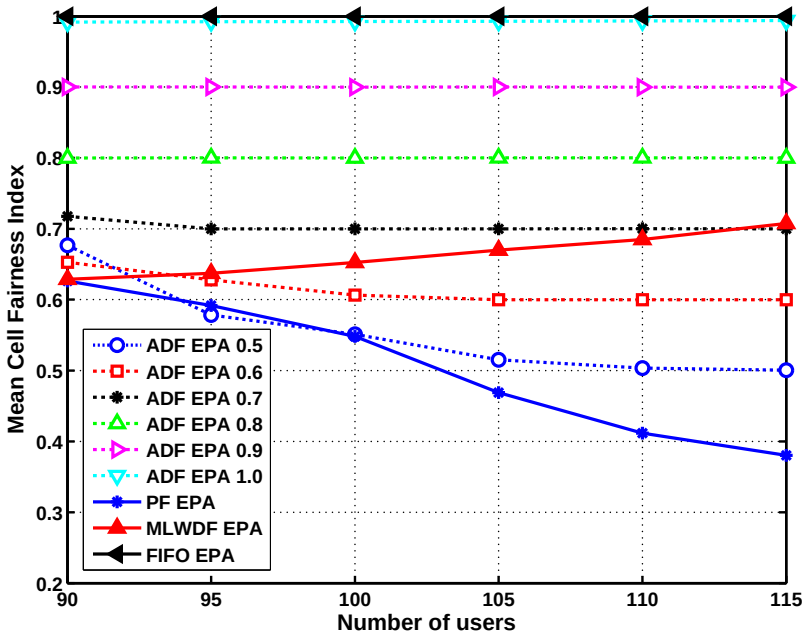


Figure 4.17.: Mean cell fairness index as a function of the number of users for the utility-based beta-rule framework

In particular, PF shows a monotonically behavior, where the fairness index decreases with the increase of the number of users. When there are more users in the system, PF is able to better explore the multi-user diversity and a more opportunistic allocation causes a fairness decrease. FIFO presents the highest fairness, which means that it is the best policy at equalizing the values of the HOL delays of the users. M-LWDF and ADF have an intermediate behavior between FIFO

and PF, in accordance with our expectations.

Notice that M-LWDF presents a fairness curve that slightly increases with the system load. This is due to the fact that the system is becoming congested for the range of loads considered in the simulations. Overload situations in a scenario with RT services is characterized by high delay-based fairness and poor QoS (compromise between fairness and QoS). This is the case of the FIFO policy that provides the maximum fairness at the expense of unacceptable delays. Since PF uses the resources more efficiently, it is able to avoid congestion for the range of loads considered. We have that M-LWDF is a static trade-off between FIFO and PF, so it presents worse QoS and higher delay-based fairness as the system becomes more congested. On the other hand, the ADF policy is very precise at controlling the fairness levels and efficient at preventing the system from being congested, as can be observed in Fig. 4.17. Due to the structure of the utility-based beta-rule framework and the limited range for the adaptation of the parameter β, the performance of the ADF policy is constrained by the performances of FIFO (maximum β) and PF (minimum β).

Table 4.9 shows that the utility-based APA algorithm has a little impact on the fairness results. Considering the particular case of ADF (first two lines of the Table), it can be concluded that the delay-based fairness control can be achieved using solely the corresponding utility-based DSA algorithm.

Table 4.9.: Mean cell fairness index as a function of the number of users for the utility-based beta-rule framework

RRA Policies	Number of users					
	J=90	**J=95**	**J=100**	**J=105**	**J=110**	**J=115**
ADF EPA 0.8	0.800	0.800	0.800	0.800	0.800	0.800
ADF Joint 0.8	0.803	0.800	0.800	0.800	0.800	0.800
PF EPA	0.626	0.592	0.548	0.469	0.412	0.380
PF Joint	0.652	0.609	0.566	0.503	0.467	0.407
M-LWDF EPA	0.629	0.637	0.652	0.670	0.685	0.708
M-LWDF Joint	0.639	0.641	0.652	0.651	0.684	0.708
FIFO EPA	1.000	1.000	1.000	1.000	1.000	1.000
FIFO Joint	1.000	1.000	1.000	1.000	1.000	1.000

Efficiency Analysis

In Fig. 4.18, the total cell throughput for the RRA policies considering equal power allocation is given. We use again the cell capacity as an indicator for the efficiency in the usage of the radio resources. As expected, the worst performance is presented by FIFO, because it only takes the delay information into account, which may lead to an inefficient resource allocation. The other policies, which give more importance to the channel quality and use the resources more efficiently, show similar behavior with higher system capacity. Notice that the ADF policy configured with a CFT equal to 1.0 presents a cell capacity much higher than FIFO, even though both are equivalent

in terms of delay-based fairness (see Fig. 4.17).

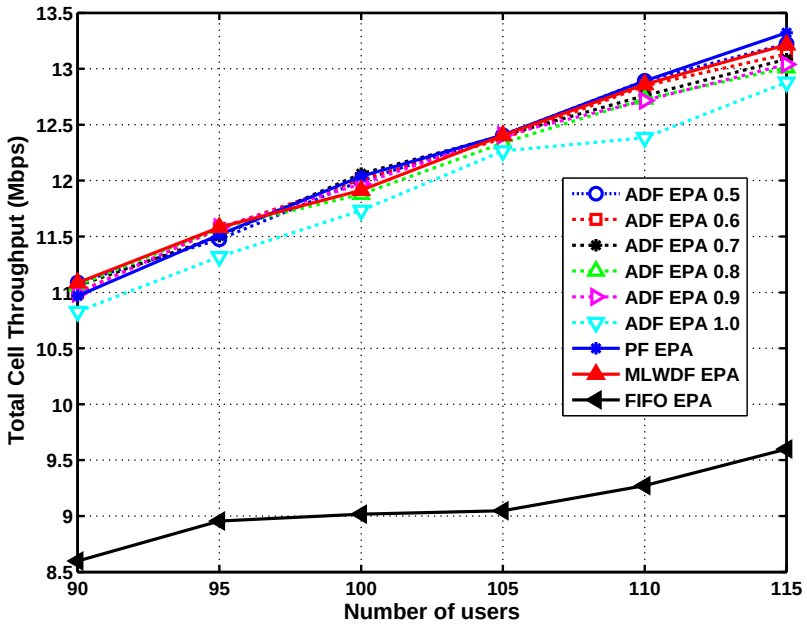


Figure 4.18.: Total cell throughput as a function of the number of users for the utility-based beta-rule framework

Notice in Table 4.10 that the performance of the joint and EPA approaches is also very similar in terms of total cell throughput, with a small advantage of the former over the latter.

Table 4.10.: Total cell throughput in Mbps as a function of the number of users for the utility-based beta-rule framework

RRA Policies	Number of users					
	J=90	J=95	J=100	J=105	J=110	J=115
ADF EPA 0.8	11.063	11.576	11.877	12.336	12.730	13.008
ADF Joint 0.8	11.231	11.714	12.114	12.559	13.007	13.404
PF EPA	10.968	11.516	12.038	12.408	12.892	13.322
PF Joint	11.015	11.518	12.007	12.439	12.952	13.318
M-LWDF EPA	11.085	11.581	11.914	12.404	12.860	13.215
M-LWDF Joint	11.174	11.675	12.110	12.566	12.910	13.327
FIFO EPA	8.598	8.955	9.017	9.047	9.273	9.599
FIFO Joint	8.728	8.833	9.087	9.504	9.638	9.577

Fig. 4.19 depicts the efficiency-fairness planes of the adaptive ADF policy in its both variants: joint and EPA. The efficiency indicator is the cell capacity and the fairness indicator is the delay-based CFI. The performance of the three classic policies, PF, M-LWDF and FIFO, is also shown (they are represented as single markers in the plane). A system load of 105 active RT users is assumed. As expected, the classic policies reveal the conflict between resource efficiency and user fairness based on delay. PF uses the radio resources more efficiently but does not present so high delay-based fairness values, while FIFO provides maximum fairness in the delay distribution but is very inefficient in the resource usage. M-LWDF presents a good static trade-off, with intermediate fairness and cell throughput as high as the one presented by PF.

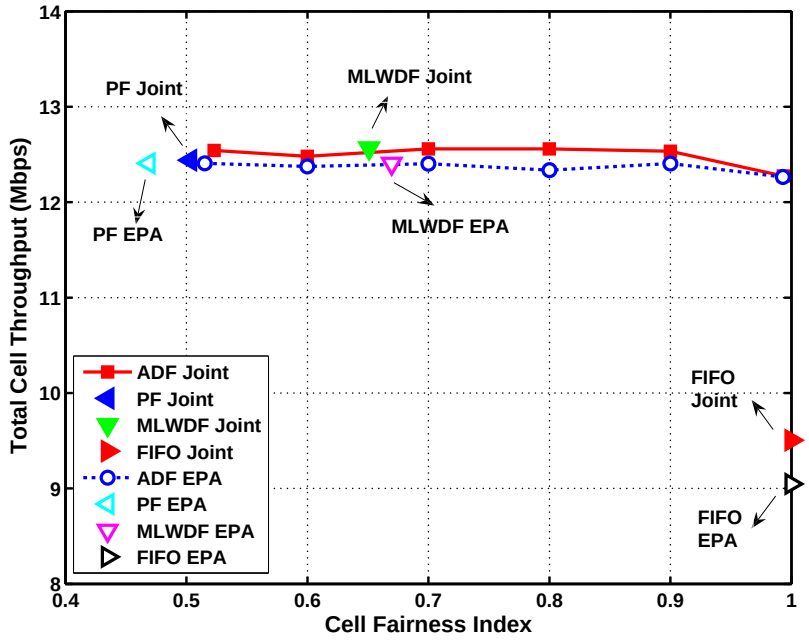


Figure 4.19.: Efficiency-Fairness plane for the utility-based beta-rule framework

On the other hand, for the same CFIs, the adaptive ADF policy provides equal or better cell throughput than the classic policies. Furthermore, the results demonstrate that it is able to meet successfully the cell fairness targets defined in the simulations (0.5, 0.6, 0.7, 0.8, 0.9 and 1.0). The fairness target of 0.5 is approximately met because of the structure of the utility-based beta-rule framework, which stipulates that the action and corresponding performance of ADF is lower bounded by PF. It is important to highlight that ADF is able to cover the whole range of fairness indexes while maintaining a cell capacity as high as PF, which is the policy that uses the resource more opportunistically within the beta-rule framework.

Similarly to the case of the ATF policy in Fig. 4.10, the ADF EPA policy shows a system capacity as high as the joint case. This indicates that a controllable trade-off can be properly achieved by applying the ADF policy only to the DSA algorithm and next applying equal power allocation among the sub-carriers.

Satisfaction Analysis

The satisfaction of the RT users, which is based on the percentage of packets discarded due to excessive delay, is presented in Fig. 4.20 for the RRA policies using the EPA approach. The QoS degradation provoked by the FIFO policy is clear. Although FIFO takes into account the delay in its allocation criterion, it is the one that presents the highest packet delays. This shows that the fact of not exploiting the OFDMA diversities is not beneficial in terms of QoS. Furthermore, when the system load increases, it causes the system to become stuck, i.e. the majority of the packets are discarded because they have a delay greater than 100 ms. All other RRA policies present similar performance regarding user satisfaction. Although the ADF policy with CFT equal to 1.0 shows a performance similar to FIFO in terms of delay-based fairness, it provides a much better satisfaction.

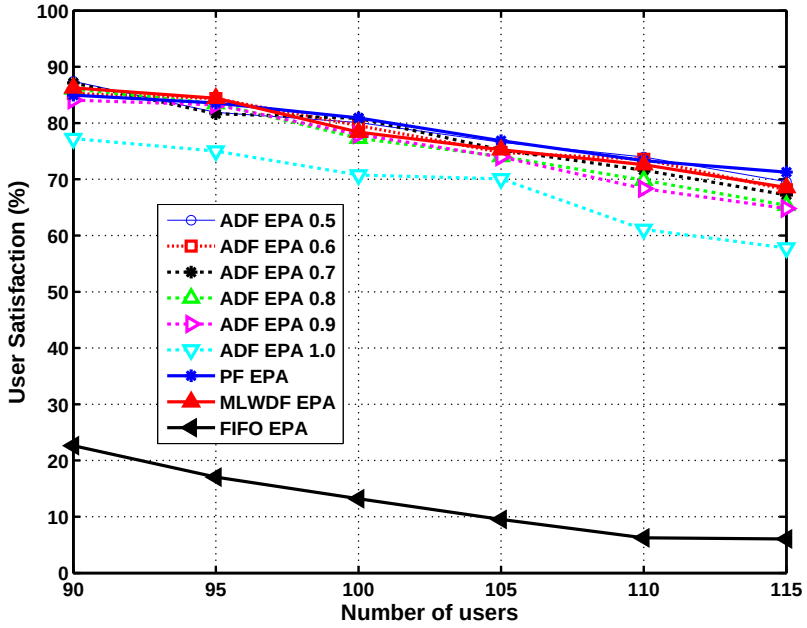


Figure 4.20.: User satisfaction as a function of the number of users for the utility-based beta-rule framework

In general, the joint approach presents slightly better satisfaction results than the EPA approach, as indicated by Table 4.11. This fact can also be observed in Fig. 4.21, which shows the satisfaction-fairness plane for the RRA policies envisaged by the beta-rule framework. This graphic assumes again a fixed system load of 105 active RT users. The adaptive ADF policy is able to guarantee a very good user satisfaction at the same level as M-LWDF and PF for the whole range of fairness indexes considered in the simulations.

Table 4.11.: Percentage of user satisfaction as a function of the number of users for the utility-based beta-rule framework

RRA Policies	Number of users					
	J=90	J=95	J=100	J=105	J=110	J=115
ADF EPA 0.8	86.037	83.474	77.300	74.064	69.849	65.362
ADF Joint 0.8	91.148	87.228	82.367	78.032	75.576	70.870
PF EPA	84.963	83.579	80.900	76.857	73.303	71.275
PF Joint	83.889	82.175	80.067	76.000	74.212	70.058
M-LWDF EPA	86.259	84.421	78.400	75.270	72.606	68.551
M-LWDF Joint	90.037	86.456	82.267	79.206	73.455	70.377
FIFO EPA	22.630	17.053	13.200	9.524	6.242	6.058
FIFO Joint	18.963	12.702	11.100	9.302	6.606	4.580

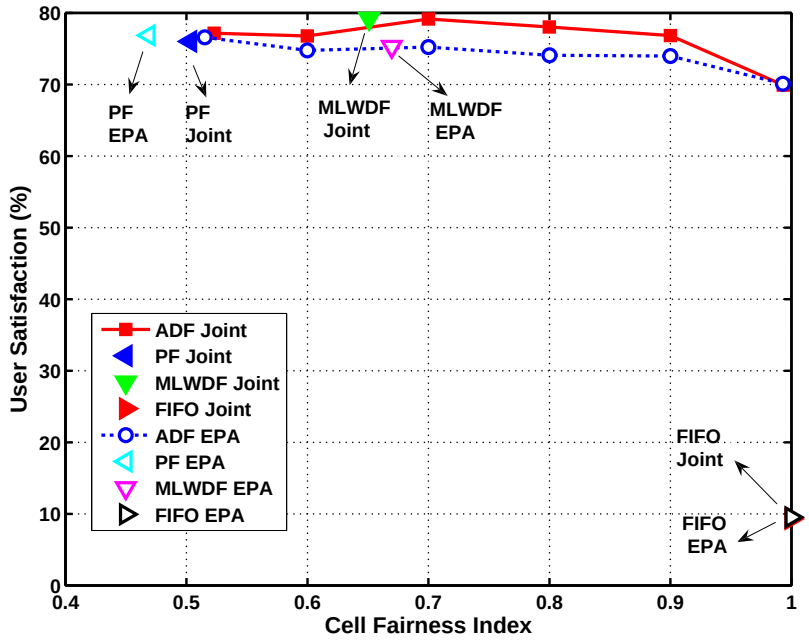


Figure 4.21.: Satisfaction-Fairness plane for the utility-based beta-rule framework

CPU Time Analysis

Fig. 4.22 compares the computational complexity of the joint and EPA approaches of the beta-rule framework by presenting the CPU time-fairness plane. It can be seen that the joint approach requires approximately 50% more computational power than the EPA approach, which does not compensate the slightly better results in terms of cell capacity and user satisfaction (see Figs. 4.19 and 4.21).

Furthermore, the ADF EPA policy demands approximately only 20% more processing time than the classic policies, which is a reasonable price to pay for the flexibility of the fairness control provided by its adaptive mechanism.

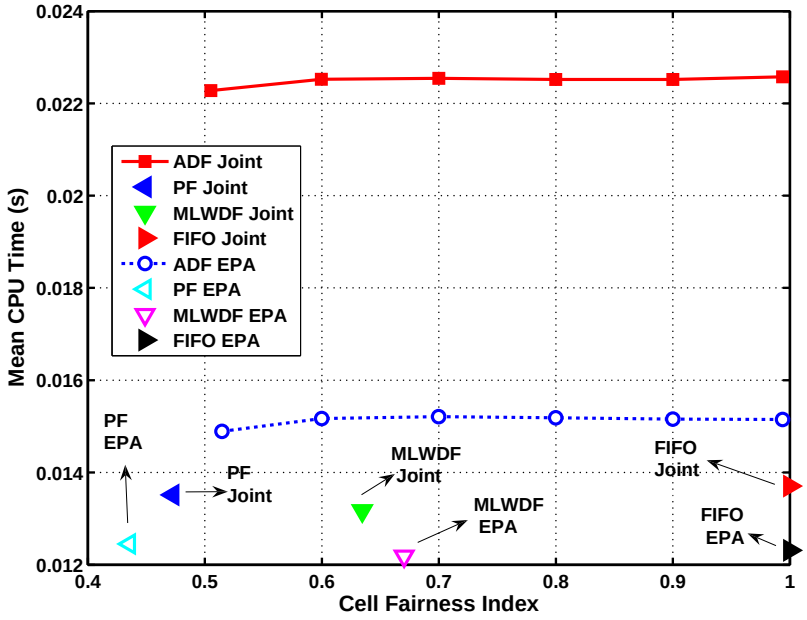


Figure 4.22.: CPU time-Fairness plane for the utility-based beta-rule framework

Opportunistic Allocation Analysis of the ADF Policy

This section assess how the fairness control provided by the ADF policy impacts in the way that opportunistic allocation is used. Fig. 4.23 shows the delay performance of the inner and outer groups when the ADF EPA policy is used with several CFTs.

One more time, it is evident that the shortest the difference between the performance of the groups, the highest the fairness. However, as the fairness increases (higher CFTs), notice that the QoS degradation of the inner group is not transformed in generalized QoS improvement for the outer group. It seems that only few users from the outer group are benefited. Thus, the consideration of the HOL delay in the criterion of the RRA policies must be carefully used in order to not achieve too degraded delay-based performance. It was shown in Figs. 4.19 and 4.21 that the ADF policy is robust in terms of cell capacity and user satisfaction for all CFTs considered in our study. However, as Fig. 4.23 suggests, the network operator must be careful when using CFTs very close to 1.0 due to the high delays involved.

4.5.3. Performance Comparison between Fairness/Rate Adaptive and Utility-Based Resource Allocation in a Scenario with NRT Services

In this thesis, we propose two ways to control the fairness in a macrocell network. The first approach is an instantaneous fairness control (short-term) based on fairness/rate adaptive techniques, which are described in chapter 3. The proposed fairness/rate adaptive RRA policies consider in their respective optimization problems a new fairness constraint based on the concept of a CFI. The

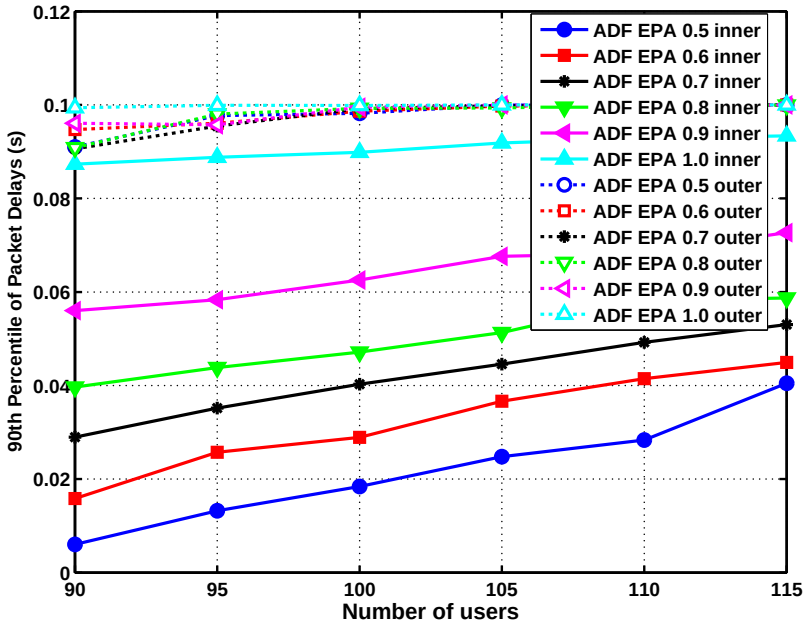


Figure 4.23.: 90th percentile of packet delays as a function of the number of users for the ADF policy considering inner and outer groups

network operator sets a desired CFT, and the proposed techniques assure that this fairness level is guaranteed every TTI by using iterative RRA algorithms that perform sub-carrier re-assignments and power re-allocations in order to decrease or increase the system fairness in accordance with the desired CFT.

The second approach is an average fairness control (mid-term) that uses utility-based RRA strategies, which are described in the present chapter. In particular, we propound two policies called ATF and ADF, which are derived from the alpha- and beta-rule frameworks, respectively (see sections 4.4.1 and 4.4.2). The alpha- and beta-rule frameworks consider samples of the throughput- and delay-based CFIs filtered over a time window, respectively, so an average fairness control is achieved.

The consideration of a time window in the resource allocation has some advantages. Without a time window, the optimization problem must guarantee fairness in each time slot period, which is a more complex task. However, when the time window is used, the fairness requirement is relaxed to a time-window length. This provides more flexibility to improve the spectral efficiency. Furthermore, the radio channel of mobile communications systems has a coherence time that characterizes the variation of the channel in the time domain. Because of that, it is likely that the CFI will not vary so fast from one TTI to another, and this slow-varying time series can be smoothed with a proper low-pass filtering. As a result, the corresponding resource allocations in consecutive time slots will also be smoothed.

According to the logical reasoning explained above, we can expect that the average fairness

control is better than the instantaneous control. However, it would be interesting to see a direct comparison between the proposed fairness/rate adaptive strategies and the ATF technique derived from the utility-based alpha-rule framework. Notice that they can be compared because both are suitable for NRT services. In order to make this comparison, we consider their variants using EPA. The simulation results presented in the following consider 16 active NRT users in the macrocell.

Fig. 4.24 compares the performance of the fairness/rate adaptive techniques[1] and the utility-based ATF technique by depicting the efficiency-fairness plane. Firstly, one can observe that all proposed adaptive RRA strategies are able to cover the whole path between the extreme points in the efficiency-fairness plane, drawing a complete efficiency-fairness curve, which is a meaningful advantage in comparison with the static classic RRA techniques. Looking at the efficiency-fairness regions (area below the curves), it can be concluded that FMMR-P is the worst technique, followed by FSRM-P, and finally FSRM and ATF. Note that the performance of FSRM and ATF are very similar, where the former performs better in the low fairness range, while the latter presents higher system capacity for the same CFIs in the high fairness range.

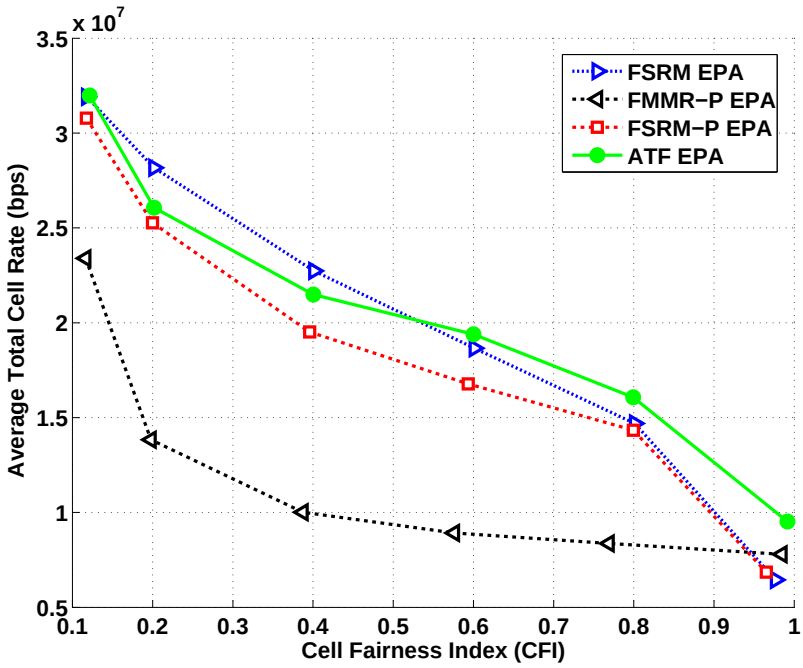


Figure 4.24.: Comparison between fairness/rate adaptive and utility-based RRA techniques for NRT services (efficiency-fairness plane)

It is also very important to make a user satisfaction analysis in order to see the impact of the fairness control in the compromise between fairness and QoS (see chapter 1 for a brief explanation about this compromise). The user satisfaction for various CFI values (satisfaction-fairness plane)

[1]The fairness/rate adaptive techniques proposed in this thesis are: Fairness-Based Sum Rate Maximization (FSRM), Fairness-Based Max-Min Rate with Proportional Rate Constraints (FMMR-P) and Fairness-Based Sum Rate Maximization with Proportional Rate Constraints (FSRM-P) (see chapter 3 for more details).

is presented in Fig. 4.25. Since the specific satisfaction results of the fairness/rate adaptive and utility-based techniques were already presented and discussed in chapter 3 and section 4.5.1, we focus here on the comparison of the ATF technique with respect to the other adaptive techniques. As can be observed in the range of low and intermediate fairness, FSRM-P provides the largest satisfaction-fairness region, followed by ATF. However, ATF outperforms all other techniques in the high fairness range, which makes it an interesting alternative also regarding user satisfaction.

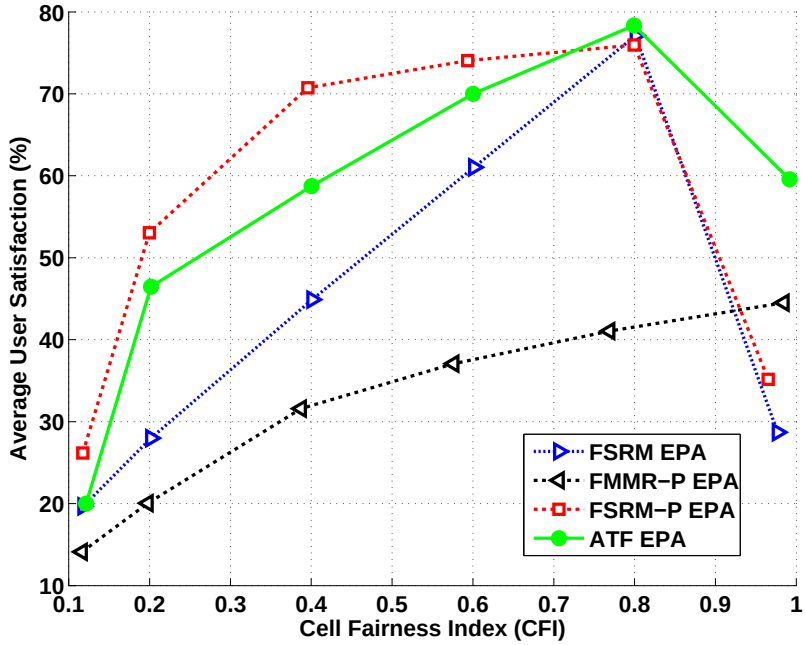


Figure 4.25.: Comparison between fairness/rate adaptive and utility-based RRA techniques for NRT services (satisfaction-fairness plane)

As commented before, it is logical to think that the instantaneous fairness control performed by the fairness/rate adaptive strategies is more complex than the control carried out by the utility-based ATF technique, which uses a feedback control loop based on filtered samples of the CFI in a time window. This supposition is demonstrated in Fig. 4.26, which shows the average CPU time needed to execute the respective techniques in a logarithmic scale against the CFI. As expected, among all considered RRA strategies, ATF requires the least processing time to perform the fairness control. Moreover, it uses approximately the same computational resources no matter the fairness range, which is not true for the fairness/rate adaptive techniques.

4.6. Conclusions

In this chapter, two utility-based RRA frameworks are proposed: alpha-rule and beta-rule. These frameworks are suitable for NRT and RT services, respectively. The former can be configured to work as any of three classic RRA policies, namely MR, PF and MMF, as well as an adaptive policy

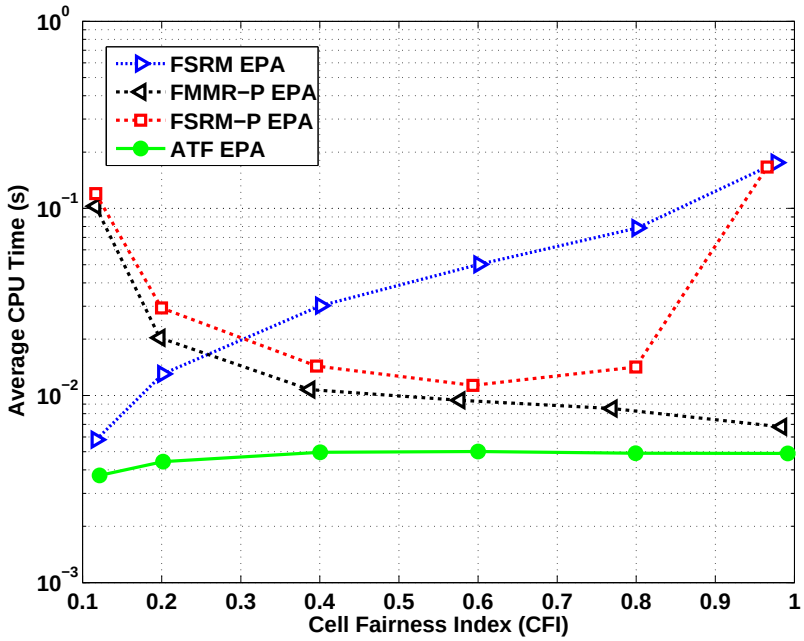


Figure 4.26.: Comparison between fairness/rate adaptive and utility-based RRA techniques for NRT services (CPU time-fairness plane)

called ATF. The latter comprises other three classic policies, which are PF, M-LWDF and FIFO, and also an adaptive policy called ADF. These policies rule the functioning of the DSA and APA algorithms, which are responsible for the assignment of sub-carriers and the allocation of power, respectively.

The utility-based alpha-rule and beta-rule frameworks are able to provide several degrees of fairness based on throughput and HOL delay, respectively. The adaptive ATF and ADF policies use a novel network management paradigm: control the fairness in the system by setting a fairness index target and letting the utility-based adaptive policies control the fairness dynamically. In this way, the network operator can manage the trade-off between resource efficiency and user fairness in cellular networks.

The main conclusions taken from the performance evaluation of the frameworks are summarized as follows:

- Opportunistic policies, such as MR for the case of NRT services or PF for the case of RT services, privileges users with good channel conditions (inner group) and does not give opportunities for the users in bad situation (outer group) to access the resources. On the other hand, policies that give much importance to the equality of QoS provision, such as MMF or FIFO, do not use the radio resources efficiently. Therefore, policies that are able to provide a trade-off between resource efficiency and user fairness are desired. PF for the case of NRT services and M-LWDF for the case of RT services are able to balance these two conflicting objectives, but the generated trade-offs are static. We generalized this concept and proposed

the adaptive ATF and ADF policies, which are able to provide dynamic trade-offs according to the interests of the network operator.

- The ATF and ADF policies dynamically adapt the parameters α and β of the alpha-rule and beta-rule frameworks, in order to achieve a desired fairness distribution in terms of throughput or HOL delay, respectively. The adaptation of the parameters α and β are governed by the step sizes η_{nrt} and η_{rt}, respectively, which control the convergence speed and the oscillation error of the filtered fairness index in the steady-state.

- The ATF and ADF policies are able to converge the cell fairness index to any target fairness index defined by the network operator. This fairness control is bounded by the structure of the alpha-rule and beta-rule frameworks, i.e. the minimum and maximum fairness performance depends on the allowed range of values for the parameters α and β. On one hand, in the alpha-rule framework, minimum and maximum α correspond to the MR and MMF policies, respectively. On the other hand, in the beta-rule framework, minimum and maximum β correspond to the PF and FIFO policies, respectively.

- ATF and ADF are able to provide equal or better cell capacity than the respective classic policies for the same cell fairness indexes. Furthermore, they are also able to provide dynamic trade-offs covering their respective efficiency-fairness planes. This is a remarkable strategic advantage to the network operators, because they can now control the aforementioned trade-off and decide in which point on the plane they want to operate.

- In terms of user satisfaction, the RRA policies that provide a trade-off between efficiency and fairness achieve equal or better results than the classic policies located in the extremes of the fairness index range.

- The joint approach, which uses utility-based resource allocation in both DSA and APA algorithms, provides slightly better results in terms of cell capacity and user satisfaction in comparison with the EPA approach, which uses utility-based DSA followed by equal power allocation. The small gain of the joint approach does not compensate the extra computational processing time demanded by the more complex power allocation step.

- The ATF and ADF policies control the fairness distribution in the system by re-allocating resources between the inner and outer groups. When the fairness in the system needs to be increased, resources are taken from the inner group and are given to the outer group. The outer group is more clearly benefited when ATF is used. The delay-based fairness implemented by ADF must be carefully managed when high fairness index targets are desired. In this situation, the packet delays of both groups approach the RT delay budget, and attention must be paid to not allow excessive packet losses due to high delays.

- The performance of the fairness/rate adaptive techniques proposed in chapter 3 was compared with the utility-based ATF technique propounded in the present chapter in a scenario with NRT services in a macrocell network. Taking into account important network metrics, such as system capacity, user fairness, user satisfaction and algorithm complexity, the ATF technique, which performs an average fairness control, is the best alternative for the management of the trade-off between resource efficiency and user fairness. Not only does it provide an efficiency-fairness region as large as FSRM, but it presents a satisfaction-fairness region as large as FSRM-P as well as. Moreover, these excellent performance results are achieved with the lowest computational burden among all considered RRA strategies.

Chapter 5

Interference Avoidance for Femtocell Networks

5.1. Introduction

Chapters 3 and 4 studied in detail and evaluated the fundamental trade-off between resource efficiency and user fairness in Orthogonal Frequency Division Multiple Access (OFDMA)-based macrocellular networks. In few words, the main reason for the appearance of this Radio Resource Management (RRM) problem is because the Mobile Terminals (MTs) have different Quality of Service (QoS) depending on their wireless channel conditions. For example, MTs that are close to the Macrocell Base Station (MBS) are more likely to perceive better channel quality than the ones far from the MBS. On one hand, if channel-based opportunistic Radio Resource Allocation (RRA) techniques exploit this multi-user diversity and allocate more resources to the best users, the system spectral efficiency will be maximized at the expense of unfairness among users. On the other hand, if RRA strategies that prioritize fairness are used, the QoS distribution among MTs will be more equalized, but as a consequence, capacity will be degraded because the performance will be flatted considering a worst-case scenario. In chapters 3 and 4 flexible RRA frameworks that could control this trade-off adaptively have been proposed.

The aforementioned trade-off is even more critical when there are indoor users in the macrocell coverage area. Due to indoor losses, these users could be located in coverage holes inside the macro network where the signal level is too low. One efficient way to diminish the appearance of the trade-off in the downlink of cellular networks is to shorten the distance between the transmitter (base station) and the receivers (users). Some possible solutions to accomplish this objective are: relay nodes, repeaters, Distributed Antenna Systems (DASs), radiating or leaky cable and picocells. Some of these solutions are specially suitable for the indoor coverage problem. For instance, when the transmitter and receiver antennas are close to each other, the signal quality is enhanced and both capacity and fairness can be increased at the same time. On one hand, capacity is increased

because higher data rates and higher area spectral efficiency can be achieved. On the other hand, the variance of the multi-user diversity is decreased with the proximity of the transmitter and receivers, which allows the majority of users to have good QoS and provides a fairness increase.

In the present chapter, we consider a technology that has the potential to approximate the network performance to the optimal operational point of the efficiency-fairness plane[1] characterized by high efficiency and high fairness: *femtocells*. Femtocell Access Points (FAPs) are small low-power cellular base stations that are deployed by the end-users on their home or office premisses. They were initially designed by the mobile operators to extend indoor coverage, aiming to solve the problem of coverage holes, improve overall system capacity, enhance signal quality, and offload data traffic from the MBS, allowing the mobile operator to be mainly focused on outdoor and mobile users. The main advantages of using femtocells instead of the other solutions commented before were presented in section 2.7 of chapter 2.

A massive deployment of FAPs is expected for the coming years. Therefore, it is of utmost importance to not only guarantee that this deployment will not have a negative impact on the macro network but also allow a seamless co-existence of neighboring FAPs in the femtocell tier. In that sense, interference management is one of the main RRM problems to be solved.

It is interesting to see that even though femtocells can help to solve the problem of the trade-off between resource efficiency and QoS-based fairness in cellular networks, the trade-off between system capacity and resource-based fairness appears when the radio resources must be distributed among several FAPs. This observation strengthens the idea that there is a fundamental trade-off between efficiency and fairness in several RRA problems in wireless networks.

We propose in this thesis three novel interference avoidance policies based on the concept of frequency partitioning for the femtocell tier, which are formulated as an RRA optimization problem and solved using Dynamic Frequency Planning (DFP) and Femtocell Sub-carrier Allocation (FSA) algorithms. These policies try to balance spectral efficiency and fairness in the resource distribution among FAPs.

The chapter is organized as follows. The main assumptions regarding the considered femtocell network are presented in section 5.2. The state-of-the-art revision about interference management in femtocell networks as well as the main contributions of this thesis regarding this subject are presented in section 5.3. In section 5.4, the characteristics of the specific scenario considered in this study are described. The novel interference avoidance policies proposed in this work are described in section 5.5, while the DFP and FSA algorithms are formulated in sections 5.6 and 5.7, respectively. In section 5.8, we assess how the proposed techniques can be implemented in a centralized or distributed network. Section 5.9 presents a discussion about the optimality of the DFP and FSA solutions, a performance comparison among the proposed interference avoidance policies, as well as an analysis about the required signaling overhead and the latency of the algorithms. Finally, the conclusions of the work are listed in section 5.10.

[1] A detailed explanation of the efficiency-fairness plane is given in section 3.2 of chapter 3 (see Fig. 3.1 on page 39).

5.2. Network Assumptions

Section 2.7 of chapter 2 describes the possible access modes that the femtocells can use: closed access, open access and hybrid access. It also presents the two main approaches for spectrum allocation between the macro- and the femtocell tiers: spectrum splitting and spectrum sharing.

In this thesis, we decided to tackle the interference avoidance problem only on the femtocell tier deployed on a closed access mode and using a dedicated spectrum. The reasons for that decision were already commented in section 2.7 of chapter 2 and are recapitulated in the following.

It is likely that the majority of the femtocell consumers will prefer the closed access mode. Since the FAPs are paid for by subscribers, they will not accept nonsubscribers as users of their own femtocells [71].

When closed access is considered, the presence of a femtocell may cause coverage holes or dead zones in the operator's macrocell network. For instance, consider a macrocell user located at the cell edge and close to a femtocell. Since it is not admitted to connect to the femtocell, it could be subject to strong interference if spectrum sharing is used, and may not be able to access the macrocell. Similarly, reception at the femtocell may be severely impacted by uplink transmissions from this macrocell user. Therefore, if the femtocells are configured for closed access operation, it is preferable to use a separate carrier, i.e. spectrum splitting between macro and femtocells, in order to maintain the overall performance of the radio access network.

Another reason to use the spectrum splitting approach is the industry prediction that by the year 2012, 150 million people around the world will be served by 70 million FAPs installed in their homes or offices [71]. It is of utmost importance that the mobile operators prevent their macrocell networks from being negatively impacted by such a large femtocell deployment. Many operators will choose the orthogonal frequency deployment (spectrum splitting) in order to avoid an uncontrollable QoS and capacity degradation of their macrocell networks.

Since a dedicated spectrum for the femtocells eliminates the cross-tier interference between the FAPs and the MBS, the main RRM problems left to be solved will reside in the femtocell tier. Moreover, it is expected that the majority of FAPs will be installed in a dense urban scenario. Such a scenario can be visualized as a crowded group of multi-flat apartments in a dense city center, which has a potential for high interference between neighboring femtocells. In this case, the femto-to-femto interference avoidance problem is the main challenge to be overtaken.

The ad hoc nature of the femtocell deployment demands the use of self-organization techniques [79]. Not only is self-organization essential for the FAPs to integrate into the macrocell network but also to co-exist with the neighboring FAPs without causing femto-to-femto interference. Moreover, each FAP must be able to automatically negotiate available resources with its neighbors and resolve interference conflicts.

5.3. Interference Management in Femtocell Networks

Due to the uncoordinated deployment of the FAPs, the mobile operator cannot apply classic network planning and optimization techniques to avoid interference, such as definition of the FAP location, fixed frequency reuse planning, etc. However, some kind of control over important network metrics such as spectral efficiency in the femtocell tier and fairness in the resource distribution among FAPs is desirable.

In that sense, this work presents a novel interference management approach for the femtocell tier. This thesis faces the femtocell RRM problem from a hierarchical perspective: first determine the resource allocation to the FAPs following a given policy, and next decide the resource assignment to the end-users. Therefore, instead of looking directly into the short-term resource assignment to the end-users, we propose to firstly prepare a mid/long-term action plan for the femtocell resource allocation:

1. Make a frequency planning for the FAPs following a given policy that is able to balance spectral efficiency and resource-based fairness in the femtocell tier. This is performed by the DFP algorithm, which decides how many frequency resources must be allocated to each FAP.

2. Allocate the specific frequency resources to the FAPs according to the frequency planning specified before and avoiding interference between neighboring FAPs. This is done by the FSA algorithm.

Once the frequency resources are allocated to the FAPs, each of them decides the resource assignment to the users. The algorithms that assign the sub-carriers to the users can be simplified, because the interference was already avoided by the frequency planning and it is expected that the users will have good channel conditions for the majority of the sub-carriers due to the proximity to the FAP antenna. Therefore, we focus our attention on the resource allocation to the FAPs and leave the resource assignment to the users out of the scope of this chapter. Furthermore, resource assignment to the users was already studied and evaluated in details in chapters 3 and 4.

Three novel interference avoidance policies are proposed in this work: Sum Frequency Allocation Maximization (SFAM), Max-Min Frequency Allocation (MMFA) and Layered Max-Min Frequency Allocation (LMMFA). These policies are formulated as an RRA optimization problem, which is solved by using two separate algorithms, DFP and FSA, such as it has been mentioned above. Firstly, the DFP algorithm decides how many frequency resources must be allocated to each FAP. Next, the FSA algorithm chooses the particular set of sub-carriers that each FAP must use.

It is not worthless to mention here that most of the interference avoidance techniques for femtocell networks based on sub-carrier allocation found in the literature try to solve the problem of cross-tier interference between FAPs and the macro users in a co-channel scenario (spectrum sharing). Furthermore, most of them tackle the resource allocation problem by considering the resource assignment to the end-users without making any high-level planning of the resource usage. This

algorithmic design approach usually requires that a huge amount of information regarding the channel conditions of all sub-carriers with respect to the macro and femto users is signalled in the network. Some examples are [71, 74–76, 80, 128–130].

Based on the arguments explained before, the present thesis focuses on the interference management only on the femtocell tier. Some works have investigated this specific problem [78, 79, 81]. Although the cross-tier interference management was the objective in [80], the authors formulated an algorithm that could also be applied in the femtocell tier. The objective was to minimize the overall network interference constrained by the interference restriction matrix and a required number of sub-channels per FAP. A simple randomized interference avoidance was proposed in [78], where each FAP accesses a random subset of the available frequency resources in order to avoid persistent collisions with neighboring FAPs. Two interference avoidance techniques were proposed in [79]: in the first one, the FAP prioritizes the usage of its sub-carriers based on a quality indicator, and the second technique aims to minimize the sum of the overall interference suffered by the users connected to the FAPs.

None of the techniques proposed in [78–80] are able to guarantee a complete interference avoidance. Furthermore, there is a lack of interference avoidance policies able to provide a high-level planning of the frequency resources to be used by the FAPs. An interesting example of such a high-level resource planning is presented in [81], where a component carrier selection algorithm is proposed for the 3GPP Long Term Evolution (LTE)-Advanced system. In this work, each femtocell selects a set of primary and secondary component carriers, which are slices of 20 MHz of spectrum, in such a way to avoid interference between neighboring femtocells. However, frequency planning in the granularity of sub-carriers is still desirable.

The main contributions of this thesis related to RRM for femtocell networks are:

1. ***Novel approach for interference management in the femtocell tier***: interference avoidance based on a high-level, mid/long-term frequency planning based on a granularity of sub-carriers.

2. ***Balance between efficiency and fairness***: the proposed interference avoidance policies are able to balance spectral efficiency and resource-based fairness in the femtocell tier.

3. ***Self-organizing RRA framework***: the proposed RRA techniques completely avoid interference and allow the seamless co-existence of several FAPs in any interference topology.

4. ***RRM paradigm based on the concept of clusters of FAPs***: the policies formulated in this thesis explore the interference topology of groups of neighboring FAPs (clusters) and are suitable for scenarios with any FAP densities, particularly dense urban scenarios.

5. ***Interference avoidance techniques that generate very low signaling overhead***: most of the works found in the literature propose RRA schemes that need a lot of information, such as the Channel State Information (CSI) of the users in all sub-carriers with respect to the

MBS and the FAPs. Our proposed techniques do not need channel information of the users and/or sub-carriers; they are based only on the cluster interference topology information.

6. ***Implementation in centralized or distributed network architectures***: the interference avoidance techniques proposed in this chapter can be implemented in a centralized (RRA broker and FAPs) or distributed (only FAPs) network approaches.

It is worth pointing out that an important advantage of the interference avoidance techniques proposed in this thesis is that the involved algorithms do not need to be executed so frequently. The resource allocation defined by the DFP and FSA algorithms for a given group of FAPs is valid until the interference pattern of that group changes. The interference topology varies only if the owner of a FAP moves it or a FAP is turned on or off. Since these changes are not frequent events, the resource allocation determined by the proposed techniques will remain valid in the mid/long term.

5.4. Scenario Characterization

Our object of study is a closed access femtocell tier that uses a separate spectrum band. Therefore, we consider only femto-to-femto (co-tier) interference. We assume that the FAPs that interfere mutually are organized in clusters, and there may exist several isolated clusters of FAPs, as illustrated in Fig. 5.1. A FAP can be aware that it belongs to a cluster using different strategies. Some options are [12, 131]:

1. The FAP has cognitive radio capabilities and sense its environment in order to find its neighbors;

2. Mobile terminals in the overlapping area of two or more FAPs inform the existence of mutual interference.

A detailed mechanism of how the FAPs know their neighbors is out of the scope of this thesis.

In this work, we are interested at studying the interference avoidance problem in each of these clusters of FAPs. Since there may exist a plurality of cluster topologies, the objective of this thesis is to propose general interference avoidance techniques that are valid for any cluster topology.

The interference avoidance techniques are formulated in a general way so that any type of frequency resources can be contemplated, for example sub-carriers in a general OFDMA system, Physical Resource Blocks (PRBs) in the 3GPP LTE system or chunks in the IEEE Worldwide Interoperability for Microwave Access (WiMAX) system. Without loss of generality, the terms frequency resource and sub-carrier will be used interchangeably in this chapter.

The interference avoidance techniques can be implemented either in a centralized or distributed network architecture. In the centralized approach, the processing entity that executes the DFP and FSA algorithms is located at an RRA broker, which is responsible for controlling several isolated

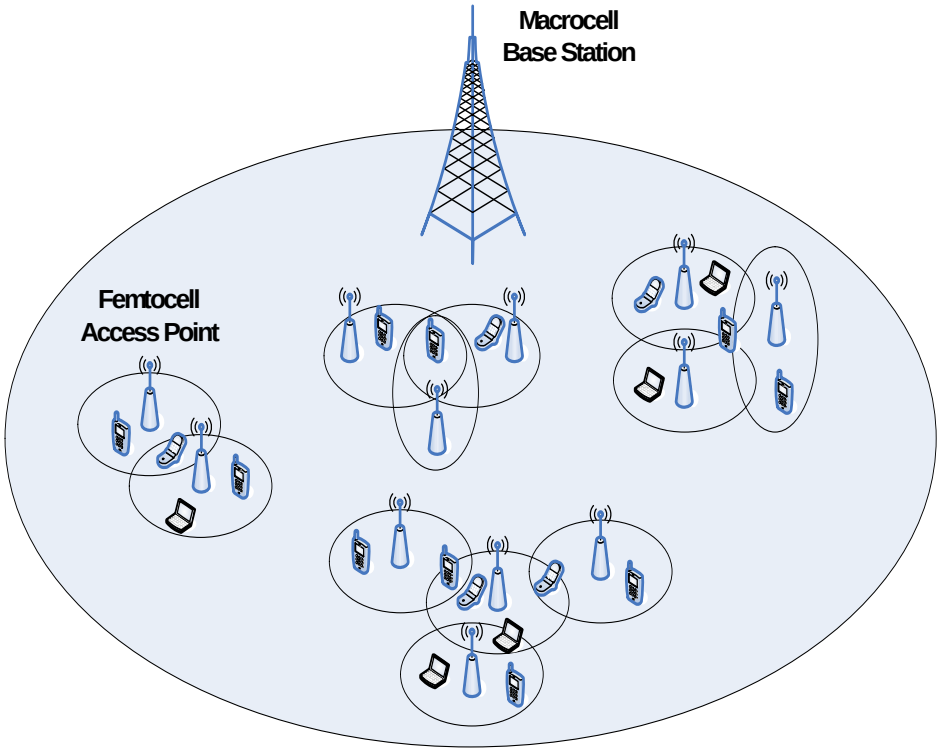


Figure 5.1.: Femtocell tier with several clusters of Femtocell Access Points (FAPs)

clusters of FAPs. According to the system architecture considered in this thesis, this RRA broker could be located in the FAP Gateway (see section 2.5 of chapter 2). It is assumed that the RRA broker communicates with the FAPs by means of a backhaul connection, such as Digital Subscriber Line (DSL), cable modem or optical fiber. The RRA broker receives the necessary information from the clusters, calculates the resource allocation and inform the decisions to the FAPs.

In a distributed approach, there are several processing entities that execute the DFP and FSA algorithms. Each of these entities is located in a different FAP. In this way, all clusters in the femtocell tier can calculate their resource allocation in a parallel and autonomous way. It is assumed that the FAPs that form a cluster communicate among them using the air interface to send broadcast messages and/or measurement reports [12]. In that sense, notice that a given FAP can only communicate directly with its interfering neighbor, since their coverage areas overlap, but it cannot communicate directly with a non-interfering FAP that belongs to the same cluster. A flooding procedure is a possible way to solve this problem. Some examples of naive and optimized flooding protocols for wireless mesh networks and ad hoc networks can be found in [132] and [133], respectively. Since these protocols were already well studied in the literature, they are out of the scope of this work.

5.5. Interference Avoidance Policies

In this section, we formulate the interference avoidance problem in an OFDMA-based femtocell tier as an integer optimization problem, as indicated below.

$$\max_{F_n} f(F_n), \tag{5.1}$$

$$\text{subject to } \bigcup_{n=1}^{N} \mathcal{S}_n = \mathcal{S}, \tag{5.2}$$

$$\mathcal{S}_i \bigcap \mathcal{S}_n = \varnothing, \quad i \nleftrightarrow n, \quad \forall i, n = 1:N, \tag{5.3}$$

$$\sum_{n \in G_m} F_n \leq K, \quad \forall m = 1:M, \tag{5.4}$$

$$F_n > 0, \quad \forall n = 1:N, \tag{5.5}$$

where N is the number of FAPs in a cluster; K is the total number of frequency resources (sub-carriers) dedicated to the femtocell tier; F_n is the number of frequency resources allocated to the nth FAP; $\mathcal{S}$ is the total set of sub-carriers dedicated to the femtocell tier; $\mathcal{S}_n$ is the subset of sub-carriers allocated to the nth FAP; M is the total number of possible groups of mutual interfering FAPs; G_m is the set of indexes of the FAPs belonging to a given group of mutual interfering FAPs; and $f(F_n)$ is the objective function that depends on F_n and must be maximized. The symbol $\nleftrightarrow$ indicates that two FAPs interfere with each other.

Constraint (5.2) states that all sub-carriers available in the bandwidth dedicated to the femtocell tier must be used. Perfect interference avoidance among different FAPs is assured by constraint (5.3), which guarantees that the sets of sub-carriers allocated to mutual interfering FAPs are disjoint. In a cluster of FAPs there may exist different groups G_m where their components interfere mutually inside the group. Constraint (5.4) requires that the FAPs inside each of these groups must not share the same sub-carriers and the sum of these frequency resources must not surpass the total bandwidth of the femtocell tier. A part or the total amount of the sub-carriers used by one group can be reused by other groups. Finally, the last constraint says that each FAP must be allocated at least one frequency resource.

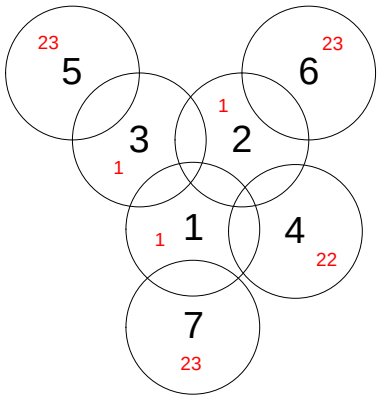
Based on the general formulation presented above, three different interference avoidance policies are proposed: SFAM, MMFA and LMMFA. Some characteristics and their respective objective functions are presented in sections 5.5.1, 5.5.2 and 5.5.3, respectively.

5.5.1. Sum Frequency Allocation Maximization (SFAM)

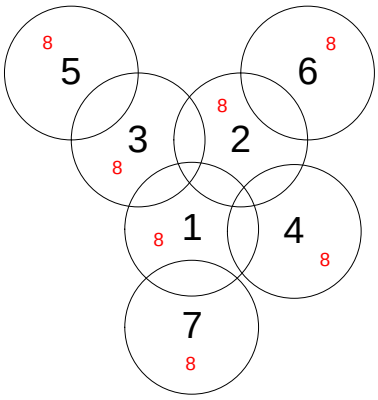
The objective of the SFAM policy is maximize the total number of allocated frequency resources in the femtocell tier, and consequently the network capacity. Its objective function is given by

$$f(F_n) = \sum_{n=1}^{N} F_n, \quad \forall n = 1:N. \tag{5.6}$$

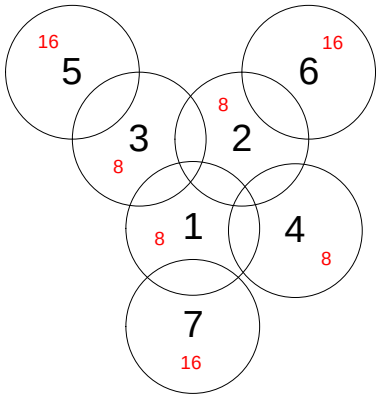
Fig. 5.2 depicts examples of frequency planning of different interference avoidance policies, including SFAM (see Fig. 5.2(a)). This example shows a cluster of N=7 FAPs, which have to share K=24 frequency resources. The big black number inside each circle is the FAP index while the small red number is the number of frequency resources allocated to that FAP (F_n). One can notice that there are a total of M=5 mutual interfering groups G_m, which are given by $G_1 = \{1, 2, 3\}$, $G_2 = \{1, 2, 4\}$, $G_3 = \{1, 7\}$, $G_4 = \{2, 6\}$ and $G_5 = \{3, 5\}$. It can be seen that the frequency planning of the SFAM policy shown in Fig. 5.2(a) is the one that maximizes the spectral efficiency in the femtocell tier at the expense of a low fairness in the resource distribution among FAPs. This is due to the fact that SFAM is an interference-based opportunistic policy, i.e. the FAPs that suffer less interference are allocated more resources while the most interfered ones receive the minimum allowed number of frequency resources. It can also be observed that the total bandwidth is limited to 24 frequency resources, so some specific sub-carriers must be reused in order to reach the number of 94 frequency resources indicated by the frequency planning.



(a) Sum Frequency Allocation Maximization (SFAM)



(b) Max Min Frequency Allocation (MMFA)



(c) Layered-Max Min Frequency Allocation (LMMFA)

Figure 5.2.: Examples of frequency planning of different interference avoidance policies - $N = 7$ **and** $K = 24$

5.5.2. Max-Min Frequency Allocation (MMFA)

Looking for a policy that could take fairness among FAPs into account, we propose the MMFA policy, whose objective function is

$$f(F_n) = \min_n F_n, \quad \forall n = 1:N. \tag{5.7}$$

According to (5.7), the MMFA policy maximizes the minimum number of frequency resources used by the FAPs in the cluster. Fig. 5.2(b) shows an example of the MMFA allocation where the minimum number of frequency resources used by any FAP in the cluster is determined by the groups of mutual interferers $G_1 = \{1,2,3\}$ and $G_2 = \{1,2,4\}$. Each of these groups is formed by three FAPs that interfere among them. According to an interference-free frequency planning, FAPs 1-4 must use 8 resources each. Even though FAPs 5-7 have less interfering neighbors, they still use 8 frequency resources in accordance with (5.7), indicating that MMFA is a non-opportunistic policy with regard to interference.

It can be clearly seen that the MMFA policy provides the maximum fairness in the frequency planning, since all FAPs will use the same number of frequency resources. This advantage comes at the expense of a lower total number of allocated frequency resources in the cluster, i.e. lower capacity. In the respective example shown in Fig. 5.2(b), this total number is 56.

5.5.3. Layered Max-Min Frequency Allocation (LMMFA)

The LMMFA policy aims to be a trade-off between SFAM and MMFA. To this end, the objective function of the LMMFA policy is defined as

$$f(F_n) = \prod_{l=1}^{L} \min_{n \in V_l} F_n, \quad \forall n = 1:N, \tag{5.8}$$

where L is the number of layers in the cluster of FAPs and V_l is the set of indexes of the FAPs that form the lth layer. A layer is defined as the group of FAPs that have the same maximum number of mutual interfering neighbors. For example, in the cluster presented in Fig. 5.2(c) we have 2 layers given by $V_1 = \{1,2,3,4\}$ and $V_2 = \{5,6,7\}$. The former is composed of FAPs that have two mutual interfering neighbors each, and the latter is formed by FAPs that have only one mutual interfering neighbor.

As a trade-off between the SFAM and MMFA policies, LMMFA is fair in the resource distribution among the FAPs that belong to the same layer and at the same time reuse the sub-carriers in an efficient manner so that the achieved total number of allocated frequency resources is high. It can be noticed in Fig. 5.2(c) that by using the LMMFA policy, the inner layer V_1 is allocated less resources than the outer layer V_2. This is due to the fact that the FAPs that belong to the inner layer suffer more interference from their neighbors, and so they have to use less resources in order to avoid co-channel interference. On the other hand, the FAPs of the outer layer have more freedom

to use more resources, reusing those ones that were forbidden for their neighbors of the inner layer. The frequency planning of the LMMFA policy is such that the product of the minimum number of allocated frequency resources on each layer is maximized, as indicated in (5.8). The product function was chosen as the objective function because it has the property of providing the highest fairness between layers. Considering a set of numbers whose sum is constant, the product between two close numbers is higher than the product of two distant numbers. The closer the minimum number of allocated resources on each layer, the higher the fairness among them.

5.6. Dynamic Frequency Planning

As said before, we propose to solve the interference avoidance problems presented in section 5.5 by using in conjunction DFP and FSA algorithms. The main objective of the DFP algorithm is to decide the number of frequency resources that must be allocated to each FAP. This section describes the DFP algorithms suitable for each of the interference avoidance policies shown in section 5.5. There is a common framework behind all the proposed DFP algorithms, which is shown in section 5.6.1. Two different mathematical/algorithimic tools are used to formulate the DFP algorithms: heuristics and Branch and Bound (BnB) technique. The BnB-based DFP algorithms are very computationally complex. Since they are able to always find the optimum solution, they are used as a reference for comparison with the sub-optimum heuristics-based DFP algorithms. The algorithms based on heuristics are described in section 5.6.2, while the ones based on the BnB technique are presented in section 5.6.3. A brief overview of the BnB technique is presented in section D.1 of appendix D.

5.6.1. General Framework

At this point of the DFP algorithm description, it is assumed that the processing entity that executes the DFP procedure has access to all information it needs. Further practical implementation issues in centralized and distributed architectures are discussed in section 5.8. Without loss of generality, we assume that the DFP allocation is decided for a single cluster of FAPs.

Before deciding the final DFP allocation, all DFP algorithms proposed in this work need to process some information. These pre-processing steps are presented in Fig. 5.3.

Firstly, the processing entity has to receive topology interference information from all FAPs in the cluster. With this information in hands, the processing entity creates an interference topology matrix $\mathbf{A}$, representing the interference pattern of the cluster. This is an N x N matrix, where N is the number of FAPs in the cluster, and each element a_{ij} of the matrix $\mathbf{A}$ indicates whether the ith FAP interferes with (is a neighbor of) the jth FAP or not.

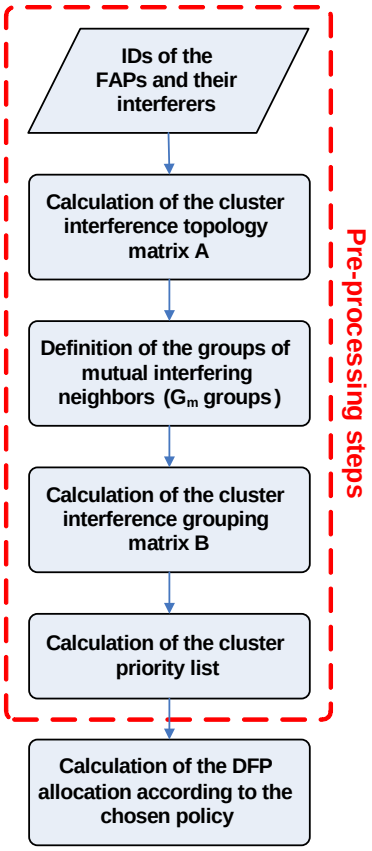


Figure 5.3.: Pre-processing steps of the Dynamic Frequency Planning (DFP) algorithms

In the example of the cluster shown in Fig. 5.2, such a matrix would be

$$\mathbf{A} = \begin{pmatrix} 0 & 1 & 1 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}. \tag{5.9}$$

Notice that the interference topology matrix is symmetric and the elements of the diagonal are all zeros. For example, looking at the first line of the matrix, one can see that FAP 1 interferes with FAPs 2, 3, 4 and 7.

Following that, the processing entity finds the groups G_m of mutual interfering FAPs, which have an important role on the constraints of the optimization problems of the interference avoidance policies. In the example shown in Fig. 5.2, one can notice that there are a total of M=5 mutual

interfering groups G_m, which are given by $G_1 = \{1, 2, 3\}$, $G_2 = \{1, 2, 4\}$, $G_3 = \{1, 7\}$, $G_4 = \{2, 6\}$ and $G_5 = \{3, 5\}$.

If the cluster size is small, the processing entity can easily make these computations. However, if the cluster is big, the computation burden cannot be neglected and an efficient algorithm is necessary. An efficient way to find the G_m groups using a dynamic search tree based on the BnB technique is described in section D.2 of appendix D.

The next step is to define the cluster interference grouping matrix $\mathbf{B}$ with dimension N x N, where each element b_{ij} indicates the number of G_m groups of size j that the ith FAP participates. If $b_{ij} = 0$, the ith FAP does not participate on any G_m group of size j.

Taking into account the cluster exemplified in Fig. 5.2, matrix $\mathbf{B}$ would be

$$\mathbf{B} = \begin{pmatrix} 0 & 1 & 2 & 0 & 0 & 0 & 0 \\ 0 & 1 & 2 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}. \tag{5.10}$$

Notice that the first line of matrix $\mathbf{B}$ indicates that FAP 1 belongs to 2 groups of size 3, $G_1 = \{1, 2, 3\}$ and $G_2 = \{1, 2, 4\}$, and one group of size 2, $G_3 = \{1, 7\}$.

Using matrices $\mathbf{A}$ and $\mathbf{B}$, the processing entity is able to calculate the cluster priority list in which the FAPs must be processed to define the frequency planning. The calculation of this priority list is a critical issue for the correctness of the proposed DFP algorithms. A reliable way to define the priority order suitable for our purposes is to look at the columns of matrix $\mathbf{B}$ from the right to the left. The FAPs that belong to larger groups are the ones that suffer more interference and so must be the first ones to define their frequency planning. If two or more FAPs have the same b_{ij} in the rightmost column, the casting vote will be given by the columns on the left. If two or more FAPs have a parity in all instances, the final decision is taken based on a FAP-specific random priority factor.

Once this pre-processing steps are done, the processing entity is able to calculate the DFP allocation according to the chosen interference avoidance policy, namely SFAM, MMFA or LMMFA.

5.6.2. Heuristic-Based DFP Algorithms

In this section, three DFP algorithms are formulated using heuristics. These three algorithms correspond to the interference avoidance policies investigated in this work.

The main idea of this algorithm is that the processing entity must test some possible DFP allo-

cations for each FAP and choose the best one according to the objective function of the respective interference avoidance policy. A possible DFP allocation is represented by a vector F with N elements, where N is the number of FAPs in the cluster. The order of the elements follow the cluster priority list, extracted from matrix $\mathbf{B}$. Assuming that n is the index of the FAP whose frequency allocation is being decided now, we have that the vector F is composed of the following elements:

- F_n^{prev}: vector containing the DFP allocations (number of allocated frequency resources) of the previous FAPs with priority higher than the nth FAP that already defined their allocations (predecessors of the nth FAP). This vector is fixed no matter the options for DFP allocation of the nth FAP.
- F_n: number that represents the DFP allocation of the nth FAP. Different options of F_n values are tested by the algorithm.
- F_n^{next}: vector containing bounded predictions of the DFP allocations of the next FAPs in the priority list (succeeders of the nth FAP). This vector changes depending on the value of F_n.

Fig. 5.4 depicts the block diagram and Algorithm 5.1 presents the detailed pseudo-code of the general framework of the heuristic-based DFP algorithms.

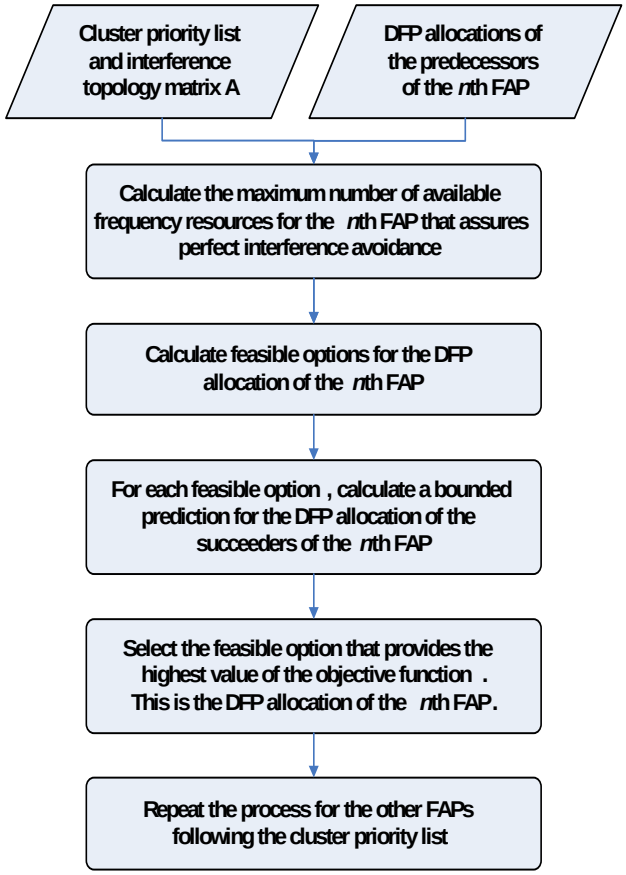


Figure 5.4.: Block diagram of the heuristic-based DFP algorithms

Algorithm 5.1 General heuristic-based Dynamic Frequency Planning (DFP) algorithm

Initialization

1: $\mathcal{Z} \leftarrow$ Cluster priority list // Set of indexes of all FAPs in order of priority
2: $\mathbf{A} \leftarrow$ Cluster interference topology matrix
3: $DFPSolution \leftarrow \varnothing$ // Vector of the solution of the DFP allocation for all FAPs

Options of DFP allocation for all FAPs

4: **for** $n = 1$ to N **do** // Loop of FAPs
5: $F_n \leftarrow 0$ // DFP solution for the nth FAP
6: $\mathcal{Z}_n^{\text{prev}} \leftarrow \mathcal{Z}(1 : n-1)$ // Set of indexes of all predecessors of the nth FAP in the priority list
7: $F_n^{\text{prev}} \leftarrow$ Solution of the DFP algorithm for the predecessors of the nth FAP in the priority list
8: $\mathcal{Z}_n^{\text{next}} \leftarrow \mathcal{Z}(n+1 : N)$ // Set of indexes of all succeeders of the nth FAP in the priority list
9: $F_n^{\text{next}} \leftarrow \varnothing$ // Vector of upper bounds of the DFP allocation for the succeeders of the nth FAP in the priority list
10: $F \leftarrow \varnothing$ // Concatenation of F_n^{prev}, F_n and F_n^{next}
11: $F_n^{\max} \leftarrow$ Maximum number of available frequency resources for the nth FAP that assures perfect interference avoidance // Depends on F_n^{prev} and $\mathbf{A}$
12: $\mathcal{O}_n \leftarrow$ Feasible options of DFP allocation for the nth FAP // These feasible options are based on heuristics that are dependent on the interference avoidance policy being considered
13: $Incumbent \leftarrow -\infty$ // Best value of the objective function for the nth FAP
14: $Solution \leftarrow 0$ // Best allocation for the nth FAP
15: **for** $i = 1$ to $\|\mathcal{O}_n\|$ **do** // $\|\cdot\|$ is the set cardinality operator
16: $F_n \leftarrow \mathcal{O}_n(i)$ // Every option of DFP allocation for the nth FAP will be tested as a candidate solution
17: $F \leftarrow F_n^{\text{prev}} \cup F_n$ // Concatenate the vectors
18: **for all** $j \in \mathcal{Z}_n^{\text{next}}$ **do**
19: $F_n^{\text{next}}(j) \leftarrow$ Bounded prediction of the DFP allocation of the succeeders of the nth FAP // This prediction is bounded by constraints imposed by $\mathcal{Z}$, F, $\mathbf{A}$ and the interference avoidance policy being used
20: **end for**
21: $F \leftarrow F \cup F_n^{\text{next}}$ // Concatenate the vectors
22: **if** $f(F) > Incumbent$ **then** // $f(F)$ is the value of the objective function for a candidate solution F and depends on the interference avoidance policy
23: $Incumbent \leftarrow f(F)$ // Save the current best value of the objective function
24: $Solution \leftarrow F_n$ // Save the current best option
25: **end if**
26: **end for**
27: $DFPSolution \leftarrow DFPSolution \cup Solution$ // Save the DFP allocation of the nth FAP
28: **end for**

In the initial step of the heuristic-based DFP algorithm (line 11), the nth FAP calculates how many frequency resources are available for its allocation. This calculation takes into account the number of resources already allocated to its predecessors (F_n^{prev}) and the interference constraints

imposed by the interference topology matrix $\mathbf{A}$.

Next on line 12, some feasible options of DFP allocation for the nth FAP are determined (different values of F_n). These feasible options are based on heuristics that are dependent on the interference avoidance policy under consideration.

For each of the different values of F_n being tested, we need to make bounded predictions about the DFP allocations of the succeeders of the nth FAP (lines 18-20). Each interference avoidance policy considered in this work requires a different way to make these predictions, as will be explained later. These bounded predictions are recorded on F_n^{next}, which together with F_n^{prev} and F_n, will compose the complete vector F. Finally, each vector F corresponding to a particular choice of F_n will be evaluated using an objective function $f(\cdot)$, which is dependent on the interference avoidance policy being used (lines 22-25). The value of F_n that maximizes the considered objective function is chosen as the DFP solution of the nth FAP. The DFP algorithms suitable for the SFAM, MMFA and LMMFA policies use the objective functions given by (5.6), (5.7) and (5.8), respectively.

In the following, particular aspects of the heuristic-based DFP algorithms adapted to the different interference avoidance policies are presented.

Heuristic-Based DFP for the SFAM Policy

The SFAM policy aims to maximize the number of frequency resources allocated in the cluster of FAPs while guaranteeing that each FAP will be allocated at least one frequency resource. Based on that, suitable options O_n and the bounded prediction vector F_n^{next} for the SFAM frequency allocation of the nth FAP are given by (5.11) and (5.12), respectively.

$$\mathcal{O}_n = \{1, F_n^{\max}\}, \tag{5.11}$$

$$F_n^{\text{next}}(j) = F_j^{\max}, \quad \forall j \in \mathcal{Z}_n^{\text{next}}, \tag{5.12}$$

where $F_n^{\max}$ and $F_j^{\max}$ are the maximum number of available frequency resources for the nth and jth FAPs that assure perfect interference avoidance, and $\mathcal{Z}_n^{\text{next}}$ is the set of indexes of all FAPs that will still execute the DFP algorithm after the nth FAP (in order of priority).

Heuristic-Based DFP for the MMFA Policy

According to our DFP formulation, the first FAP in the cluster priority list suffers the highest level of interference, since it always belongs to the G_m group that has the maximum number of elements among all M groups. Thus, in the heuristic-based DFP algorithm suitable for the MMFA policy, the first FAP has a special treatment because it is the one that directly determines the minimum number of frequency resources that are allocated to the FAPs in the cluster.

Having that into account, (5.13) and (5.14) show the suitable options O_n and the bounded prediction vector F_n^{next} for the MMFA frequency allocation of the nth FAP, respectively, depending

whether this FAP is the first in the priority list or not.

$$\mathcal{O}_n = F_n = \left\{ \begin{array}{c} \left\lfloor K/\max\left(\left\|G_m^{(n)}\right\|\right)\right\rfloor, \text{ if } n = 1,\ m = 1:M, \\ \min\left(F_n^{\text{prev}}\right), \quad \text{if } n \neq 1, \end{array} \right\}, \tag{5.13}$$

$$F_n^{\text{next}}(j) = \left\{ \begin{array}{c} F_n, \text{ if } n = 1,\ \forall j \in \mathcal{Z}_n^{\text{next}}, \\ \min\left(F_n^{\text{prev}}\right), \text{ if } n \neq 1,\ \forall j \in \mathcal{Z}_n^{\text{next}}. \end{array} \right\}, \tag{5.14}$$

where $\|\cdot\|$ is the cardinality operator, $\lfloor\cdot\rfloor$ is the floor operator, K is the total number of frequency resources in the femtocell tier, M is the total number of possible groups of mutual interfering FAPs, $G_m^{(n)}$ is the set of indexes of a given group of mutual interfering FAPs in which the nth FAP takes part, and F_n^{prev} is the vector of DFP allocations for the predecessors of the nth FAP. Notice that there is no need to test several options for the allocation variable F_n since $\mathcal{O}_n$ has only one element.

Heuristic-Based DFP for the LMMFA Policy

The DFP allocation options suitable for the LMMFA optimization problem are presented in (5.15), while the proposed expression for the bounded prediction vector F_n^{next} is given by (5.16).

$$\mathcal{O}_n = \left\{\left\lfloor K/\max\left(\left\|G_m^{(n)}\right\|\right)\right\rfloor, \ldots, F_n^{\max}\right\},\ m = 1:M, \tag{5.15}$$

$$F_n^{\text{next}}(j) = \min\left(\min\left(F_k\right), F_j^{\max}\right),\ \forall j \in \mathcal{Z}_n^{\text{next}},\ \forall k \in V_l^{(j)}. \tag{5.16}$$

Notice that the vector $\mathcal{O}_n$ given by (5.15) is limited by two extreme values with increments of one frequency resource between them. More options are needed for LMMFA because it has to deal with allocations for different layers of FAPs inside the cluster. The bounded prediction for the jth FAP in the $\mathcal{Z}_n^{\text{next}}$ set is the minimum between the following two values: 1) the minimum DFP allocation F_k among the FAPs that already defined their allocations and belong to the same layer $V_l^{(j)}$ of the jth FAP; and 2) the maximum number of available frequency resources for the jth FAP that assures perfect interference avoidance.

5.6.3. Branch and Bound-Based DFP Algorithms

The BnB technique [26–29] is a mathematical/algorithimic tool widely used to solve large scale combinatorial optimization problems, including NP-hard problems. Not only is the BnB an optimization tool but also an algorithm paradigm that must be adapted to the characteristics of the specific problem. In the scope of this thesis, we use the BnB technique to find the optimum DFP allocation for the maximization problems indicated in sections 5.5.1, 5.5.2 and 5.5.3. A brief overview of the BnB technique is presented in section D.1 of appendix D.

The BnB algorithm searches the complete space of solutions of a given problem for the optimum solution. The use of branching and bounding procedures enables the algorithm to search parts of the solution space only implicitly, and so accelerate the finding of the optimum solution.

In order to explain the branching procedure, we present Fig. 5.5, which illustrates the mapping of our DFP problem into the structure of a dynamic search tree. For didactic reasons, we consider a simple example characterized by a cluster of two interfering FAPs ($N = 2$) who must share four sub-carriers ($K = 4$).

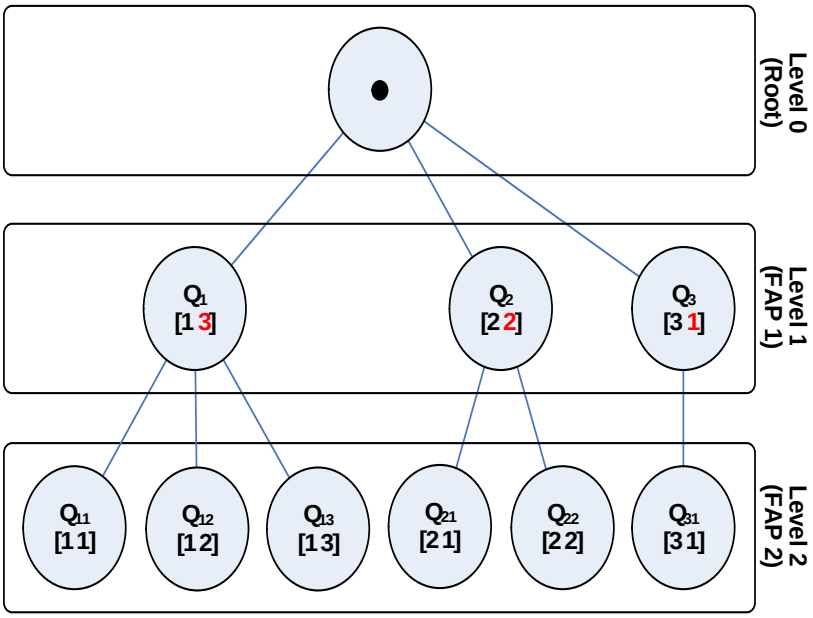


Figure 5.5.: Example of a dynamic search tree of the BnB-based DFP algorithm considering two interfering FAPs and four sub-carriers

As illustrated in Fig. 5.5, this search tree has three levels when completely built ($N + 1$ levels, where N is the number of FAPs). Level 0 only represents the root node, i.e. the complete solution space. The other levels correspond to each one of the N FAPs in the cluster in the order indicated by the cluster priority list. Each node of the tree is represented by a vector Q with N elements. In the case of an internal node located in the nth level (associated to the nth FAP), the vector Q corresponding to this node is defined as $Q = [q_1, \ldots, q_n, q_{n+1}, \ldots, q_N] = [x_1, \ldots, x_n, y_{n+1}, \ldots, y_N]$. Let us assume a vector $X_n \subseteq Q \mid X_n = [x_1, \ldots, x_n]$, which represents the possible frequency allocations of the predecessors of the nth FAP in the priority list, including itself (elements in black color in Fig. 5.5). Let us also define a vector $Y_n \subseteq Q \mid Y_n = [y_{n+1}, \ldots, y_N]$, which represents an upper bound for the possible frequency allocations associated to the succeeders of the nth FAP in the priority list (elements in red color in Fig. 5.5). As can be seen, the vector Q is the union of vectors X_n and Y_n. In the example shown in the figure, the nodes of the first level associated with FAP 1 are internal nodes represented by a vector $Q = [x_1, y_2]$.

The root father node is branched into several children nodes located at the first level by applying the constraints associated with FAP 1. This is equivalent to divide the complete solution space in several subspaces, where each subspace is represented by a different children node. FAP 1 can hypothetically choose any number of resources for its own allocation, from 1 to 3, since it has

the highest priority and is the first to define its allocation. This range of values is determined by the optimization constraints of our interference avoidance policy. In the exemplary case, FAP 1 is interfered by FAP 2, which must use at least one frequency resource. Therefore the maximum number of resources to be allocated to FAP 1 shall be 3 in this example. Each possible frequency allocation for FAP 1 corresponds to a different node in the first level of the tree, yielding 3 different nodes. Considering the first level of the tree, the vector X_1 of possible frequency allocations has only one element and there are 3 different vectors X_1, one for each of the nodes in the first level. The elements of the vector Y_1 associated to each node of the first level are upper bounds of the possible frequency allocations for FAP 2 taking into account the optimization constraints and the frequency allocation already fixed for FAP 1. That is why the vector Y of each node depends on the vector X defined for that specific node.

The same reasoning applies to the other levels. In the second level, vector X_2 of possible frequency allocations is composed of two elements: the first element is defined by the corresponding father node located in the first level, and the second element represents one possible frequency allocation for FAP 2 (branching procedure). Since in our example we consider only two FAPs, the second level is the last level of the tree and the nodes belonging to this level are terminal nodes (leaves). There are no vectors Y_N associated with terminal nodes because it is not necessary to compute upper bounds for the frequency allocation in this case. There are two reasons for that: 1) there are not succeeders of FAP N because it is the last in the priority list to be processed; and 2) all terminal nodes represent feasible solutions X_N since the frequency allocations represented by these nodes take into account the allocation and interference constraints of all FAPs in the cluster.

Fig. 5.5 shows a complete dynamic search tree for illustration purposes. In the execution of the BnB algorithm, it is not necessary to create all the branches of the tree. In order to reduce the search of the optimum solution, a bounding procedure is necessary to determine which branches of the search tree can possibly contain the optimum solution, discarding those branches that cannot contain it. The bounding procedure is explained in the following.

Let us define a bounding function $g(\cdot)$, which is an upper bound of the objective function $f(\cdot)$ such that $g(\cdot) \geq f(\cdot)$ over the region of feasible solutions. When a terminal node of the tree is processed for the first time, the feasible solution X_N represented by that node is saved as the current best solution, which is called incumbent. During the execution of the BnB algorithm, the nodes of the search tree are processed. For a generic node Q, a bounding procedure is applied by calculating $g(Q)$. This upper bounded value is compared with the incumbent. If the bounded value is lower than the incumbent, it means that the subspace associated to node Q cannot contain the optimum solution and so the subproblem is discarded (or fathomed). Otherwise, the node Q is saved in the pool of live subproblems for further branching. The search terminates when there is no unexplored parts of the solution space, and then the optimal solution is the one recorded as the incumbent.

Another important aspect of the BnB algorithm is the selection of the next node to be processed.

The main selection strategies are: *Breadth First*, *Best First* and *Depth First* [28]. More details about these search strategies can be found in section D.1 of appendix D.

Algorithm 5.2 sketches the pseudo-code of a general BnB-based DFP algorithm, where the three main components of the BnB technique can be observed: 1) a strategy for selecting the live subspace to be investigated in the current iteration (line 6); 2) a branching rule that subdivides a father subspace into two or more children subspaces to be investigated in the current iteration (line 8); and 3) a bounding procedure that provides for a given subspace an upper bound for the best solution obtainable in that subspace (lines 9-21).

Algorithm 5.2 General BnB-based Dynamic Frequency Planning (DFP) algorithm

Initialization

1: $Incumbent \leftarrow -\infty$ // Current best value of the objective function
2: $Solution \leftarrow 0$ // Current best solution
3: $UB(Q_0) \leftarrow g(Q_0)$ // Upper bound of the initial node Q_0
4: $Live \leftarrow \{Q_0, UB(Q_0)\}$ // Save the node Q_0 and its upper bound value on the set of alive nodes

Processing of the nodes in the search tree

5: **repeat**
6: Select the node Q from the set *Live* to be processed according to the *Best First* selection strategy [28] (see appendix D for more details)
7: $Live = Live \setminus \{Q, UB(Q)\}$ // Remove the father node Q from the set of alive nodes
8: Branch on the father node Q generating the children nodes $Q_1, \ldots, Q_k$
9: **for** $i = 1$ **to** k **do**
10: $UB(Q_i) \leftarrow g(Q_i)$ // Upper bound of the node Q_i
11: **if** $UB(Q_i) = f(X_N)$ **and** $f(X_N) > Incumbent$ **then** // $f(X_N)$ is the value of the objective function for a feasible solution X_N
12: $Incumbent \leftarrow f(X_N)$ // Save the current best value of the objective function
13: $Solution \leftarrow X_N$ // Save the current best solution
14: **continue** // Jump to the next iteration of the *for* loop
15: **end if**
16: **if** $UB(Q_i) \leq Incumbent$ **then**
17: Fathom Q_i // Discard the node Q_i
18: **else**
19: $Live \leftarrow Live \cup \{Q_i, UB(Q_i)\}$ // Save the node Q_i and its upper bound value on the set of alive nodes
20: **end if**
21: **end for**
22: **until** $Live = \varnothing$

The BnB-based DFP algorithm has to be adapted to the different interference avoidance policies presented in sections 5.5.1, 5.5.2 and 5.5.3. What differentiates the particular DFP algorithms are the objective and bounding functions. Therefore, the bounding functions given by (5.17), (5.18)

and (5.19) correspond to the SFAM, MMFA and LMMFA policies, respectively.

$$g(Q) = \sum_{i=1}^{N} q_i, \tag{5.17}$$

$$g(Q) = \min(q_i), \quad \forall i = 1:N, \tag{5.18}$$

$$g(Q) = \prod_{l=1}^{L} \min_{i \in V_l}(q_i), \quad \forall i = 1:N. \tag{5.19}$$

The objective functions $f(\cdot)$ of the BnB-based DFP algorithms suitable for the interference avoidance policies are equal to the bounding functions given by (5.17), (5.18) and (5.19). The difference is that the bounding function is used for any node Q of the search tree, while the objective function is used only for the feasible solutions X_N corresponding to the terminal nodes (leaves of the tree).

Now, let us summarize the main steps of the BnB technique applied to our DFP problem taking into account the example presented in Fig. 5.5. We show this example again in Figs. 5.6(a) and 5.6(b), which present the branching and bounding procedures considering two node selection strategies, *Breadth First* and *Best First*, respectively. In the former, the nodes are processed following the order of creation inside each level of the tree. In the latter, the order of processing is determined by the bounding function $g(\cdot)$. The nodes with highest bound values are processed first. In both cases, we consider the DFP algorithm suited to the MMFA policy, whose objective function is given by (5.7) (see page 140) and bounding function is given by (5.18).

According to Fig. 5.6(a), the BnB-based DFP algorithm using the *Breadth First* search strategy executes the following steps:

1. Branch the root node into the children nodes Q_1, Q_2 and Q_3.

2. Bound node Q_1, yielding $g(Q_1) = 1$. Since $g(Q_1)$ is higher than the current incumbent $(-\infty)$, the node Q_1 will remain alive. Do the same for nodes Q_2 and Q_3. Since $g(Q_2) = 2$ and $g(Q_3) = 1$ are higher than the current incumbent $(-\infty)$, nodes Q_2 and Q_3 will also remain alive.

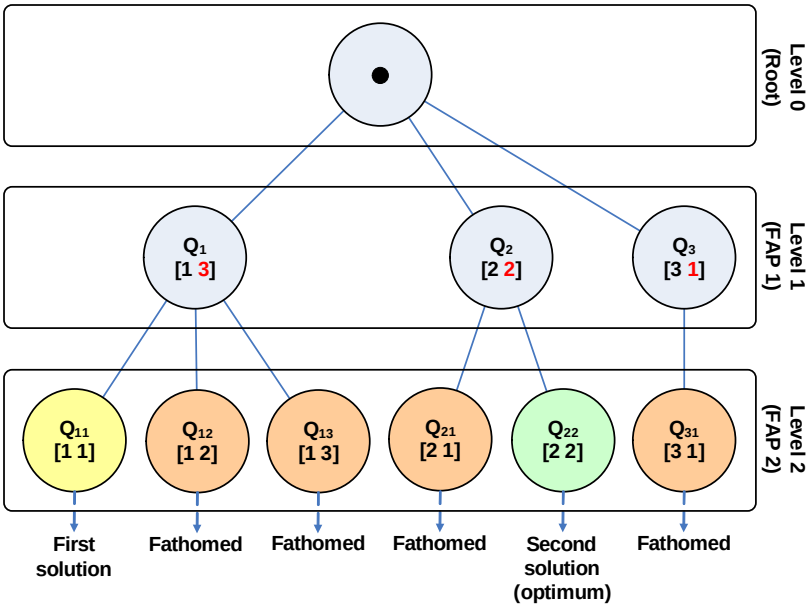
3. Select and branch the father node Q_1 into the children nodes Q_{11}, Q_{12} and Q_{13}.

4. Bound node Q_{11}, yielding $g(Q_{11}) = 1$. Since $g(Q_{11})$ is higher than the current incumbent $(-\infty)$ and Q_{11} is a terminal node (leaf), this node is saved as the current best solution and the current incumbent is set to $g(Q_{11}) = 1$.

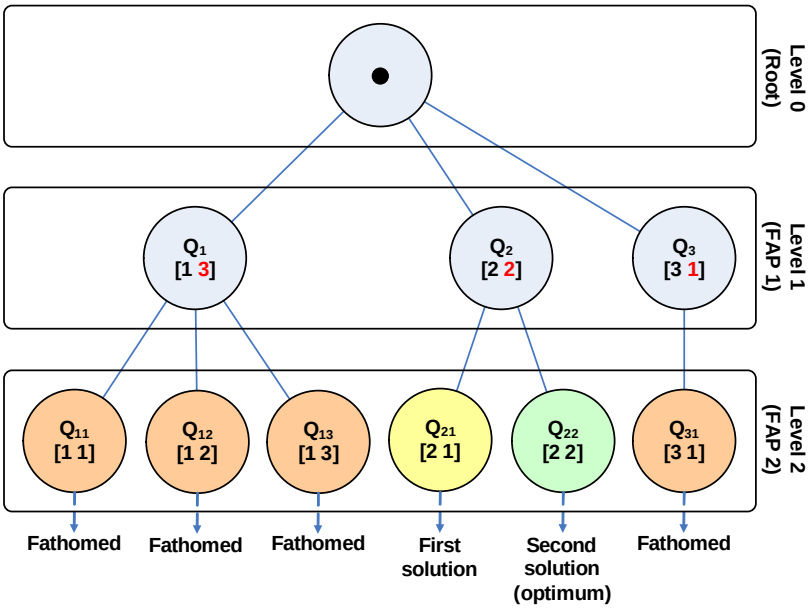
5. Bound node Q_{12}, yielding $g(Q_{12}) = 1$. Since $g(Q_{12})$ is lower or equal to the current incumbent (1), this node is discarded (fathomed). The same occurs for the node Q_{13}.

6. Select and branch the father node Q_2 into the children nodes Q_{21} and Q_{22}.

7. Bound node Q_{21}, yielding $g(Q_{21}) = 1$. Since $g(Q_{21})$ is lower or equal to the current incumbent (1), this node is discarded (fathomed).



(a) Breadth search



(b) Best search

Figure 5.6.: Examples of branching and bounding procedures on a dynamic search tree of a BnB-based DFP algorithm considering two interfering FAPs and four sub-carriers

8. Bound node Q_{22}, yielding $g(Q_{22}) = 2$. Since $g(Q_{22})$ is higher than the current incumbent (1) and Q_{22} is a terminal node (leaf), this node is saved as the current best solution and the current incumbent is set to $g(Q_{22}) = 2$.

9. Select and branch the father node Q_3 into the children node Q_{31}.

10. Bound node Q_{31}, yielding $g(Q_{31}) = 1$. Since $g(Q_{31})$ is lower or equal to the current incumbent (2), this node is discarded (fathomed).

11. The algorithm ends informing that the node $Q_{22} = [2\ 2]$ is the optimum solution for the MMFA DFP allocation, i.e. the one that maximizes the minimum number of allocated frequency resources in the cluster.

When the *Best First* search strategy is used, the steps executed by the BnB-based DFP algorithm are slightly different, as illustrated in Fig. 5.6(b). In the *Best First* strategy, the nodes with better bound values have priority in the selection process. In this way, node Q_2, with $g(Q_2) = 2$, is branched before nodes Q_1 and Q_3, which have a bound value equal to $g(Q_1) = g(Q_3) = 1$. In this way, node Q_{21}, with $g(Q_{21}) = 1$, turns out to be the first found solution of the DFP problem. Immediately after that, the algorithm bounds node Q_{22}, with $g(Q_{22}) = 2$, and finds out that its respective solution is better than the current one given by node Q_{21}. Therefore, node Q_{22} is saved as the current best solution and its bound value is recorded as the current incumbent. Notice that Q_1 and Q_3 are still alive nodes, and according to the implementation of the BnB technique (see lines 6-21 of algorithm 5.2), an alive node is first selected, after that it is branched, and the generated children nodes are bounded. In this way, after calculating the bound values of the children nodes Q_{11}, Q_{12}, Q_{13} and Q_{31}, they are compared with the current incumbent. Once it is verified that they have lower bounds, these children nodes are fathomed.

Notice that the BnB algorithm is able to find the same optimum solution, no matter which search strategy is used. However, the *Best First* strategy finds the optimum solution earlier than the *Breadth First* strategy, which can save a lot of computational time in large complex problems because more internal nodes will be discarded by the algorithm. In the simple example illustrated in Fig. 5.6, this advantage is not so clear. However, it becomes evident for more complex problems with many levels of the search tree. For that reason, the *Best First* search is the one chosen for the BnB-based DFP algorithm proposed in this thesis.

5.7. Femtocell Sub-carrier Allocation

The FSA algorithm has a fundamental importance in the solution of the interference avoidance problems proposed in section 5.5. The DFP algorithms described in section 5.6 perform the first part of the solution, defining the number of frequency resources that must be allocated to each FAP according to a chosen interference avoidance policy. The FSA algorithm does the second part of the job: chooses the particular set of sub-carriers that each FAP must use. This section proposes a single FSA algorithm that is suitable for all interference avoidance policies considered in this thesis.

The solutions calculated by the DFP algorithms can only be assured if a proper FSA algorithm that takes into account the interference topology of the femtocell tier is used. Notice that all interference avoidance policies require that the sets of sub-carriers allocated to the mutual interfering FAPs must be disjoint in order to accomplish perfect interference avoidance (see expression (5.3) on

page 138). It may there exist intricate cluster topologies, with many interrelated interference connections among FAPs. Therefore, a general FSA algorithm capable of avoiding frequency resources already used by previous interfering FAPs as well as minimizing the impact on the allocation of the subsequent FAPs is required. In that sense, it is proposed in this work the novel Max Frequency Reutilization (MFR) FSA algorithm, which maximizes the frequency reutilization in the femtocell tier. It is general and perfectly matches the frequency allocations of all DFP algorithms proposed in this thesis in a way that assures their feasibility. Fig. 5.7 depicts its general block diagram and Algorithm 5.3 sketches its detailed pseudo-code.

The algorithm operates on the sets of sub-carriers allocated to each FAP ($\mathcal{S}_i, i = 1 : N$). The set operations used in the FSA MFR algorithm are union ($\bigcup$), intersection ($\bigcap$) and relative complement ($\setminus$), which is defined as follows. Considering the sets of available sub-carriers $\mathcal{A}_i$ and $\mathcal{A}_j$ of the ith and jth FAPs, respectively, the set-theoretic difference (relative complement) of $\mathcal{A}_j$ and $\mathcal{A}_i$, is the set of elements in $\mathcal{A}_j$, but not in $\mathcal{A}_i$, i.e. $\mathcal{A}_j \setminus \mathcal{A}_i = \{a \in \mathcal{A}_j \mid a \notin \mathcal{A}_i\}$.

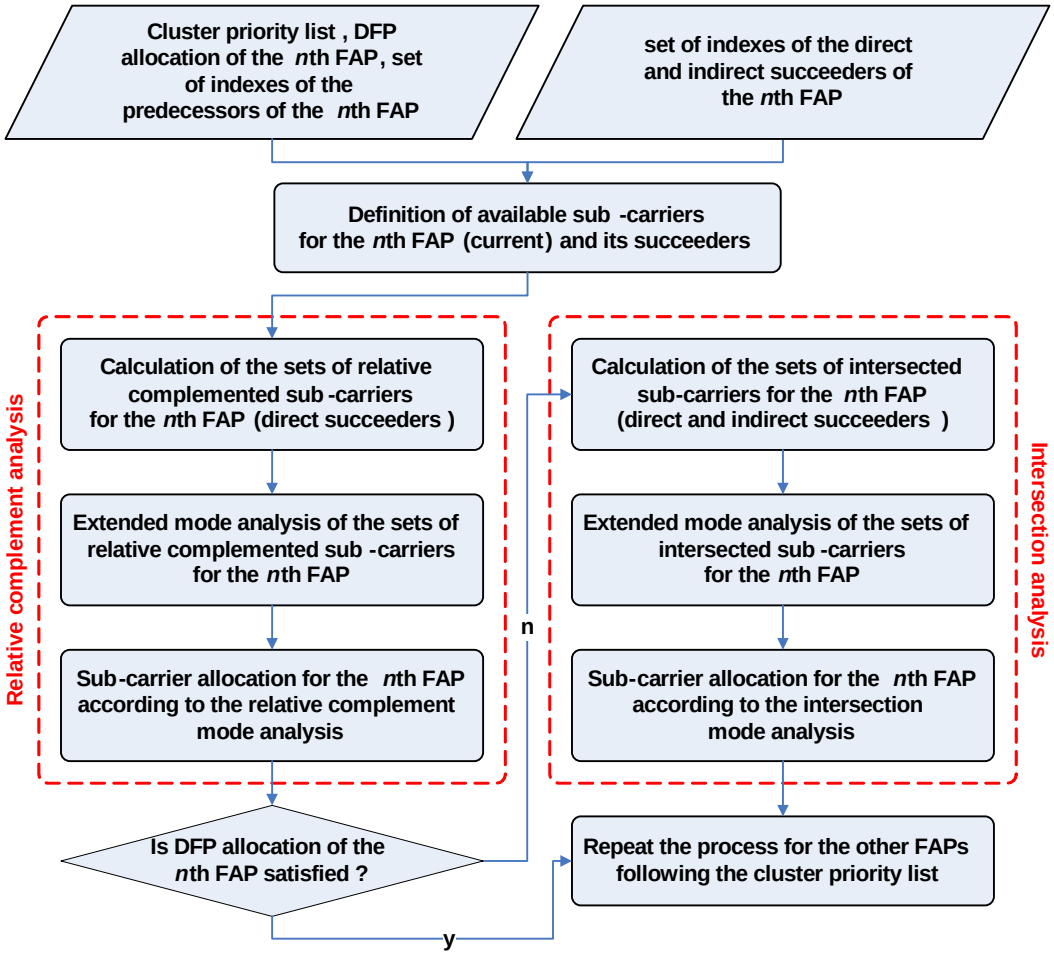


Figure 5.7.: Block diagram of the Max Frequency Reutilization (MFR) Femtocell Sub-carrier Allocation (FSA) algorithm

Algorithm 5.3 Max Frequency Reutilization (MFR) FSA algorithm

Initialization

$\mathcal{Z} \leftarrow$ Cluster priority list // Set of indexes of all FAPs in order of priority
$n \leftarrow$ index of the FAP being processed
$\mathcal{Z}_n^{\text{prev}} \leftarrow \mathcal{Z}(1:n-1)$ // Set of indexes of the predecessors of FAP n
$\mathcal{Z}_n^{\text{next}} \leftarrow \mathcal{Z}(n+1:N)$ // Set of indexes of the succeeders of FAP n
$\mathcal{Z}_{n,\text{direct}}^{\text{next}} \leftarrow \mathcal{Z}_n^{\text{next}}(i), \forall i \leftrightsquigarrow n$ // Succeeders of FAP n that interfere directly with it
$\mathcal{Z}_{n,\text{indirect}}^{\text{next}} \leftarrow \mathcal{Z}_n^{\text{next}}(i), \forall i \nleftrightsquigarrow n$ // Succeeders of FAP n that do not interfere directly with it
$F_n \leftarrow$ Number of frequency resources to be assigned to FAP n (defined by the DFP algorithm)
$\mathcal{S}_i \leftarrow \varnothing$ // Set of sub-carriers assigned to a generic FAP i
$\mathcal{S} \leftarrow$ Complete set of sub-carriers available in the femtocell tier
$\mathcal{A}_i \leftarrow \mathcal{S}$ // Available sub-carriers for a generic FAP i

Definition of available frequency resources

for all $i \in \mathcal{Z}$ **do**
  for all $j \in \mathcal{Z}_n^{\text{prev}}$ **do**
    $\mathcal{A}_i \leftarrow \mathcal{A}_i \setminus \mathcal{S}_j$, if $i \leftrightsquigarrow j$ // If FAPs i and j interfere with each other, remove the sub-carriers assigned to FAP j from the set of available sub-carriers of FAP i
  end for
end for

Relative complement analysis

for all $i \in \mathcal{Z}_{n,\text{direct}}^{\text{next}}$ **do** // Loop of direct succeeders
  $\mathcal{C}_i \leftarrow \mathcal{A}_n \setminus \mathcal{A}_i$ // Set of relative complemented sub-carriers of FAP i with respect to FAP n
end for
$\mathcal{C}_{\text{mode}}^{f} \leftarrow \text{Mo}(\mathcal{C}_i), i \in \mathcal{Z}_{n,\text{direct}}^{\text{next}}$ // Mode analysis of sets of relative complemented sub-carriers
for $f^* = \max(f)$ to $\min(f)$ **do** // f is the frequency of appearance of the elements in the sets
  $\mathcal{S}_n \leftarrow \mathcal{S}_n \cup \mathcal{C}_{\text{mode}}^{f^*}(1 : F_n - \|\mathcal{S}_n\|)$ // The selection of sub-carriers within this allowed set may follow different policies, such as random selection or max channel gain selection
end for

Intersection analysis

for all $i \in \mathcal{Z}_{n,\text{direct}}^{\text{next}}$ **do** // Loop of direct succeeders
  $\mathcal{I}_{n,i} \leftarrow \mathcal{A}_n \bigcap \mathcal{A}_i$, if $n \leftrightsquigarrow i$
  for all $j \in \mathcal{Z}_{n,\text{indirect}}^{\text{next}}$ **do** // Loop of indirect succeeders
    $\mathcal{I}_{n,i,j} \leftarrow \mathcal{I}_{n,i} \bigcap \mathcal{A}_j$, if $n \leftrightsquigarrow i$, and if $i \leftrightsquigarrow j$
  end for
end for
$\mathcal{I}_{\text{mode}}^{f} \leftarrow \text{Mo}(\mathcal{I}_{n,i}, \mathcal{I}_{n,i,j}), i \in \mathcal{Z}_{n,\text{direct}}^{\text{next}}, j \in \mathcal{Z}_{n,\text{indirect}}^{\text{next}}$ // Mode analysis of sets of intersected sub-carriers
for $f^* = \max(f)$ to $\min(f)$ **do**
  $\mathcal{S}_n \leftarrow \mathcal{S}_n \cup \mathcal{I}_{\text{mode}}^{f^*}(1 : F_n - \|\mathcal{S}_n\|)$ // See comment on line 21
end for

Assuming that n is the index of the FAP currently defining its FSA allocation, we define three sets of FAPs' indexes: 1) $\mathcal{Z}$, which is the set of indexes of all FAPs in order of priority (priority list explained in section 5.6.1); 2) $\mathcal{Z}_n^{\text{prev}} = \mathcal{Z}(1:n-1)$, which is the set of indexes of all FAPs with priority higher than the nth FAP, i.e. predecessors of the nth FAP (in order of priority); and 3) $\mathcal{Z}_n^{\text{next}} = \mathcal{Z}(n+1:N)$, which is the set of indexes of all FAPs that will still execute the FSA algorithm after the nth FAP, i.e. succeeders of the nth FAP (in order of priority). Moreover, $\mathcal{Z}_n^{\text{next}}$ is subdivided into two other groups: 1) $\mathcal{Z}_{n,\text{direct}}^{\text{next}} = \mathcal{Z}_n^{\text{next}}(i), \forall i \nleftrightarrow n$, which is composed of the indexes of the FAPs in $\mathcal{Z}_n^{\text{next}}$ that interfere directly with the nth FAP; and 2) $\mathcal{Z}_{n,\text{indirect}}^{\text{next}} = \mathcal{Z}_n^{\text{next}}(i), \forall i \nleftrightarrow n$, which comprises the indexes of the FAPs in $\mathcal{Z}_n^{\text{next}}$ that do not interfere directly with the nth FAP.

The FSA MFR algorithm must be executed for all FAPs following the cluster priority list. Considering the nth FAP, the algorithm follows the following steps: definition of available frequency resources (section 5.7.1) and sub-carrier allocation (section 5.7.2). The second step is further subdivided in two procedures: relative complement analysis and intersection analysis. Each procedure is explained in details in the following.

5.7.1. Definition of Available Frequency Resources

The first step is to determine the set of available sub-carriers for all FAPs in the cluster, taking into account the already fixed allocation of the predecessors of the nth FAP. Next following an iterative procedure and starting with the higher priority FAP, the sets of sub-carriers already assigned, $\mathcal{S}_j \mid j \in \mathcal{Z}_n^{\text{prev}}$, are removed from the pool of available resources for the FAPs that will still determine their assignments. Notice that only the sub-carriers belonging to a previous assigned FAP that mutually interferes with the current nth FAP are subtracted. Therefore, the set of sub-carriers $\mathcal{S}_n$ to be assigned to the nth FAP will be taken from its pool of available resources $\mathcal{A}_n$, which guarantees that the nth FAP will not interfere with its predecessors and a perfect interference avoidance will be achieved.

5.7.2. Sub-carrier Allocation Procedure

The FSA calculation is based on set operations and is composed of two sequential phases: relative complement analysis and intersection analysis.

Relative Complement Analysis

The main idea of the relative complement analysis is to allocate to the nth FAP the sub-carriers that will cause least interference to its direct succeeders (FAPs with lower priority that interfere directly with it). This is accomplished by performing relative complement operations between the sets of available sub-carriers of the nth FAP and its direct succeeders (lines 16-18), forming the sets $\mathcal{C}_i$'s. We join all sets $\mathcal{C}_i$'s and an extended mode[2] statistical analysis is done, saving all elements of the

[2]The mode is the value that occurs most frequently in a data set.

data set and their frequencies of appearance in a variable called $\mathcal{C}^{f}_{\text{mode}}$ (line 19). The elements that appear most frequently are those sub-carriers that the majority of the direct succeeders cannot use due to interference avoidance constraints. Nothing more natural than assigning these most frequent sub-carriers to the nth FAP, since they will not have a negative impact on the assignment of its direct succeeders. The order of preference of this resource assignment is given by the variable f, which is the frequency of appearance of the elements of the data set. The higher f, the fewer direct succeeders will be affected negatively by the sub-carrier assignment of the nth FAP.

Special attention must be given to the sub-carriers that appear only once in the data set. These are the sub-carriers that will not impact negatively on only one direct succeeder. It is wise to select these remaining sub-carriers from the corresponding sets $\mathcal{C}_i$'s in a way as fair as possible, in order to have an equal impact on the sub-carrier assignment of the other FAPs.

If the number of sub-carriers allowed by the relative complement analysis is higher than the required by the DFP algorithm (F_n), the nth FAP has the freedom to use a policy to determine which sub-carriers are the best for it. Examples of such policies are random selection or max channel gain-based selection.

Otherwise, if the number of sub-carriers allowed to the nth FAP by the relative complement analysis is lower than the required by the DFP algorithm, it is necessary to pass to the next phase, which is the intersection analysis.

Intersection analysis

After finishing the possibilities of the relative complement analysis, it is not possible to avoid any negative impact on the direct succeeders. Any assignment from now on will decrease the number of available sub-carriers for the next FAPs. However, there is still an efficient manner to allocate the remaining sub-carriers: the way that most benefit the indirect succeeders.

The indirect succeeders of the nth FAP are those ones that follow in the priority list and do not interfere directly with it. However, they interfere with the FAPs that are direct interferers of the nth FAP. In theory, the indirect succeeders can reuse all the sub-carriers assigned to the nth FAP. The main objective of the intersection analysis is to identify the sub-carriers that can be reused by the majority of the indirect succeeders, and so maximize the frequency reutilization in the femtocell tier.

The intersection analysis initiates with the formation of the sets of intersected sub-carriers between the nth FAP and its direct succeeders ($\mathcal{I}_{n,i} \mid i \in \mathcal{Z}^{\text{next}}_{n,\text{direct}}$), and between the nth FAP and its indirect succeeders ($\mathcal{I}_{n,i,j} \mid i \in \mathcal{Z}^{\text{next}}_{n,\text{direct}}, j \in \mathcal{Z}^{\text{next}}_{n,\text{indirect}}$), as can be seen in lines 23-28 of Algorithm 5.3. After that, an extended mode statistical analysis is done on the data formed by all $\mathcal{I}_{n,i}$ and $\mathcal{I}_{n,i,j}$ sets. The elements and their frequency of appearance are saved in the variable $\mathcal{I}^{f}_{\text{mode}}$ (see line 29). The elements that appear most frequently are those sub-carriers that the majority of the indirect succeeders could reuse. It is a good decision to allocate these most frequent sub-carriers to the nth FAP, since they will have a more positive impact on the assignment of its indirect succeeders. The

order of preference of this sub-carrier assignment is given by the variable f, which is the frequency of appearance of the elements of the data set. The higher f, the more indirect succeeders will be benefited by the sub-carrier assignment of the nth FAP.

Similarly to the case of the relative complement analysis, special attention must be given to the sub-carriers that appear only once in the data set of the intersection analysis. These are the sub-carriers that will benefit only one indirect succeeder. It is advised to select these remaining sub-carriers from the corresponding sets $\mathcal{I}_{n,i,j}$'s in a way as fair as possible, in order to have an equal impact on the sub-carrier assignment of the indirect succeeders.

It may happen that during the sub-carrier assignment based on the intersection analysis, the number of allowed sub-carriers is higher than the number of sub-carriers left to be assigned to the nth FAP. In this case, the nth FAP is again free to use a suitable policy to determine which sub-carriers are the best for it.

Notice that the last FAPs to define their allocations in a given layer of the cluster (lower priority FAPs) will not have the freedom to choose their sub-carriers, because their set of available sub-carriers is limited in order to avoid co-channel interference with their predecessors. However, we assume that this is not a critical issue for the final resource assignment to the users because the majority of the sub-carriers will have good channel conditions due to the proximity of the users and the FAP antennas. Even if the user observes a not so good channel condition due to shadowing fading, an appropriate power control on the FAP can compensate this fading effect.

5.7.3. Simple Example of the FSA MFR Algorithm

In order to better understand the FSA MFR algorithm, let us consider a simple example given by the cluster illustrated in Fig. 5.2(c) (see page 139). In this example, we have $N = 7$ FAPs in the cluster, $K = 24$ sub-carriers available, and we use the LMMFA interference avoidance policy. Notice that the FSA algorithm has to allocate the sub-carriers according to the amount of frequency resources determined by the LMMFA DFP algorithm (check Fig. 5.2(c) to see the frequency planning). For this specific scenario, the steps indicated below are taken in accordance with algorithm 5.3:

1. ***Sub-carrier allocation of FAP 1***: Since FAP 1 is the first to define its allocation, it can choose any sub-carriers from the pool. Let us assume that FAP 1 chooses $\mathcal{S}_1 = \{1, \ldots, 8\}$.

2. ***Sub-carrier allocation of FAP 2***: The group of predecessors of FAP 2 is $\mathcal{Z}_2^{\text{prev}} = \{1\}$. The sets of direct and indirect succeeders of FAP 2 are given by $\mathcal{Z}_{2,\text{direct}}^{\text{next}} = \{3, 4, 6\}$ and $\mathcal{Z}_{2,\text{indirect}}^{\text{next}} = \{5, 7\}$, respectively. Taking into account the sub-carriers already allocated to FAP 1, the step of definition of available sub-carriers yields: $\mathcal{A}_2 = \mathcal{A}_3 = \mathcal{A}_4 = \mathcal{A}_7 = \{9, \ldots, 24\}$, and $\mathcal{A}_5 = \mathcal{A}_6 = \{1, \ldots, 24\}$. According to the relative complement analysis, we have $C_i = \varnothing, \forall i \in \mathcal{Z}_{2,\text{direct}}^{\text{next}}$. In this case, the relative complement analysis is not helpful. Considering the intersection analysis, we have that $\mathcal{I}_{2,i} = \{9, \ldots, 24\}, \forall i \in \mathcal{Z}_{2,\text{direct}}^{\text{next}}$ and $\mathcal{I}_{2,i,j} = \{9, \ldots, 24\}, \forall i \in \mathcal{Z}_{2,\text{direct}}^{\text{next}}$ and $\forall j \in \mathcal{Z}_{2,\text{indirect}}^{\text{next}}$. Performing the extended mode analysis, we observe that the sub-

carriers that are most frequent in the sets of intersected sub-carriers are $\{9, \ldots, 24\}$. Therefore, FAP 2 can choose any 8 sub-carriers from this pool. Let us assume that $\mathcal{S}_2 = \{9, \ldots, 16\}$.

3. ***Sub-carrier allocation of FAP 3***: Regarding FAP 3, we have: $\mathcal{Z}_3^{\text{prev}} = \{1, 2\}$, $\mathcal{Z}_{3,\text{direct}}^{\text{next}} = \{5\}$, and $\mathcal{Z}_{3,\text{indirect}}^{\text{next}} = \{4, 6, 7\}$. Taking into account the allocations already defined for FAPs 1 and 2, we have that $\mathcal{A}_3 = \mathcal{A}_4 = \{17, \ldots, 24\}$, $\mathcal{A}_5 = \{1, \ldots, 24\}$, $\mathcal{A}_6 = \{1, \ldots, 8, 17, \ldots, 24\}$ and $\mathcal{A}_7 = \{9, \ldots, 24\}$. FAP 3 has only one direct succeeder (FAP 5), and its set of relative complemented sub-carriers is $C_5 = \varnothing$. In this case, we have to go to the phase of intersection analysis. In this phase, we have $\mathcal{I}_{3,5} = \{17, \ldots, 24\}$, and $\mathcal{I}_{3,5,4} = \mathcal{I}_{3,5,6} = \mathcal{I}_{3,5,7} = \{17, \ldots, 24\}$. Performing the extended mode analysis of the intersected sub-carriers shows that the most frequent sub-carriers, and therefore the allocation of FAP 3, is given by $\mathcal{S}_3 = \{17, \ldots, 24\}$.

4. ***Sub-carrier allocation of FAP 4***: For the case of FAP 4, we have that $\mathcal{Z}_4^{\text{prev}} = \{1, 2, 3\}$, $\mathcal{Z}_{4,\text{direct}}^{\text{next}} = \varnothing$, and $\mathcal{Z}_{4,\text{indirect}}^{\text{next}} = \{5, 6, 7\}$. Considering the allocations of FAPs 1, 2 and 3 yields $\mathcal{A}_4 = \{17, \ldots, 24\}$, $\mathcal{A}_5 = \{1, \ldots, 16\}$, $\mathcal{A}_6 = \{1, \ldots, 8, 17, \ldots, 24\}$ and $\mathcal{A}_7 = \{9, \ldots, 24\}$. Since there are not direct succeeders of FAP 4, i.e. it s the last in its group of mutual interfering FAPs to define the sub-carrier allocation, the relative complement and intersection analyses are not needed. The allocation solution is choose directly from the pool of available resources the amount stipulated by the DFP algorithm, i.e. 8 sub-carriers. Therefore, $\mathcal{S}_4 = \mathcal{A}_4 = \{17, \ldots, 24\}$. Notice that as a natural consequence of the FSA MFR algorithm, FAP 4 reuses the same sub-carriers already allocated to FAP 3. This maximizes the frequency reutilization in the cluster at the same time that perfect interference avoidance is achieved.

5. ***Sub-carrier allocations of FAPs 5, 6 and 7***: Such as the case of FAP 4, we have that FAPs 5, 6 and 7 do not have direct succeeders. In this situation, their allocations are readily defined from their respective pools of available sub-carriers. Their final allocations are: $\mathcal{S}_5 = \mathcal{A}_5 = \{1, \ldots, 16\}$, $\mathcal{S}_6 = \mathcal{A}_6 = \{1, \ldots, 8, 17, \ldots, 24\}$ and $\mathcal{S}_7 = \mathcal{A}_7 = \{9, \ldots, 24\}$.

Notice that the FSA MFR algorithm was able to exactly follow the frequency planning determined by the DFP algorithm (see Fig. 5.2(c)), maximizing the frequency reutilization and completely avoiding the interference between neighbor FAPs.

5.8. Network Architectural Issues

In the description of the DFP and FSA algorithms presented in sections 5.6 and 5.7, respectively, we assumed that the processing entity responsible for executing these algorithms had all the information it needed. However, where the processing entity is located and how the network elements communicate with each other depends directly on the network architecture. In this work, two processing and communication approaches are envisaged: centralized and distributed. Some assumptions regarding these approaches were already explained in section 5.4. The centralized and distributed approaches are described in sections 5.8.1 and 5.8.2, respectively.

5.8.1. Centralized Approach

Fig. 5.8 depicts a basic processing and communication protocol considering a centralized network architecture. In the centralized approach, the processing entity is located at an RRA broker. For didactic reasons, this figure shows only two FAPs that are assumed to belong to the same cluster. The DFP and FSA algorithms are meant to determine the frequency allocation and sub-carrier assignment for a single cluster of FAPs. If the RRA broker is responsible for more than one cluster, it has to find the DFP and FSA solutions for all clusters in a sequential way.

The execution and communication flow of the DFP and FSA algorithms in a centralized approach is described below:

1. Firstly, measurement reports are used by the FAPs to sense the environment [12]. Then the interference topology information is sent from the FAPs to the RRA broker. This information consists of their own Identification Numbers (IDs) and the IDs of their direct interferers (steps a and b in the figure).

2. The RRA broker executes the pre-processing steps of the DFP algorithm (steps c-f). For more details about the pre-processing steps, see Fig. 5.3 and section 5.6.1.

3. The RRA broker runs the DFP algorithm in accordance with the chosen interference avoidance policy, i.e. SFAM, MMFA or LMMFA (see section 5.5 for more details). The DFP algorithm can be based on heuristics or the BnB technique, as described in sections 5.6.2 and 5.6.3, respectively (step g).

4. The RRA broker finds the FSA solution according to the algorithm described in section 5.7 (step h).

5. Finally, the RRA broker sends the DFP and FSA decisions to the FAPs (steps i and j in the figure).

5.8.2. Distributed Approach

A basic processing and communication protocol assuming a distributed network approach is illustrated in Fig. 5.9. We assume that the three FAPs presented in this figure belong to the same cluster. Each FAP executes the DFP and FSA algorithms independently. In the following, a list of steps representing the execution and communication flow of the DFP and FSA algorithms in a distributed approach is described.

1. Firstly, the FAPs exchange local interference topology information between them using a flooding protocol (see comments on section 5.4). In this way, after some messages, every FAP in the cluster can have interference information about all other FAPs in that cluster (step a in the figure).

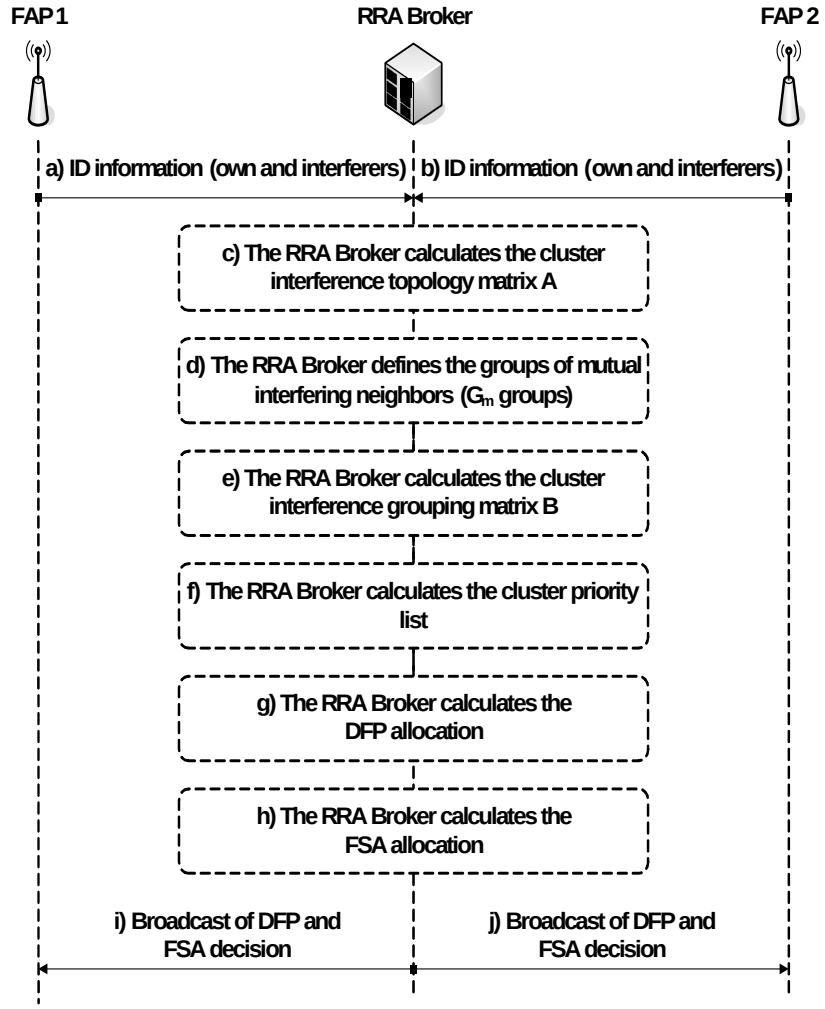


Figure 5.8.: Basic processing and communication protocol of the interference avoidance techniques in a centralized architecture

2. With this information in hands, each FAP can calculate the cluster interference topology matrix **A** and its own groups of mutual interfering neighbors (G_m groups) (steps b an c). Next, all FAPs in the cluster communicate with each other by means of flooding to exchange information about their G_m groups and their own random priority factors (step d). The G_m groups of all FAPs are necessary to calculate the cluster interference grouping matrix **B** (step e). If matrix **B** is ambiguous, i.e. it is not possible to extract a unique priority list from it, the FAPs' random priority factors will be used for the casting votes. Since every FAP is able to calculate the same matrix **B**, and knows the priority factors of the other FAPs, each FAP is also able to calculate the same cluster priority list (see Fig. 5.3 and section 5.6.1 for more details) (step f). After calculating the same cluster interference topology matrix **A** and the same cluster priority list, each FAP is now able to calculate its own DFP and FSA allocations in a totally distributed way (steps g-l).

3. Each FAP in the cluster has to wait its turn in the priority list in order to determine its DFP and FSA allocations. Once a given FAP defines its own DFP and FSA solutions, it has to inform that allocation decision (number of allocated frequency resources and set of assigned sub-carriers) to the next FAPs in the cluster priority list. Since they can possibly not interfere directly with each other, flooding messages carrying this allocation information are necessary (steps g-l).

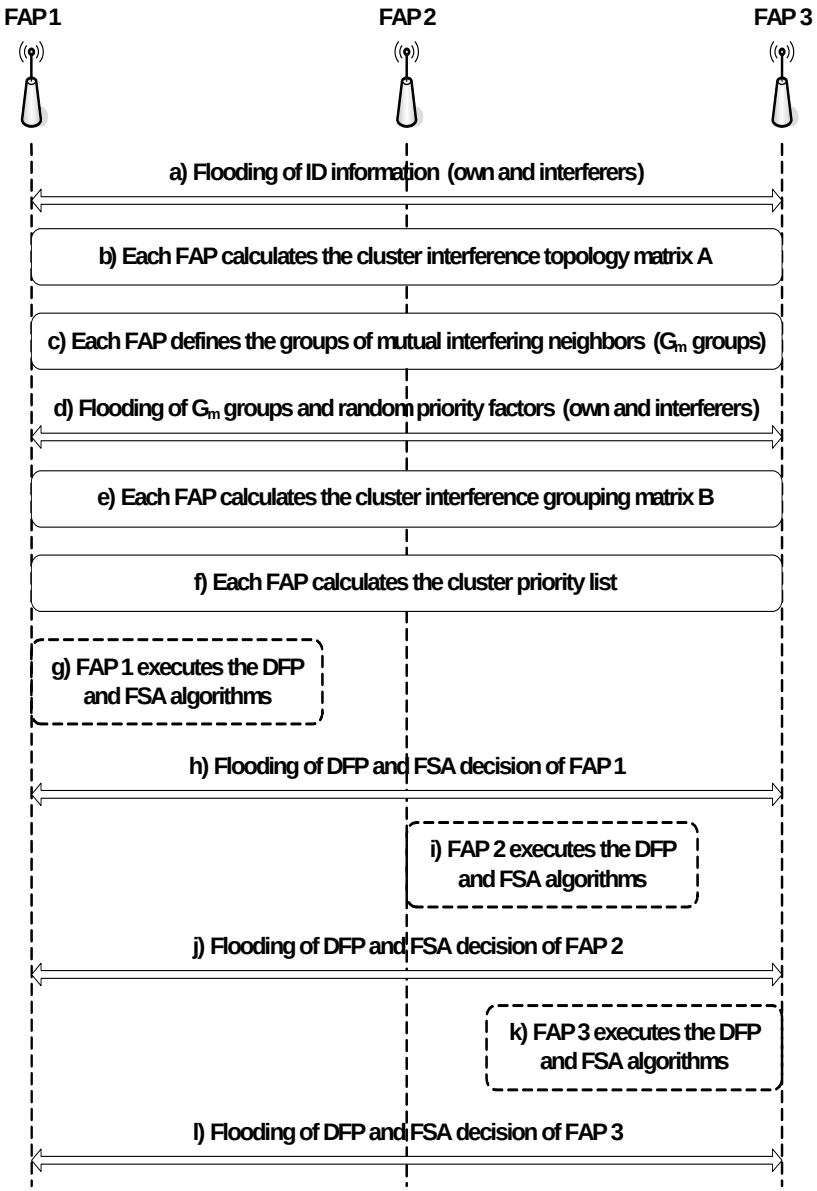


Figure 5.9.: Basic processing and communication protocol of the interference avoidance techniques in a distributed architecture

5.9. Discussion and Results

In this section, some important aspects related to the performance evaluation of the proposed interference avoidance techniques are addressed. The optimality of the DFP and FSA algorithms is evaluated in section 5.9.1, while section 5.9.2 presents a comparison between the interference avoidance policies using different metrics. Finally, we discuss in section 5.9.3 about the signaling overhead and the latency associated with the proposed algorithms in a centralized or distributed network architecture.

The numerical results presented in this section were taken from simulations of different clusters of FAPs. We considered several cluster sizes ($N = 4, \ldots, 7$) and different total numbers of frequency resources in the femtocell tier ($K = 15$ or 25). The chosen values for K were based on the number of PRBs standardized for the LTE system considering a bandwidth of 3 and 5 MHz, respectively. Fixing N and K, each interference avoidance policy was evaluated over 500 different cluster topologies. The simulation results presented in this section are average values considering all these snapshots.

5.9.1. Optimality Analysis

Dynamic Frequency Planning

As explained in section 5.6.2, the DFP algorithms based on heuristics are able to find sub-optimum solutions of the frequency planning. The BnB-based DFP algorithms proposed in section 5.6.3 are able to find the optimum planning of frequency allocations that maximizes the objective functions of the corresponding interference avoidance policies proposed in section 5.5. Table 5.1 presents the comparison between these two DFP approaches for different interference avoidance policies by presenting the percentage of times that the heuristic-based solution matched the optimum solution found by the BnB technique.

Table 5.1.: Percentage of matching (%) between the optimum solution of the BnB-based DFP and the sub-optimum solution of the heuristic-based DFP

	N=4		**N=5**		**N=6**		**N=7**	
	K=15	**K=25**	**K=15**	**K=25**	**K=15**	**K=25**	**K=15**	**K=25**
SFAM	100	100	100	100	100	100	98.0	98.6
MMFA	100	100	100	100	100	100	100	100
LMMFA	100	100	100	100	99.6	100	96	99.4

It can be observed that the heuristic-based DFP algorithms are also able to find the optimum solution in most of the times (matching percentage higher than 96% in the worst case). It is interesting to notice that for simpler cluster topologies (lower number of FAPs), the heuristic-based DFP showed an exact correspondence to the solutions found by the BnB-based DFP. Furthermore, since the frequency planning of the MMFA policy can be solved by simple and efficient heuristics, the corresponding heuristic-based DFP algorithm was able to find the optimum solution in all

scenarios and snapshots.

When the heuristic solutions did not match the optimum solutions, the percentaged difference between the corresponding values of the objective functions was calculated. Suppose that the values of the objective function for the sub-optimum heuristic- and optimum BnB-based solutions are $f(x)$ and $f(y)$, respectively. The percentaged difference is given by $|f(y) - f(x)|/f(x)$. This result is presented in Table 5.2 and gives us an idea how far from the optimum the heuristic-based solution is. The maximum percentaged difference was 11.11% for the case of the MMFA policy with $N = 6$ and $K = 15$, which is a reasonable value. This indicates that when the heuristic-based is not able to find the optimum solution, the calculated sub-optimum is still roughly close to the optimum. Notice that these percentages were calculated only for the cases of mismatches, which are rare events as indicated in Table 5.1.

Table 5.2.: Mean percentaged difference (%) between heuristic-based and BnB-based DFP considering only the cases where the solutions do not match

	N=6		**N=7**	
	K=15	**K=25**	**K=15**	**K=25**
SFAM	0	0	10.47	8.58
MMFA	0	0	0	0
LMMFA	11.11	0	6.30	7.09

A great advantage of the heuristic-based algorithms compared with the ones based on the BnB technique is the required processing time. As shown in Table 5.3, the difference in Central Processing Unit (CPU) processing time can reach three orders of magnitude. Comparing the interference avoidance policies, in general LMMFA is the one that requires the highest computational power, followed by SFAM and finally MMFA. This is due to the intrinsic complexity of the corresponding optimization problems.

Table 5.3.: Mean CPU time in seconds of the heuristic-based and BnB-based DFP algorithms

		N=4		**N=5**		**N=6**		**N=7**	
		K=15	**K=25**	**K=15**	**K=25**	**K=15**	**K=25**	**K=15**	**K=25**
SFAM	**Heur.**	0.011	0.009	0.014	0.015	0.021	0.023	0.037	0.033
	BnB	0.124	0.352	0.564	3.6	2.5	22.9	17.4	61.0
MMFA	**Heur.**	0.002	0.001	0.002	0.002	0.002	0.003	0.003	0.003
	BnB	0.235	0.669	0.498	2.120	0.943	4.8	3.8	12.5
LMMFA	**Heur.**	0.012	0.010	0.012	0.016	0.017	0.026	0.034	0.037
	BnB	2.7	17.4	14.5	34.1	22.1	84.2	28.7	166.0

Based on what was explained above, it can be concluded that the heuristic-based DFP algorithms are preferred to the ones based on the BnB technique. In most of the scenarios and cluster topologies

evaluated, the heuristic-based algorithms were able to find the optimum solution. In the cases it did not succeed, it was able to find a sub-optimum solution very close to the optimum. Moreover, these algorithms have a simple implementation and require a computational time much lower than their counterparts based on the BnB technique.

Femtocell Sub-carrier Allocation

As commented in section 5.7, the FSA algorithm has a significant importance on the feasibility of the solution given by the DFP algorithm. If an unproper FSA is used, an optimum frequency planning cannot be guaranteed, which can result in a sub-optimum solution for the interference avoidance problem.

In order to demonstrate the optimality of the MFR FSA algorithm, it is compared with the Available Frequency Reutilization (AFR) algorithm, which is a simplified version of MFR, where only its first step, related to the definition of the set of available sub-carriers of each FAP, is executed. The AFR algorithm chooses the sub-carriers from the pool of available resources without taking into account the impact of this resource allocation on the other FAPs that will still define their allocations. This is a critical issue in more intricate cluster topologies.

Table 5.4 presents the mean percentage of frequency resources that would suffer co-channel interference if AFR were used. This mean value is calculated considering all simulation snapshots. If a sub-carrier suffers interference as a result of the resource assignment, it means that the corresponding DFP allocation was not respected.

Table 5.4.: Mean percentage of interfered sub-carriers (%) over all snapshots using the AFR[a] FSA[b] algorithm and considering the heuristic-based DFP algorithm

	N=6		N=7	
	K=15	**K=25**	**K=15**	**K=25**
SFAM	0.03	0.02	0.11	0.04
MMFA	0.09	0.12	0.11	0.08
LMMFA	0.18	0.16	0.96	0.66

[*a*] For smaller clusters (N = 4 or 5), the AFR algorithm was able to completely avoid interference on the sub-carriers.
[*b*] The MFR FSA algorithm was able to completely avoid interference on the sub-carriers for all simulated DFP allocations and cluster topologies.

Although Table 5.4 shows the results of the heuristic-based DFP algorithms, similar results were found for the BnB-based DFP. Notice that for higher values of N, i.e. more complex clusters, the AFR algorithm is not able to guarantee complete co-channel interference avoidance, as required by the optimization problems of the interference avoidance policies. On the other hand, it was

observed in the simulations that the MFR FSA is able to assign all the sub-carriers according to the frequency planning defined by the DFP algorithms without causing any co-channel interference. Therefore, considering various scenarios and cluster topologies, the MFR algorithm was always able to guarantee the feasibility of the corresponding DFP solutions.

However, notice that the values presented in Table 5.4 are very small, i.e. the performance of the AFR algorithm is very close to the optimum provided by MFR. Although MFR is an efficient algorithm that runs fast, the advantage of AFR is that it requires a fraction of the computational time demanded by MFR.

5.9.2. Comparison of Interference Avoidance Policies

Three metrics are used to compare the performance of the proposed interference avoidance policies: sum of frequency allocations, which is a metric of spectral efficiency, minimum number of frequency allocations, and a fairness index. The index used in this work is based on the general fairness index proposed by [86] (see more details in section A.7.3 of appendix A). The fairness index of a given cluster is defined as

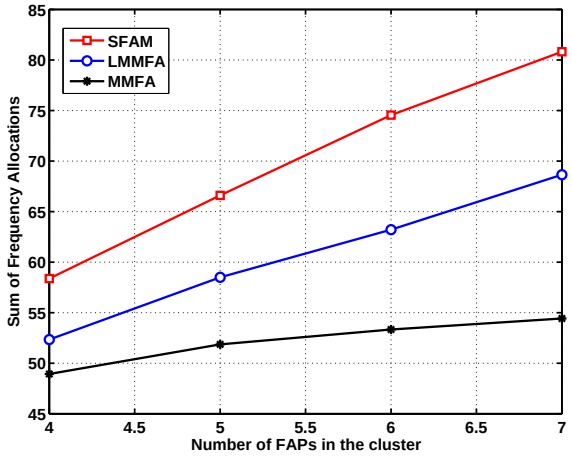
$$\Phi_{\text{cluster}} = \frac{\left(\sum_{n=1}^{N} F_n\right)^2}{N \cdot \sum_{n=1}^{N} \left(F_n\right)^2}, \tag{5.20}$$

where N is the number of FAPs in the cluster and F_n is the number of frequency resources allocated to the nth FAP. This index indicates the degree of fairness in the distribution of frequency resources within the cluster of FAPs. Notice that $1/N < \Phi_{\text{cluster}} \leq 1$. A perfect fair allocation is achieved when $\Phi_{\text{cluster}} = 1$, which means that the number of frequency resources allocated by the DFP and FSA algorithms are equal for all FAPs in the cluster. The unfairest allocation occurs when Φ_{cluster} approaches the lower bound $1/N$, which means that almost all frequency resources were allocated to only one FAP. Remember that the interference avoidance policies require that at least one frequency resource must be allocated to each FAP (see section 5.5).

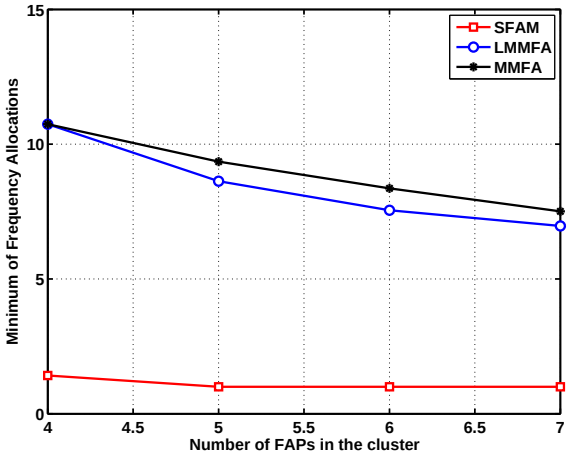
The following figures assume different values of N and a fixed value of $K = 25$, which is the total number of available frequency resources. The results were generated assuming the heuristic-based DFP algorithms and the MFR FSA algorithm.

Fig. 5.10(a) depicts the sum of the frequency allocations for each interference avoidance technique. Since this is the objective function of the SFAM policy, it is the one that presents the highest sum values, and consequently, the highest spectral efficiency. The higher spectral efficiency is a consequence of the interference-based opportunistic allocation of the SFAM policy: allocate more frequency resources to the FAPs that suffer less interference. In order to achieve a high percentage of frequency reutilization, SFAM allocates the minimum allowed number of frequency resources to the FAPs that suffer more interference. This action decreases the value of the minimum number of resources allocated by SFAM, as indicated in Fig. 5.10(b). In the opposite extreme, we have the MMFA policy, which does not use an opportunistic allocation. As a consequence, it presents a low number of frequency reutilizations (see Fig. 5.10(a)). However, as Fig. 5.10(b) shows, this

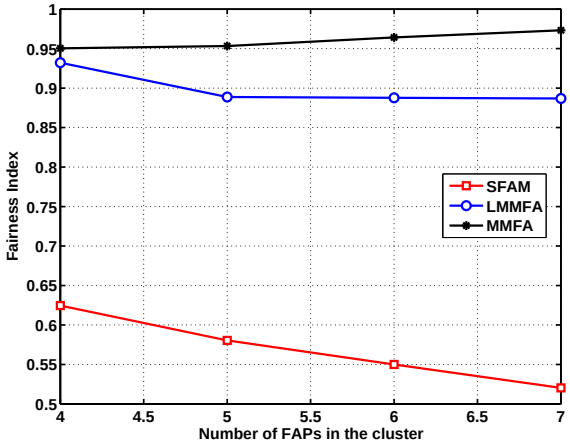
disadvantage is counterbalanced by the fact that a good resource allocation for all FAPs is assured. The LMMFA policy is a trade-off between SFAM and MMFA: the minimum number of frequency allocations is as high as the one presented by MMFA, while a good spectral efficiency is achieved.



(a) Sum of frequency allocations



(b) Minimum number of frequency allocations



(c) Fairness index

Figure 5.10.: Comparison of the interference avoidance techniques

The conflicting relation between the sum of frequency reutilizations and the minimum number of frequency allocations is clearly reflected in the fairness index, which is plotted in Fig. 5.10(c). As expected, MMFA is the fairest policy because all FAPs are allocated approximately the same number of resources, as indicated by a fairness index close to 1. On the other hand, the SFAM policy is the unfairest on the resource distribution. In its attempt to maximize the sum of frequency allocations in the cluster, some FAPs will use almost the total bandwidth dedicated to the femtocell tier, while others will be allocated the minimum number of resources allowed by the optimization problem, which is 1. Finally, the LMMFA policy is able to provide a high fairness among the FAPs within each layer of the cluster, and the product function, used in the optimization objective, also

allows a good fairness between different layers. As a consequence, we have that the fairness index of the whole cluster given by the LMMFA policy is also high.

Looking at Figs. 5.10(a) and 5.10(c), one can see the important trade-off between spectral efficiency and fairness. The SFAM policy provides the highest spectral efficiency at the expense of an unfair resource distribution. The MMFA policy is not spectrally efficient but assures a good minimum resource allocation for all FAPs. The LMMFA turns out to be the best alternative: it provides high fairness among FAPs and also presents a good performance regarding spectral efficiency.

5.9.3. Signaling Overhead and Latency Analysis

Section 5.8 briefly described how the interference avoidance techniques proposed in this work could be implemented in a centralized or distributed network architecture. Therefore, the signaling overhead generated by these proposed techniques must be also analyzed as well as the latencies involved in the resource allocation decisions. In the analysis, we have made some assumptions regarding the cluster topologies in order to deal with worst case scenarios.

It is assumed that there are C isolated clusters of FAPs in the macrocell network, where each cluster has N FAPs. We consider a worst case scenario where each FAP interferes with the other $N-1$ FAPs inside the cluster. There are K frequency resources (sub-carriers) dedicated to the femtocell tier. We assume that all variables that need to be transmitted in the network can be represented by one byte (possible values are $0-255$). Packet headers are not taken into account.

Firstly, let us consider the centralized approach illustrated in Fig. 5.8. The RRA broker communicates with the FAPs by means of a backhaul connection. It is assumed that before the execution of the DFP algorithm, the FAPs sensed the environment and discovered the IDs of their interfering neighbors. The signaling between the RRA broker and the FAPs can be described as:

- Communication FAP $\Rightarrow$ RRA broker (Uplink): Initially, every FAP must send to the RRA broker N bytes (its own ID and the IDs of its interferers). This results in a transmission overhead of N bytes for each FAP and a reception overhead of $C \cdot N^2$ for the RRA broker.

- Communication RRA broker $\Rightarrow$ FAPs (Downlink): The RRA broker has to inform to the FAPs the decisions of the DFP and FSA algorithms. The DFP allocation of each cluster can be represented by a vector F of N elements, which requires N bytes. The FSA allocation of a cluster can be represented by N vectors $\mathcal{S}_n$. Each vector $\mathcal{S}_n$ is the set of indexes of the sub-carriers assigned to the nth FAP. In total, the sum of the number of elements of all $\mathcal{S}_n$ sets is K, which is the total number of frequency resources. Based on that, the transmission overhead for the RRA broker is given by $C \cdot (N + K)$ bytes, while the reception overhead for each FAP is $N + K$ bytes.

Table 5.5 presents an estimation of the signaling overhead generated by the interference avoidance techniques considering different values of N and K and a period $T = 1$ second, i.e. the DFP and

FSA algorithms are executed every second. It is assumed that the RRA broker controls 50 clusters of FAPs[3] (C = 50). Notice that in the centralized approach, it makes no sense to refresh the interference avoidance resource allocation in a period lower than the Round Trip Time (RTT) between the RRA broker and the FAPs.

Table 5.5.: Estimated signaling overhead in kbps considering a centralized network architecture

		N=5		N=6		N=7	
		K=15	K=25	K=15	K=25	K=15	K=25
UL RRA Broker	$\frac{C \cdot N^2}{T}$	10.0	10.0	14.4	14.4	19.6	19.6
DL RRA Broker	$\frac{C \cdot (N+K)}{T}$	8.0	12.0	8.4	12.4	8.8	12.8
UL FAP	$\frac{N}{T}$	0.040	0.040	0.048	0.048	0.056	0.056
DL FAP	$\frac{(N+K)}{T}$	0.160	0.240	0.168	0.248	0.176	0.256

It can be concluded that the signaling overhead in the centralized approach is very small. The bigger values related to the RRA broker are not a problem due to its higher data transfer capability. It is important to emphasize that a remarkable advantage of the interference avoidance techniques proposed in this work is that the involved algorithms do not need to be executed so often. Due to the fixed/nomadic nature of the FAPs deployment, the interference pattern of the cluster does not change frequently. Therefore, the proposed interference avoidance techniques can be executed in a period of time in the order of seconds (or even more), which makes its signaling overhead negligible.

The latency of the RRA techniques depends on the processing time of the DFP and FSA algorithms in the RRA broker and the RTT time between the RRA broker and the FAPs. It was shown in section 5.9.1 that the heuristic-based DFP algorithms are very fast to run. Furthermore, we assume that the RTT time in a Fourth Generation (4G) femtocell network, such as LTE or WiMAX, is also reduced. Therefore, it can be concluded that the latency of the proposed techniques suits the requirements for implementation in a centralized 4G femtocell network.

[3] According to [134], a dense urban scenario can be characterized by 500 users / Km^2. Assuming a macrocell radius of 1 km, we have approximately 400 users per macrocell. If the femtocell technology has a market penetration of 10%, we have 40 FAPs per macrocell. Assuming that half of these FAPs form clusters with N = 4, we have 5 active clusters in the macrocell area. If we assume that the RRA broker is responsible for the FAPs in the area of 10 macrocells, we reach the value of C = 50 clusters.

In the distributed network architecture (see Fig. 5.9), the FAPs within a cluster must cooperate in order to calculate their resource allocation following an interference avoidance policy. The FAPs communicate among them using broadcast messages and/or measurement reports [12]. In a cluster, there may be FAPs that cannot communicate directly because their coverage area do not overlap. In this case, we assume that flooding protocols are used so that the information needed by the DFP and FSA algorithms are spread among all FAPs in the cluster. Suppose that a message of X bytes must be transmitted by each FAP to all other $N-1$ FAPs in the cluster. So, N messages of size X must be known by all FAPs after the flooding procedure. In a rough estimation, if each FAP retransmits a received message only once to its neighbors, a naive flooding protocol is able to spread this information over the whole cluster. Thus, each FAP has to transmit its own message and also retransmit the $N-1$ messages coming from the other FAPs. The total number of bytes transmitted and received by each FAP in this flooding process would be $X \cdot N$ and $X \cdot (N-1)$, respectively.

Looking at Fig. 5.9, one can see that the following information should be flooded in the cluster in different instants in order to allow each FAP to define its own DFP and FSA allocations: ID information, G_m groups, random priority factor, and DFP and FSA decisions. Each FAP must inform its ID and the IDs of its neighbors. Supposing that the cluster is fully connected, the ID information of each FAP can be represented by a vector of N elements. The G_m groups that each FAP participates are difficult to predict, since they are highly dependent on the cluster topology. We roughly assume that all the elements of the G_m groups of a given FAP can also be represented by a vector of N elements. Each FAP has to inform to the others its random priority factor, which can be coded in one byte. The DFP decision for a given FAP is the number of frequency resources allocated to it, which can also be coded in one byte. Finally, the FSA decision is the set of sub-carriers' indexes assigned to a given FAP, which can have K elements in a worst case estimation. Based on what was explained above, the total number of bytes X that must be flooded by each FAP is $X = 2N + K + 2$. As explained before, taking into account the number of retransmissions required by the flooding process, the total number of transmitted bytes required by the interference avoidance technique is given by $X \cdot N$, which yields a transmission overhead of $2N^2 + N \cdot (K+2)$ bytes. On the other hand, the total number of received bytes required by the interference avoidance technique is estimated by $X \cdot (N-1)$, which gives us a reception overhead of $2N^2 + K \cdot (N-1) - 2$ bytes for each FAP.

An estimation of the signaling overhead generated by the interference avoidance technique in the transmission and reception links of each FAP is presented in Table 5.6. Different values of N and K are considered and it is assumed that the DFP and FSA algorithms are executed every second.

The signaling overhead in a distributed network approach is also very small compared with the transmission capability of the FAPs. The values presented in Table 5.6 were calculated considering that a new resource allocation is defined every second. Following the logical assumption that the proposed interference avoidance techniques can be executed in a even lower frequency, the signaling

overhead in the distributed approach will also be negligible.

Table 5.6.: Estimated signaling overhead in kbps considering a distributed network architecture

		N=5		N=6		N=7	
		K=15	K=25	K=15	K=25	K=15	K=25
Tx FAP	$\frac{2N^2 + N \cdot (K+2)}{T}$	1.08	1.48	1.39	1.87	1.74	2.30
Rx FAP	$\frac{2N^2 + K \cdot (N-1) - 2}{T}$	0.864	1.18	1.16	1.56	1.49	1.97

The same assumption smoothes the latency requirements of the algorithms. The operations that most contribute to the latency are the flooding of messages and the fact that the FAPs must follow an order of priority in the calculation of their DFP and FSA allocations. For example, the last FAP in the cluster priority list has to wait all its predecessors to define their RRA allocations and send this information to the other FAPs by means of a flooding procedure. The latency of the distributed approach could be prohibitive in the case that the interference avoidance technique had to be executed very frequently. However, as explained before, this is not the case because of the fixed/nomadic nature of the FAPs.

From the analysis presented above, it can be concluded that both centralized and distributed approaches are feasible regarding signaling overhead and latency issues. The centralized approach has some advantages compared with the distributed one: the radio resources of the FAPs are not used for signaling, and the latency can be estimated more accurately because it does not depend on flooding procedures. Some disadvantages are the higher economic costs and the processing and signaling burden on the RRA broker in the case it had to control many clusters of FAPs.

5.10. Conclusions

Three novel interference avoidance techniques that perform a high-level, mid/long-term frequency planning on the clusters of FAPs are proposed to balance the spectral efficiency and resource-based fairness in the femtocell tier. The interference avoidance optimization problems are solved in separate and sequential steps accomplished by the DFP and FSA algorithms, which define the allocation of frequency resources among the FAPs.

The main conclusions based on the simulations results are summarized below:

- In most of the scenarios and cluster topologies evaluated, the heuristic-based DFP algorithms

were able to find the optimum frequency planning. In the cases it did not succeed, it was able to find a sub-optimum solution very close to the optimum achieved by the DFP algorithms based on the BnB technique. Moreover, these algorithms have a simple implementation and require a short computational time.

- In all simulated scenarios, the MFR FSA algorithm was able to allocate all the sub-carriers according to the frequency planning defined by the DFP algorithms without causing any co-channel interference, i.e. it was always able to guarantee the feasibility of the corresponding DFP solutions. The AFR algorithm, which is a simplified version of MFR that defines the sub-carrier allocations based only on the set of available sub-carriers, also presents very good results. Although it is a sub-optimum algorithm, most of the times it is able to find the same solution of MFR with less computational time.

- Among the proposed interference avoidance techniques, the LMMFA strategy is the best alternative, since it is a trade-off that finds the best compromise between spectral efficiency (high frequency reutilization) and fairness in the resource distribution among FAPs.

- Since the proposed algorithms do not need to be executed so frequently, it was shown that the signaling overhead in both centralized and distributed networks is negligible and the latency involved in the resource allocation decision is adequate for implementation on a 4G femtocell network.

Chapter 6

General Conclusions and Perspectives

In this thesis, we have proposed some novel adaptive Radio Resource Management (RRM) techniques for the downlink of Fourth Generation (4G) macro- and femtocell networks using Orthogonal Frequency Division Multiple Access (OFDMA). Different RRM problems were represented using optimization-based formulations composed of an objective function and several network constraints. The consideration of different objective functions gives rise to distinct Radio Resource Allocation (RRA) policies, one for each optimization goal. Some of these optimization goals can be conflictive between them, such as the case of the joint maximization of capacity (efficiency) and fairness, or capacity and coverage, or coverage and Quality of Service (QoS), etc.

In this work, we were interested at studying the fundamental trade-off between efficiency and fairness in wireless networks in general, and OFDMA-based cellular systems in particular. Static RRA policies aiming the maximization of system capacity, maximization of fairness in the resource distribution, or a compromise between these factors, were proposed. This thesis took a step further and propounded adaptive RRA techniques that are able to control the aforementioned trade-off according to new network management paradigms.

These RRA strategies are comprised of different RRA algorithms, such as sub-carrier assignment, power allocation and interference avoidance based on frequency planning, whose functionalities were found using different optimization tools. Finally, these techniques were evaluated by means of system-level simulations, and interesting and meaningful conclusions were taken. Sections 6.1 and 6.2 summarize the main conclusions and present some perspectives of future work, respectively.

6.1. Conclusions

The last section of chapter 2 presents a detailed explanation about the intrinsic trade-off that exists in wireless systems between efficiency in the resource usage and fairness in the resource distribution.

OFDMA-based systems have multi-user and multi-cell diversities, i.e. sub-carriers are likely to be in different quality states for different users, or randomly located Femtocell Access Points (FAPs) can suffer different levels of interference depending on the topology of their vicinity. In order to use the radio resources efficiently and maximize capacity, opportunistic RRA algorithms try to explore these diversities allocating the resources to the network players (users or FAPs) with better efficiency indicators. This decision provides uneven resource distribution among these players, which will cause fairness problems. On the other hand, schemes that provide absolute fairness penalize the users with better conditions or the FAPs with less interference, which reduces the system efficiency. Trying to maximize one of these objectives causes the detriment of the other.

This important RRM problem in a macrocell scenario was addressed in this thesis using two different approaches: fairness/rate adaptive optimization, which is described in chapter 3, and Utility Theory, which is used in chapter 4. The former was used to formulate optimization problems based on the users' instantaneous rates, and so are suitable for Non-Real Time (NRT) services. The latter is a more general and flexible approach that allowed the proposition of utility-based optimization problems based on users' average data rate (throughput) and Head-Of-Line (HOL) packet delay, which are suitable for NRT and Real Time (RT) services, respectively.

In chapter 3, we have introduced the novel concepts of efficiency-fairness planes, curves and regions. These visual representations comprise a new performance evaluation tool to analyze the trade-off between efficiency and fairness, which complements the performance evaluation of RRA techniques in wireless networks. The efficiency-fairness regions are somewhat similar to the capacity regions: the higher the region below the curve, the better the technique. The usefulness of this tool was tested with success in the performance evaluation of the RRA techniques proposed in chapters 3 and 4.

In both chapters, we started studying some well-known classic RRA policies found in the literature. All these classic policies are static, because their objectives are either the maximization of capacity, or the maximization of fairness, or to achieve a trade-off between these factors. As expected, their performance evaluation using the efficiency-fairness plane showed a limited and static behavior, i.e. each classic RRA policy was represented by a single point in the plane.

This fact motivated us to propose novel adaptive RRA techniques that could control the aforementioned trade-off and provide to the network operators the possibility of operating their systems on any feasible point of the efficiency-fairness plane. In order to do that, we needed to consider a new network management paradigm. The idea behind this paradigm starts with the calculation of a cell fairness index, which is based on a well-known general fairness index found in the literature and described in details in appendix A. It was demonstrated in this thesis by means of system-level simulations that each value of the cell fairness index corresponds to a different performance in terms of efficiency in the resource usage, or consequently system capacity. Therefore, if the mobile operator is able to force the network to operate on a desired fairness level, it can control the trade-off between efficiency and fairness.

In this work, we proposed two ways to control the fairness in the system. The first is an instantaneous fairness control (short-term) based on fairness/rate adaptive techniques (see chapter 3), which use iterative sub-carrier assignment and power allocation algorithms that are able to vary the system fairness according to a desired value.

The second is an average fairness control (mid-term) that uses utility-based RRA techniques (see chapter 4). Two general utility-based RRA frameworks called alpha-rule and beta-rule, which are suitable for NRT and RT services, respectively, were proposed. The Adaptive Throughput-Based Fairness (ATF) and Adaptive Delay-Based Fairness (ADF) techniques dynamically adapt the fairness-controlling parameters α and β of the alpha-rule and beta-rule frameworks using a feedback control loop, in order to achieve a desired fairness distribution in terms of throughput (average data rate) or HOL delay, respectively.

It was demonstrated by means of performance comparison in a macrocell scenario with NRT services that the fact of considering a time window in the fairness control performed by the utility-based ATF technique is more advantageous than the instantaneous fairness control carried out by the fairness/rate adaptive techniques. ATF is able to achieve efficiency-fairness and satisfaction-fairness regions as large as the fairness/rate adaptive strategies using less computational resources.

It was concluded in chapters 3 and 4 that the fairness control executed by only the corresponding Dynamic Sub-carrier Assignment (DSA) algorithms followed by a power allocation step that divides the base station transmit power equally among the sub-carriers was sufficient. The results achieved with Equal Power Allocation (EPA) were acquired with less computational time and were similar to the ones provided by the Joint approach, where the fairness control is done by both DSA and Adaptive Power Allocation (APA) algorithms.

As shown is chapters 3 and 4, due to the high variability of efficiency indicators of NRT users[1] in a macrocell scenario, the trade-off between efficiency and fairness cannot be managed by adaptive RRA techniques in a 'win-win' approach[2] in this scenario. This gap is even more critical when there are indoor users in the macrocell coverage area, because the signal level indoors is weak due to penetration losses. One efficient way to diminish the appearance of the trade-off in the downlink of cellular networks is to shorten the distance between the transmitter (base station) and the receivers (users). With the enhanced signal quality, both capacity and fairness can be increased at the same time. This can be accomplished with the use of femtocells. Among many other advantages and benefits of using femtocells, they can help to approximate the network performance to the optimal operational point of the efficiency-fairness plane, characterized by high efficiency and high fairness.

Although femtocells can help to solve the problem of the trade-off between resource efficiency and QoS-based fairness in cellular networks, the trade-off between system capacity and resource-based

[1] As indicated in Table 2.1 on page 35, the efficiency indicator used for NRT services is the achievable sub-carrier transmission rate.

[2] As shown in section 4.5.2 of chapter 4, the same is not true for the management of the trade-off between resource efficiency and delay-based fairness in a scenario with RT services. In this case, the proposed utility-based ADF technique derived from the utility-based beta-rule framework was able to achieve several fairness levels with high spectral efficiency. In this way, the network can operate close to the optimal point of the efficiency-fairness plane.

fairness also appears when the radio resources must be distributed among several FAPs. Chapter 5 addresses at the same time two problems: 1) the seamless co-existence between neighboring FAPs in a femtocell tier using interference avoidance techniques; and 2) how the aforementioned trade-off between spectral efficiency and resource-based fairness could be balanced.

In this sense, three novel interference avoidance techniques were proposed. They are called Sum Frequency Allocation Maximization (SFAM), Max-Min Frequency Allocation (MMFA) and Layered Max-Min Frequency Allocation (LMMFA). All of them could be implemented in a centralized or distributed network architecture. These RRA strategies are based on the concept of a high-level, mid/long-term frequency planning, and are composed of Dynamic Frequency Planning (DFP) and Femtocell Sub-carrier Allocation (FSA) algorithms. Firstly, the DFP algorithm decides how many frequency resources must be allocated to each FAP. Next, the FSA algorithm chooses the particular set of sub-carriers that each FAP must use. This specific study focused on the resource allocation among the FAPs and did not consider the resource assignment to the end-users. Anyway, any of the RRA techniques that were studied in chapters 3 and 4 could be used to decide the user resource assignment.

It was concluded from the simulation results that LMMFA is the best alternative among the proposed interference avoidance techniques, because it provides high fairness among FAPs, presents a good performance regarding spectral efficiency, and also requires approximately the same processing time compared with the other strategies.

We noticed that there is no need to make an adaptive fairness control in the femtocell scenario because the clusters of FAPs are relatively small, so there is not so much freedom to control the trade-off between interference-based resource efficiency and fairness. Furthermore, the LMMFA technique was already able to find a good compromise between these objectives.

Table 6.1 summarizes the main features regarding the trade-off between efficiency and fairness of all RRA techniques studied/proposed in this thesis.

Table 6.1.: Summary of the features of the RRA techniques studied[a]/proposed[b] in this thesis regarding the trade-off between efficiency and fairness

RRA Techniques	Characteristics
Fairness/rate adaptive RRA techniques for NRT services in a macrocell scenario	
SRM	High resource efficiency and low rate-based fairness
MMR-P	Low resource efficiency and high rate-based fairness
SRM-P	Low resource efficiency and high rate-based fairness
FSRM	Dynamic trade-off between resource efficiency and rate-based fairness
FMMR-P	Dynamic trade-off between resource efficiency and rate-based fairness
FSRM-P	Dynamic trade-off between resource efficiency and rate-based fairness
Utility-based RRA techniques for NRT services in a macrocell scenario	
MR	High resource efficiency and low throughput-based fairness
MMF	Low resource efficiency and high throughput-based fairness
PF	Static trade-off between resource efficiency and throughput-based fairness
ATF	Dynamic trade-off between resource efficiency and throughput-based fairness
Utility-based RRA techniques for RT services in a macrocell scenario	
PF	High resource efficiency and low delay-based fairness
FIFO	Low resource efficiency and high delay-based fairness
M-LWDF	Static trade-off between resource efficiency and delay-based fairness
ADF	Dynamic trade-off between resource efficiency and delay-based fairness
Interference avoidance techniques in a femtocell scenario	
SFAM	High resource efficiency and low resource-based fairness
MMFA	Low resource efficiency and high resource-based fairness
LMMFA	Good trade-off between resource efficiency and resource-based fairness

[a] The RRA techniques already proposed in the literature that are studied in this thesis are: Sum Rate Maximization (SRM), Max-Min Rate with Proportional Rate Constraints (MMR-P), Sum Rate Maximization with Proportional Rate Constraints (SRM-P), Max-Rate (MR), Max-Min Fairness (MMF), Proportional Fairness (PF), First-In-First-Out (FIFO), Modified Largest Weighted Delay First (M-LWDF).

[b] The novel RRA techniques proposed in this thesis are: Fairness-Based Sum Rate Maximization (FSRM), Fairness-Based Max-Min Rate with Proportional Rate Constraints (FMMR-P), Fairness-Based Sum Rate Maximization with Proportional Rate Constraints (FSRM-P), Adaptive Throughput-Based Fairness (ATF), Adaptive Delay-Based Fairness (ADF), Sum Frequency Allocation Maximization (SFAM), Max-Min Frequency Allocation (MMFA) and Layered Max-Min Frequency Allocation (LMMFA).

6.2. Future Work

Some perspectives for the continuation of this thesis work are listed below:

- Consider other simulation models and scenarios for the evaluation of the fairness/rate adaptive and utility-based RRA techniques, such as: multiple macrocells, imperfect/incomplete Channel State Information (CSI), user mobility, discrete link adaptation, other traffic models, mixed services scenarios, user classes with different proportional rate requirements, etc.

- Study the impact on the trade-off between efficiency and fairness of considering the spatial diversity and RRA techniques that use multiple antennas, e.g. beamforming and Multiple Input Multiple Output (MIMO).

- Propose RRA policies that could explicitly map the trade-offs between fairness and QoS, or fairness and coverage, and propose adaptive RRA techniques to control these compromises.

- Propound new interference avoidance policies for the femtocell tier based on frequency planning that are able to optimize other network metrics and resources. The interference avoidance techniques proposed in this thesis could be called 'Frequency Allocation Adaptive', because they optimize the number of allocated frequency resources subject to a constraint on the total available bandwidth. We plan to propose a new family of policies called 'Bandwidth Adaptive', whose objective is to optimize the bandwidth dedicated to the femtocell tier taking into account constraints on the number of frequency resources that should be allocated to each FAP.

- Consider in the interference avoidance optimization problems the impact of the macrocell tier working on a spectrum sharing approach. After that, we plan to propose utility-based interference management techniques to mitigate the cross-tier interference.

- Modify and extend the proposed techniques for implementation in a Coordinated Multi-Point (CoMP) system. We intend to analyze how the coordination between multiple cell sites can aid the fairness control not only among users but also among base stations (inter-cell fairness).

Appendix A

System and Simulation Modeling

This appendix presents the system modeling that was considered in the formulation of the Radio Resource Management (RRM) techniques proposed in this thesis, as well as the models that were implemented in the system-level simulation tool used in this work. This simulator takes into account the most important aspects of the downlink of a cellular network in general, and the Orthogonal Frequency Division Multiple Access (OFDMA) technology in particular. Throughout the appendix, some validation results that demonstrate the correctness of the implemented models are shown.

The appendix is organized as follows. Section A.1 presents some general network assumptions. The network deployment, comprising macrocell, femtocells and users, is shown in section A.2. The propagation models are described in section A.3, while some coverage-related metrics are calculated in section A.4. Next, the link adaptation procedure and the traffic models are presented in sections A.5 and A.6, respectively. Section A.7 shows how the performance metrics considered in this work are calculated or defined. Finally, section A.8 presents the simulator structure.

A.1. General Assumptions

Some general simulation assumptions considered in this thesis are listed below.

- Although we consider frequency-selective Rayleigh fading, each sub-carrier experiences flat fading. In this way, we assume that the channel gains are constant over a Transmission Time Interval (TTI), but vary from one TTI to another.
- The Macrocell Base Station (MBS) has perfect knowledge of the Channel State Information (CSI) of all Mobile Terminals (MTs) in all sub-carriers.
- The resource allocation information (sub-carrier assignment, modulation and coding schemes, etc.) is sent to each MT in a separate control channel, so that the MTs can decode the data

in their own sub-carriers.

- The MTs are static, i.e. there is no mobility. However, in each simulation scenario several independent snapshots with different user distributions are simulated, which captures the system performance in different coverage situations.
- The downlink transmission scheme is based on Orthogonal Frequency Division Multiplexing (OFDM) using a normal cyclic prefix length with 7 symbols per time slot (TTI). The total sub-carrier bandwidth is 15 kHz, which accounts for both data and pilot symbols. In order to account for only data symbols, we consider an effective sub-carrier bandwidth of 14 kHz.
- The minimum resource block considered in the simulations is a time-frequency chunk formed by a time slot of 0.5 ms and an effective frequency bandwidth of 14 kHz (sub-carrier).

A.2. Network Deployment and User Distribution

The general high-level network deployment considered in this thesis is illustrated in Fig. A.1. In this figure, we can see a MBS with an omni-directional antenna that serves several outdoor MTs and several Femtocell Access Points (FAPs) located inside houses, offices and buildings that serve indoor MTs. Specific details about the network deployment and user distribution in the macrocell and femtocell scenarios are presented in sections A.2.1 and A.2.2, respectively.

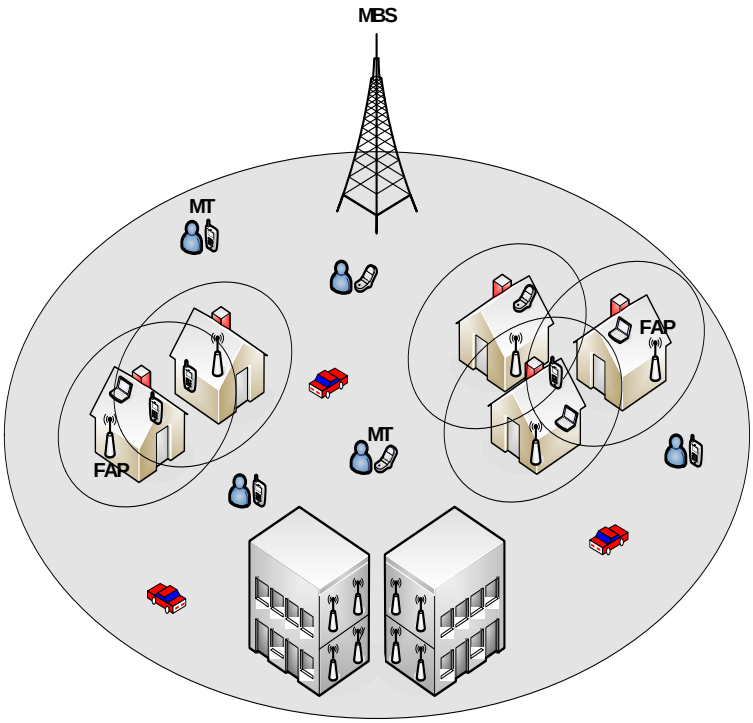
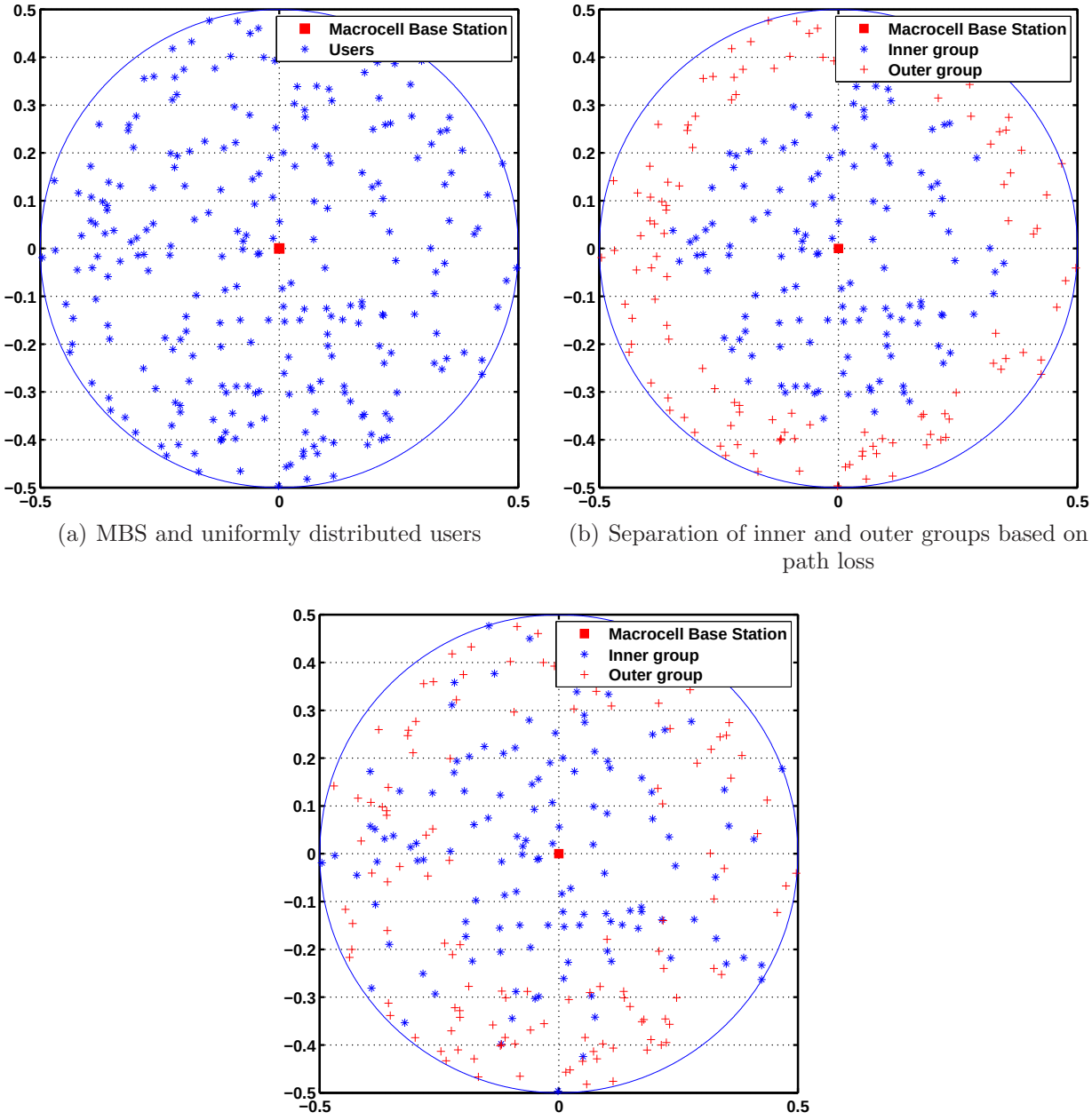


Figure A.1.: High-level network deployment

A.2.1. Macrocell Scenario

In the macrocell scenario, the users are uniformly distributed over the circular coverage area of the MBS, as indicated in Fig. A.2(a).



(a) MBS and uniformly distributed users

(b) Separation of inner and outer groups based on path loss

(c) Separation of inner and outer groups based on path loss + shadowing

Figure A.2.: Deployment of a Macrocell Base Station (MBS) and users

In order to analyze the relation between fairness and opportunistic resource allocation, it is interesting to separate the users in groups, for example inner and outer users. The inner group is comprised of users with good channel quality, while the outer group is the set of users with bad channel conditions. Figs. A.2(b) and A.2(c) depict the separation of the groups in a macrocell scenario based on the path loss, or the path loss plus shadowing, respectively. One can notice the higher variability of the channel quality when shadowing is considered. Finally, the analysis of inner and outer groups is also useful for the study of the compromise between fairness and coverage.

A.2.2. Femtocell Scenario

In the femtocell scenario, we do not perform simulations with radio propagation modeling because we do not deal with the resource allocation to the end-users. Instead of that, we evaluate interference avoidance techniques based on frequency planning, which do not take into account the CSI of the users (see chapter 5 for more details). The proposed interference avoidance policies are mainly dependent on the topology of the clusters of FAPs, which can be evaluated by means of simplified numerical simulations. Fig. A.3 depicts some examples of possible cluster topologies, where the circular coverage area of some neighbor FAPs overlap.

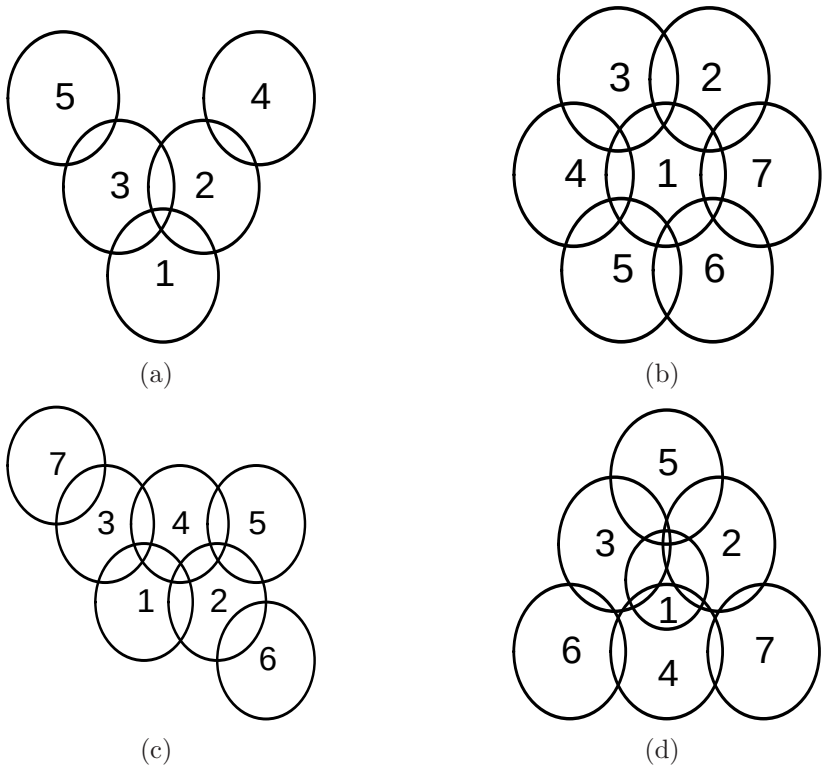


Figure A.3.: Examples of clusters of Femtocell Access Points (FAPs) with different topologies

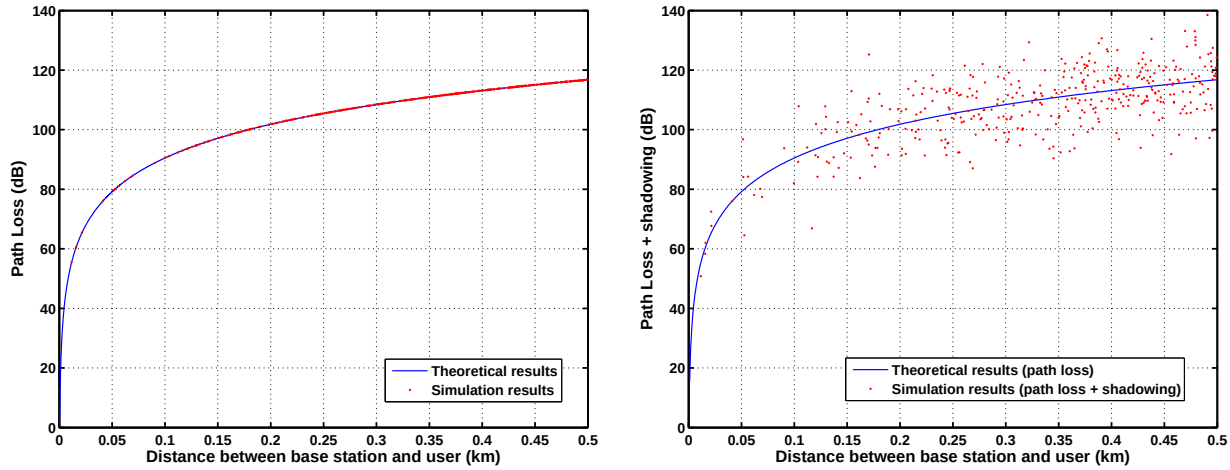
A.3. Propagation

A.3.1. Path Loss

The path loss follows the model proposed in [135] for a test scenario in urban and suburban areas. Considering a 2 GHz carrier frequency, and a mean MBS antenna height of 15 m, the equation of the path loss L_j^{path} in dB as a function of the distance d between the MBS and the jth MT in km is presented as follows:

$$L_j^{\text{path}} = 128.1 + 37.6 \log_{10} d. \tag{A.1}$$

The theoretical path loss given by (A.1) is plotted in Fig. A.4(a), as well as some samples taken from simulations. The match between both curves validates the path loss model implemented in the system-level simulator.



(a) Path loss

(b) Path loss + shadowing

Figure A.4.: Theoretical and simulation-based results for the path loss and shadowing

A.3.2. Large-Scale Fading

The modeling of the large-scale fading used in this thesis is the well-known zero-mean lognormal shadowing fading model characterized by a given standard deviation [136]. Assuming that the standard deviation is σ dB, we have that the shadowing fading with respect to the jth MT is defined as $L_j^{\text{shadow}} \sim N(0, \sigma)$ (in dB), being $N(0, \sigma)$ a normal random variable with mean equal to zero and a standard deviation equal to σ.

Fig. A.4(b) shows how the path gain composed of path loss and shadowing deviates from the theoretical path loss model. These simulation samples were taken using a log-normally distributed shadowing with standard deviation of 8 dB.

A.3.3. Small-Scale Fading

In this work, we assume that the small-scale fading (fast fading) follows a Rayleigh distribution. Rayleigh fading is a reasonable model when there are many objects in the environment that scatter the radio signal before it arrives at the receiver, and there is no dominant propagation along a Line of Sight (LOS) between the transmitter and receiver.

If the number of scatterers is sufficiently high, the central limit theorem can be used and the channel impulse response can be modeled as a Gaussian process irrespective of the distribution of the individual components. Without a dominant propagation dominant, this Gaussian process has zero mean and phase evenly distributed between 0 and 2π radians. Therefore, the envelope of the channel response will be Rayleigh distributed. One of the most popular approaches to generate the Rayleigh fading suitable for simulation purposes is the Jakes' model [137], which is the approach considered in this thesis. It comprises a method for simulating mobile radio fading that produces random phase modulation, a Rayleigh fading envelope, and a time-averaged, discrete approximation to the desired power spectrum.

With the Jakes' model, it is possible to simulate any propagation environment characterized by any power-delay profile. In this thesis, we consider the power-delay profile according to the Typical Urban (TU) model proposed by the 3rd. Generation Partnership Project (3GPP) [138].

In order to exemplify the correctness of the model implemented in the simulator, fast fading processes in the time and frequency domains are depicted in Figs. A.5(a) and A.5(b), respectively. Notice that in figure A.5(b) the abscissa axis represents the frequency in MHz. Therefore assuming a sub-carrier bandwidth of 14 KHz, it is evident that a flat fading is perceived by the different modulated sub-carriers.

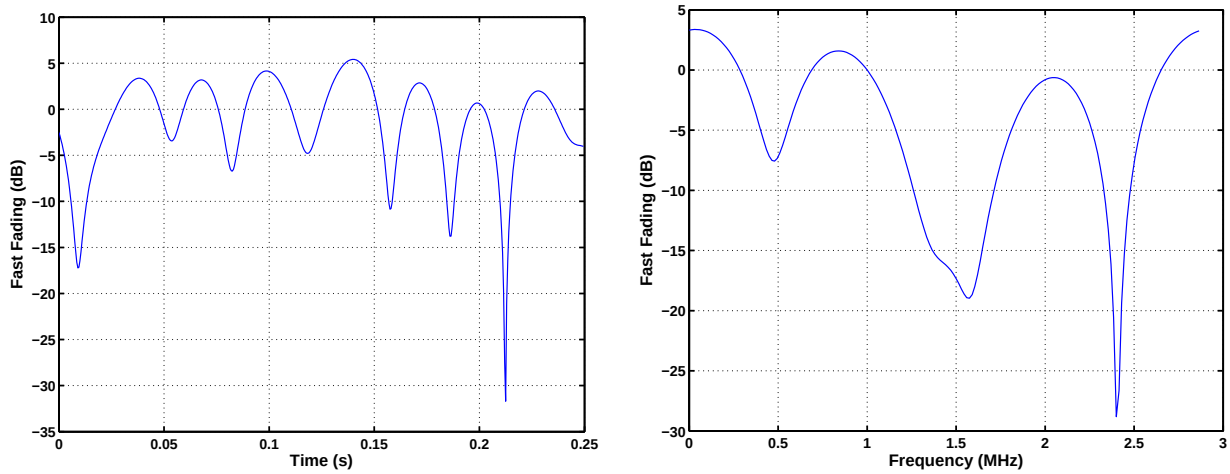
In the remainder of this appendix, we represent the fast fading gain of the jth MT in the kth sub-carrier in a given time instant as $L_{j,k}^{\text{fast}}$ (in dB).

A.4. Coverage

A.4.1. Macrocell Base Station Transmit Power Calculation

It is important to calculate which are the coverage area limits of the macrocell in order to assure that a given percentage of places receive the signal correctly. This random characterization of the coverage area is due to the large-scale fading (shadowing). The slow variation of the MT received power in each sub-carrier follows a lognormal Probability Density Function (PDF) when the power is in linear unit (e.g. mW), or a normal PDF when the power is in dBm. Assuming that the power is in dBm unit, the statistical characterization of the received power of the jth MT in the kth sub-carrier is given by the following PDF:

$$f\left(p_{j,k}^{\text{rx}}\right) = \frac{1}{\sqrt{2\pi}\sigma} \cdot e^{\frac{-\left(p_{j,k}^{\text{rx}} - \overline{p_{j,k}^{\text{rx}}}\right)^2}{2\sigma^2}}. \tag{A.2}$$



(a) Time domain (fixed sub-carrier) (b) Frequency domain (fixed time slot)

Figure A.5.: Example of small-scale fading (fast fading) taken from simulations

We have that $p_{j,k}^{\mathrm{rx}}$ is the received power of the jth MT in the kth sub-carrier, $\overline{p_{j,k}^{\mathrm{rx}}}$ is the mean received power according to a given prediction model and σ is the shadowing standard deviation.

The Signal-to-Noise Ratio (SNR) in dB of the jth MT in the kth sub-carrier is calculated as $\zeta_{j,k} = p_{j,k}^{\mathrm{rx}} - p^{\mathrm{noise}}$, where p^{noise} is the noise power. The noise power in dBm is calculated as

$$p^{\mathrm{noise}} = -174 + 10\log_{10}(\Delta f) + \vartheta, \tag{A.3}$$

where $\Delta f = B/K$ is the sub-carrier bandwidth given by the ratio between the total system bandwidth B and the total number of sub-carriers K. The variable ϑ is the noise figure in dB.

We want to guarantee that in the cell edge, ξ % of the locations will have coverage, i.e. the users located in this region will have at least a minimum quality to receive the packets correctly. This minimum quality can be defined as the lowest modulation level that can be used in the system. According to the link adaptation curve (see Fig. A.7 in section A.5), the lowest modulation level corresponds to a minimum SNR value. Since we can use the same normal PDF shown in equation A.2 to model the SNR in dB, we have that the probability of coverage can be calculated as

$$\xi = \mathrm{Prob}\left(\zeta_{j,k} > \zeta^{\mathrm{th}}\right) = \int_{\zeta^{\mathrm{th}}}^{\infty} f(\zeta)\, d\zeta = \frac{1}{2} - \frac{1}{2}\mathrm{erf}\left(\frac{\zeta^{\mathrm{th}} - \overline{\zeta_{j,k}}}{\sqrt{2}\sigma}\right), \tag{A.4}$$

where ζ^{th} is the threshold SNR in which the lowest modulation level can be used, $\overline{\zeta_{j,k}}$ is the mean SNR and erf is the error function.

The interpretation of this calculation is the following: among all the locations that have a mean SNR level of $\overline{\zeta_{j,k}}$, ξ % of them receive data correctly (an SNR level greater than ζ^{th}). After making some calculations, it is possible to isolate $\overline{\zeta_{j,k}}$ from equation A.4. Once we have this mean SNR

value, we can calculate the transmission power of each sub-carrier k taking into account the path loss model (see section A.3.1). Thus, using (A.1), we have that the transmission power in dBm in the kth sub-carrier necessary to assure a coverage probability of ξ % in the cell edge is given by

$$p_k = 128.1 + 37.6 \log_{10} R + p^{\text{noise}} + \zeta^{\text{th}} - \sqrt{2}\sigma \cdot \text{erfinv}\,(1 - 2\xi)\,, \tag{A.5}$$

where R is the macrocell radius in km and erfinv is the inverse error function. Once we have the needed transmission power on each sub-carrier, we can calculate the MBS total transmit power.

A.4.2. Signal-to-Noise Ratio Calculation

Taking into account the propagation losses described in section A.3, we have that the instantaneous received power of the jth MT in the kth sub-carrier in dBm is given by

$$p_{j,k}^{\text{rx}} = p_k - L_j^{\text{path}} - L_j^{\text{shadow}} + L_{j,k}^{\text{fast}}, \tag{A.6}$$

where p_k is the MBS transmission power on the kth sub-carrier in dBm, and L_j^{path}, L_j^{shadow} and $L_{j,k}^{\text{fast}}$ are the propagation losses due to path loss, shadowing and fast fading, respectively.

For didactic reasons, let us define the linear Channel-to-Noise Ratio (CNR) $\gamma_{j,k}$ of the jth MT in the kth sub-carrier given by the expression below:

$$\gamma_{j,k} = 10^{\left(L_j^{\text{path}} - L_j^{\text{shadow}} + L_{j,k}^{\text{fast}} - p^{\text{noise}}\right)/10}, \tag{A.7}$$

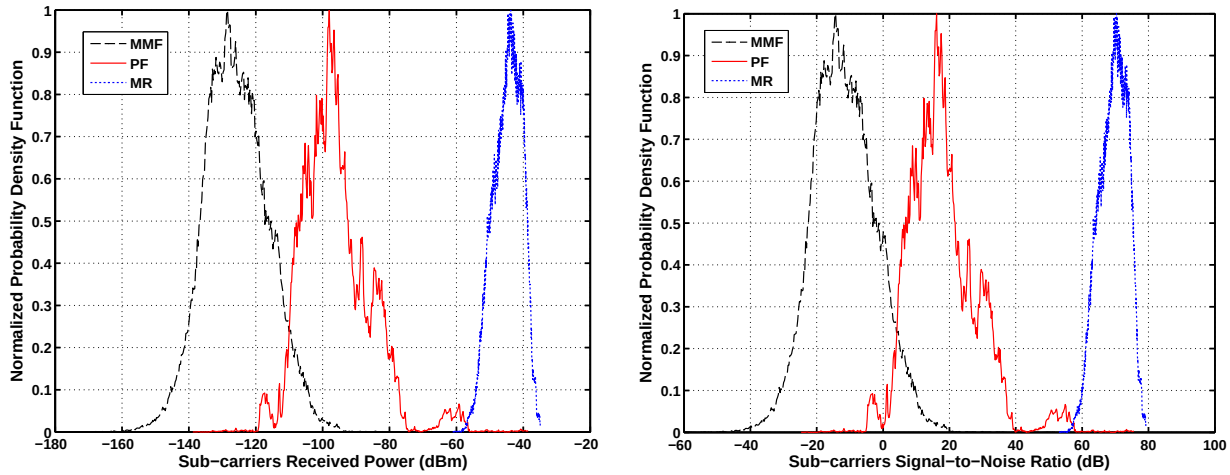
After that, the linear SNR $\zeta_{j,k}$ can be readily calculated by

$$\zeta_{j,k} = p_k \cdot \gamma_{j,k}. \tag{A.8}$$

In order to validate the calculation of the received power and SNR of the sub-carriers, Fig. A.6 is plotted. This figure depicts the normalized PDFs of some samples of the sub-carriers' received power and SNR taken from the simulations using different sub-carrier assignment algorithms. The following simulation parameters were considered in this example: macrocell radius $R = 500$ m, total MBS transmit power of 1 W, $K = 192$ sub-carriers, effective sub-carrier bandwidth $\Delta f = 14$ kHz, noise figure $\vartheta = 9$ dB, and shadowing standard deviation $\sigma = 8$ dB. The considered sub-carrier assignment algorithms are Max-Min Fairness (MMF), Proportional Fairness (PF) and Max-Rate (MR) (for more details see section 4.4.1 of chapter 4).

As expected, the PDFs of the received power and SNR have the same form, since their calculations differ only because of the noise power, which is a constant. Moreover, it can be noticed that the range of values are in accordance with expressions (A.6) and (A.8) and the simulation parameters considered in the example. Finally, it can be seen that the MR policy shows the highest values of received power and SNR, followed by PF and at last MMF. This is due to the fact that MR is

an opportunistic policy that assigns the sub-carriers to the users with better channel conditions, which results in higher received power, and consequently, higher SNR. Not only does PF take into account the CSI of the users but also their throughput, which decreases the importance of the efficiency in the resource usage, and so provides lower received power and SNR. MMF does not take into account the channel quality, so it presents the worse results regarding both metrics.



(a) Received power

(b) Signal-to-Noise Ratio

Figure A.6.: Normalized Probability Density Functions of sub-carriers' received power and SNR samples taken from simulations using different sub-carrier assignment algorithms

A.5. Link Adaptation

Depending on the channel condition, an appropriate number of bits is transmitted on each sub-carrier. This is accomplished by the link adaptation procedure. The link adaptation curve can be discrete or continuous. The former is the one used in current practical systems, while the latter is widely used in the literature because of their simplicity and ability to provide coherent results that are close to the ones achieved with the discrete curve. Furthermore, continuous link adaptation is based on the well-known theoretical Shannon's capacity formula, which is being increasingly approximated due to the use of more sophisticated physical layer techniques. Assuming continuous link adaptation, the achievable transmission rate in bits/s/Hz of the kth sub-carrier assigned to the jth MT is given by (A.9) below [112].

$$c_{j,k} = \log_2\left(1 + p_k \cdot \frac{\gamma_{j,k}}{\Gamma}\right), \tag{A.9}$$

where p_k is the transmit power of the kth sub-carrier and $\gamma_{j,k}/\Gamma$ is the effective CNR, i.e. channel quality, of the kth sub-carrier with respect to the jth MT. The constant Γ is called SNR gap, which

indicates the difference between the theoretical limit and the SNR needed to achieve a certain data transmission rate for a practical system [112]. This constant is dependent on the target Bit Error Rate (BER) and, considering an M-level Quadrature Amplitude Modulation (QAM), its value is given by $\Gamma = -[\ln(5 \cdot \mathrm{BER})]/1.5$. In this way, the continuous link adaptation given by (A.9) is a realistic Shannon's capacity model.

If discrete link adaptation is to be used, a quantization operator needs to be considered. Assuming discrete modulation levels $m = \{2, 4, 6, ...\}$, this quantization procedure is defined as $2\left\lfloor \frac{c_{j,k}}{2} \right\rfloor$, where $\lfloor x \rfloor$ is an operator that returns the largest integer less than x. The link adaptation curve used in the simulations is depicted in Fig. A.7.

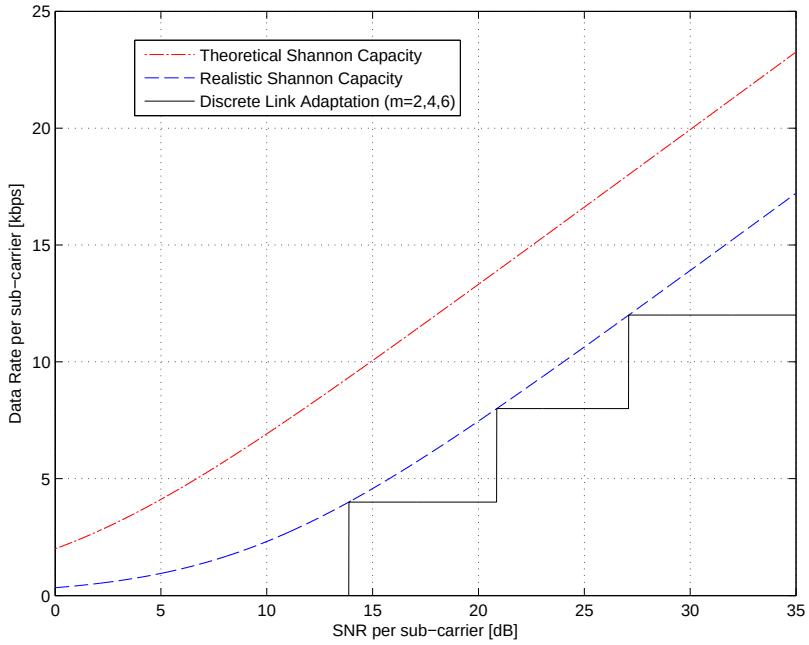


Figure A.7.: Link adaptation curve

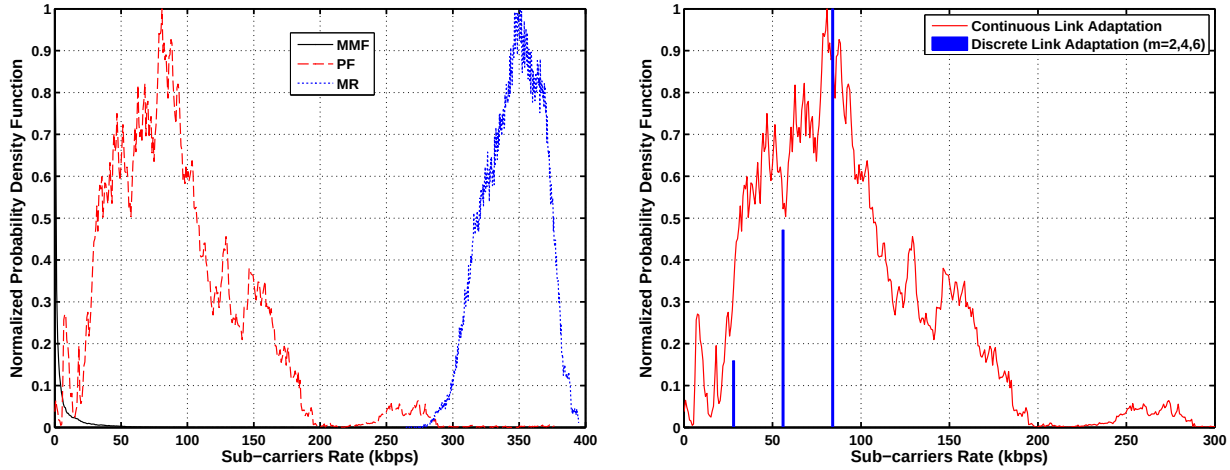
Once we have the achievable transmission rate per Hertz of each sub-carrier, the downlink data transmission rate for each MT can be calculated. In the sub-carrier assignment process, we assume that each sub-carrier can only be assigned to one single MT. Assuming that a sub-carrier set $\mathcal{S}_j$ is assigned to the jth MT, its transmission rate is calculated as

$$R_j = \sum_{k \in \mathcal{S}_j} r_{j,k} = \sum_{k \in \mathcal{S}_j} c_{j,k} \cdot \Delta f \tag{A.10}$$

where $c_{j,k}$ is the channel capacity per Hertz of the kth sub-carrier assigned to the jth MT given by (A.9) and Δf is the sub-carrier bandwidth. The total rate of the system is the sum of R_j among all MTs, as indicated below:

$$R_{\mathrm{cell}} = \sum_{j=1}^{J} R_j. \tag{A.11}$$

Fig. A.8 presents the normalized PDFs of the transmission rates in the simulation example shown in section A.4.2. Thus, the values of the data rates is the outcome of the link adaptation procedure considering the SNR values presented in Fig. A.6(b). The link adaptation curve used in this example was the realistic Shannon capacity curve presented in Fig. A.7. Fig. A.8(a) compares the rates provided by three different sub-carrier assignment algorithms, namely MMF, PF and MR. The results are coherent because it can be observed that the lower the SNR, the lower the data rate. For instance, since the MMF policy does not take into account the channel quality, it is inefficient and provides very low transmission rates. Fig. A.8(b) analyzes the PDF of the PF algorithm comparing the outcome of continuous (realistic Shannon capacity) and discrete link adaptation curves (see Fig. A.7). For the case of discrete link adaptation, three modulation schemes were considered: Quadrature Phase Shift Keying (QPSK), 16-QAM and 64-QAM ($m = \{2, 4, 6\}$, respectively).



(a) Data Rate

(b) Data Rate considering PF DSA

Figure A.8.: Normalized Probability Density Functions of sub-carriers' data rate samples taken from simulations using different sub-carrier assignment algorithms

A.6. Traffic Models

The simulator used in this thesis adopts two traffic models, where each of them are suitable for Non-Real Time (NRT) and Real Time (RT) services, respectively. NRT services, such as World Wide Web (WWW) and File Transfer Protocol (FTP), are not delay-sensitive and require an overall high throughput, while RT services, such as Voice over IP (VoIP) and video conference, require a low and bounded delay.

The traffic model used for NRT services is the full-buffer model. It assumes that the buffers of the MTs located in the MBS always have data to be transmitted. It is an assumption widely used in many works in the literature that evaluate Radio Resource Allocation (RRA) techniques

for OFDMA-based systems. The idea behind this model is that some NRT multimedia services to be provided by next-generation mobile broadband systems require the transfer of large amounts of data, for example high definition images, music and video. Furthermore, the full-buffer model characterizes a worst-case scenario regarding system load. Since all RRA techniques suitable for NRT services studied in this thesis consider the same model, the relative performance comparison remains valid.

We consider a simple traffic model for RT services, which consists on the regular generation of packets of 32 bytes every 2 ms. The delay of each packet is accounted and it must respect the RT delay budget of the radio access network. If the packet arrives at the receiver later than this delay budget, it is discarded.

A.7. Performance Metrics

A.7.1. Throughput

The throughput (average data rate) of the jth MT is calculated using a low-pass exponential filtering, as indicated in (A.12).

$$T_j[n] = (1 - f_{\text{thru}}) \cdot T_j[n-1] + f_{\text{thru}} \cdot R_j[n] \tag{A.12}$$

where $R_j[n]$ is the instantaneous data rate of the jth MT given by (A.10) and f_{thru} is a filtering constant.

A.7.2. Delay

The Head-Of-Line (HOL) delay is the time the oldest packet in the user buffer has to wait before accessing the wireless channel. Considering a generic MT j, it can be calculated approximately by the following recursive equation:

$$d_j^{\text{hol}}[n+1] = d_j^{\text{hol}}[n] + \frac{b_j^{\text{hol}}[n] - R_j[n] \cdot t_{\text{tti}}}{T_j[n-1]} \tag{A.13}$$

where $b_j^{\text{hol}}[n]$ is the current number of bits in the HOL packet, t_{tti} is the duration of the TTI in seconds, $T_j[n-1]$ is the average data rate (throughput) given by (A.12) up to the previous transmission interval and $R_j[n]$ is the instantaneous achievable transmission rate given by (A.10).

If the jth MT has not been served by any sub-carrier in the nth TTI, $R_j[n]$ is equal to zero and the HOL delay is incremented. This delay increment is calculated assuming that the remaining bits of the HOL packet will be transmitted using a rate equal to the throughput experienced so far by the MT. For sake of simplicity, we assume that the packets' interarrival time is equal to the TTI duration.

If the instantaneous transmission rate is such that all remaining bits of the HOL packet are transmitted in the current TTI, the HOL delay remains constant because the previous packet in the buffer will be the HOL packet now.

Finally, the HOL delay is decremented when the instantaneous achievable transmission rate is high enough to transmit the remaining bits of the HOL packet and some bits of the preceding packets in the queue.

In order to perform fairness control based on the HOL delay, it is advised to perform an exponential filtering, as indicated below:

$$d_j^{\text{hol,filt}}[n] = \left(1 - f_{\text{delay}}\right) \cdot d_j^{\text{hol,filt}}[n-1] + f_{\text{delay}} \cdot d_j^{\text{hol}}[n], \tag{A.14}$$

where $d_j^{\text{hol}}[n]$ is the current HOL delay of the jth MT and f_{delay} is a filtering constant.

A.7.3. Fairness Index

The well-known Jain's fairness index is a quantitative fairness measure originally proposed by Jain et al. in [86]. The general Jain's fairness function is independent of the allocation metric being used. Particularized versions of this fairness index considering different allocation metrics are widely used throughout this thesis.

Considering a generic allocation metric $\vec{x} = [x_1, \ldots, x_j, \ldots, x_J]$, the Jain's fairness function can be interpreted in terms of the variance and expected value of $\vec{x}$, as follows:

$$F(x) = \frac{[E(\vec{x})]^2}{E(\vec{x}^2)} = \frac{1}{1 + \frac{Var(\vec{x})}{[E(\vec{x})]^2}} = \frac{\left(\sum_{j=1}^{J} x_j\right)^2}{J \cdot \sum_{j=1}^{J} (x_j)^2}. \tag{A.15}$$

This index has some interesting properties [86]:

- The fairness is bounded between 0 and 1 (or 0% and 100%). A totally fair allocation (with all x_j's equal) has a fairness of 1, while a totally unfair allocation (with all resources given to only one user) has a fairness of $1/J$, which is 0 in the limit as $J \to \infty$.
- The fairness is independent of scale, i.e. unit of measurement does not matter.
- The fairness is a continuous function. Any slight change in allocation is reflected in the fairness.
- If only Q of J users share the resources equally with the remaining $J - Q$ users not receiving any resource, then the fairness is Q/J.

A.7.4. User Satisfaction

The definition of satisfaction depends on the type of service that the MT uses, i.e. NRT or RT service.

A NRT user is considered satisfied if its session throughput is higher than a threshold. The session duration depends on the time span of each independent simulation snapshot. If the RRA techniques being evaluated are based on instantaneous data rates, the duration of the simulation snapshot, and consequently the user session, is the TTI. If RRA strategies based on average data rates (throughput) are being assessed, the durations of the user session and the snapshots are longer.

A RT user is considered satisfied if its Frame Erasure Rate (FER) is lower than a threshold. In our simulation model, we assume that a frame is lost if a packet arrives at the MT receiver later than the delay budget of the RT service.

A.7.5. CPU Time

The processing time required by the RRA techniques to find the allocation solutions is a measure of algorithmic complexity. In order to compare the different techniques in a fair manner, the simulations used to calculate this performance metric were executed in the same machine. The hardware and software configuration of this machine is presented below.

- Hardware
  - ACPI Multiprocessor x64-based PC;
  - Intel Core 2 Quad CPU;
  - Q6600, 2.40 GHz;
  - 7.92 GB of RAM.
- Software
  - Microsoft Windows Server 2003 R2;
  - Enterprise x64 Edition, Service Pack 2.

A.8. Simulator Structure

A discrete-time system-level simulator was developed using the technical computing software package Matlab. The block diagram of the simulator structure is presented in Fig. A.9.

The main steps of the simulator are listed below:

1. **Simulation scenario**: Input of main parameters for scenario characterization, such as: number of users in the macrocell, type of service, RRA technique, sub-carrier assignment and power allocation algorithms, simulation duration, etc.
2. **Initialization of simulation parameters**.
3. **Generation of link adaptation curve**: Generate the curve presented in Fig. A.7.

4. **Generation of path loss and shadowing**: Generate the path loss and large-scale fading (shadowing) for the whole simulation according to sections A.3.1 and A.3.2, respectively.

5. **Loop of iterations (TTIs)**

   a) **Generation of fast fading**: Generate the fast fading samples in the time and frequency domain according to section A.3.3.

   b) **Generation of traffic**: Generation of the packets to be transmitted according to a proper traffic model, as explained in section A.6.

   c) **Fairness calculation**: Calculate the user fairness index (allocation metric) and the cell fairness index as shown in section A.7.3. This information will be needed by the RRA techniques proposed in this thesis.

   d) **Sub-carrier assignment**: Execute the corresponding sub-carrier assignment algorithm according to the RRA technique being evaluated.

   e) **Power allocation**: Define the power level of each sub-carrier depending on the chosen RRA technique.

   f) **Rate calculation**: After the sub-carrier assignment and the power allocation have been defined, we are able to calculate the SNR of all sub-carriers in the system. After calculating the SNR, we perform the link adaptation procedure according to section A.5, and calculate the transmission rates of all users.

   g) **Transmission**: Based on the transmission rates calculated in the step above, we are able to know the amount of bits that can be transmitted for each user. These bits are removed from the packets of the users' buffers.

   h) **Test of number of iterations**: Test if the number of iterations has reached the desired value. If it is true, the simulation has ended and we can go to the final step below. Otherwise, the loop of iterations continue.

6. **Save results**: In this final step, all simulation results are saved.

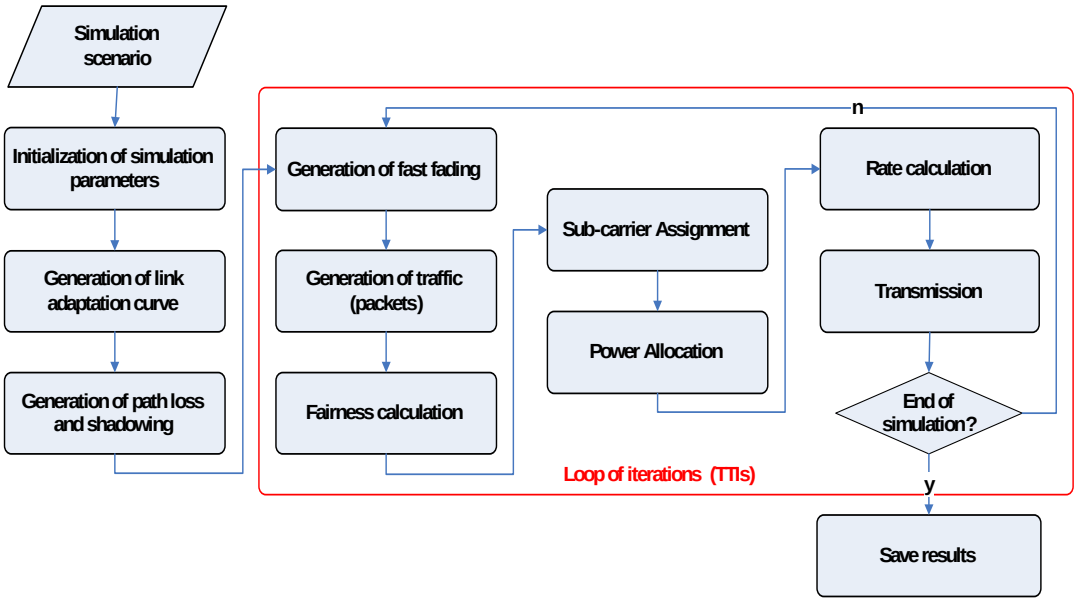


Figure A.9.: System-level simulator structure

Appendix B

Classic Rate Adaptive Policies

In this appendix, three classic rate adaptive policies are described. We have chosen three references, which are in our opinion the most representative works on each of the three Radio Resource Allocation (RRA) approaches considered in chapter 3: *Sum Rate Maximization (SRM)* [61] (section B.2), *Sum Rate Maximization with Proportional Rate Constraints (SRM-P)* [87] (section B.3) and *Max-Min Rate (MMR)* [64] (section B.4). Notice that in section B.4, a generalization of the original problem described in [64] is proposed in order to take into account proportional rate constraints. The new rate adaptive problem is called *Max-Min Rate with Proportional Rate Constraints (MMR-P)*. All classic rate adaptive policies share a common optimization background, which will be formulated in section B.1.

B.1. General Optimization Problem Formulation

All classic rate adaptive policies considered in this thesis share a common framework that can be formulated as an RRA optimization problem, as follows:

$$\max_{\rho_{j,k},p_k} \; f\left(\rho_{j,k},p_k\right), \tag{B.1}$$

$$\text{subject to } \rho_{j,k} = \{0,1\}, \quad \forall j,k, \tag{B.2}$$

$$\sum_{j=1}^{J} \rho_{j,k} = 1, \quad \forall k, \tag{B.3}$$

$$p_k \geq 0, \quad \forall k, \tag{B.4}$$

$$\sum_{k=1}^{K} p_k \leq P_{\text{total}}, \tag{B.5}$$

where J and K are the total number of Mobile Terminals (MTs) and sub-carriers, respectively; $\rho_{j,k}$ is the connection indicator, whose value 0 or 1 indicates whether the kth sub-carrier is assigned to the jth MT or not; p_k is the power of the kth sub-carrier; P_{total} is the Macrocell Base Station (MBS) total transmit power; and $f\left(\rho_{j,k}, p_k\right)$ is a general objective function that depends on the sub-carriers assignment and power allocation.

Constraints (B.2) and (B.3) say that each sub-carrier must be assigned to only one user at any instant of time, while constraints (B.4) and (B.5) state that the sub-carriers' powers must be non-negative and the sum of powers among all sub-carriers must be lower or equal to the MBS total transmit power.

The rate adaptive optimization (B.1)-(B.5) is a mixed binary integer programming problem, because it involves both binary variables $\rho_{j,k}$ and continuous variables p_k. This problem is not convex because the integer constraint (B.2) makes the feasible set nonconvex.

The general RRA problem formulated above is particularized in different classic rate adaptive policies by considering different objective functions and/or new optimization constraints, as will be shown in the following.

B.2. Sum Rate Maximization (SRM)

The objective function of the SRM policy is the maximization of system capacity, as indicated below:

$$\max_{\rho_{j,k},p_k} \sum_{j=1}^{J}\sum_{k=1}^{K} \rho_{j,k}\cdot\log_2\left(1+p_k\cdot\gamma_{j,k}\right), \tag{B.6}$$

where $\gamma_{j,k}$ is the Channel-to-Noise Ratio (CNR) of the kth sub-carrier with respect to the jth MT.

The problem (B.1)-(B.5) considering the objective function (B.6) has been solved in [61] using a two-step approach: 1) Dynamic Sub-carrier Assignment (DSA) for users; and 2) Adaptive Power Allocation (APA) for sub-carriers. The first step states that a sub-carrier should be assigned to only one user who has the best channel gain for that sub-carrier, as indicated in Algorithm B.1. The second step consists of performing waterfilling power allocation over all previously assigned sub-carriers, as shown in Algorithm B.2.

On one hand, such a solution maximizes the cell throughput but on the other hand is extremely unfair tending to privilege the users that are closest to the MBS, i.e. users with best channel conditions, and neglecting all the others.

Algorithm B.1 Dynamic Sub-carrier Assignment (DSA) of the SRM technique

Initialization

$\mathcal{M} \leftarrow \{1, 2, 3, \cdots J\}$ // Users set
$\mathcal{S} \leftarrow \{1, 2, 3, \cdots K\}$ // Sub-carriers set
for all $j \in \mathcal{M}$ and $k \in \mathcal{S}$ **do**
  $\rho_{j,k} \leftarrow 0$ // Reset connection matrix
  $\mathcal{S}_j \leftarrow \varnothing$ // Reset user sub-carrier subset
end for

Sub-carrier assignment

for all $k \in \mathcal{S}$ **do**
  $j^* \leftarrow \arg\max_j \{\gamma_{j,k}\}$ // Find user with maximum channel quality on sub-carrier k
  $\rho_{j^*,k} \leftarrow 1$ // Set the connection
  $\mathcal{S}_{j^*} \leftarrow \mathcal{S}_{j^*} \bigcup \{k\}$ // Update user sub-carrier subset
end for

Algorithm B.2 Adaptive Power Allocation (APA) of the SRM technique

Initialization

$\mathcal{S} \leftarrow \{1, 2, 3, \cdots K\}$ // Sub-carriers set

Power allocation

// After sub-carrier assignment is determined by Algorithm B.1, multi-user waterfilling is performed over all sub-carriers
for all $k \in \mathcal{S}$ **do**
  $p_k = \left[\mu - \frac{1}{\gamma_{j,k}[n]}\right]^+$ // μ is the water level and the operator $[\cdot]^+$ is defined as $[x]^+ \triangleq \max(0, x)$.
end for

B.3. Sum Rate Maximization with Proportional Rate Constraints (SRM-P)

Trying to balance the trade-off between capacity and fairness, reference [87] considered the optimization problem initially proposed by [85], which we call SRM-P. The SRM-P problem has the same objective function (B.6) and the same constraints (B.2)-(B.5) of the common RRA problem formulated before with the addition of a set of nonlinear constraints into the optimization problem, as indicated below:

$$R_i : R_j = \lambda_i : \lambda_j, \quad \forall i, j = 1 : J, i \neq j. \tag{B.7}$$

Constraint (B.7) states that the users' rates must follow the proportional rate requirements for each channel realization, which ensures the rates of different users to be proportional in any time

scale of interest. The SRM-P problem is not convex due to the integer constraint of the connection indicator $\rho_{j,k}$ and the nonlinear equality constraints shown in (B.7).

The sub-optimum solution proposed by [85] used the two-step approach (DSA and APA). In the first step, following the heuristic approach taken in [64], the sub-carriers are allocated trying to comply as much as possible with the proportional rate constraints and assuming a uniform power distribution. In the second step, having fixed the sub-carrier assignment, an optimal power allocation is performed so that the proportional rate constraints are met exactly. However, the algorithm proposed in [85] involves solving nonlinear equations, which requires computationally expensive iterative operations, and thus it is not suitable for a cost-effective real-time implementation. Reference [87] extends the work in [85] by developing a sub-carrier assignment scheme that linearizes the power allocation problem while achieving approximate rate proportionality. This sub-carrier assignment scheme is shown in Algorithm B.3.

After the sub-carrier assignment described in Algorithm B.3, the SRM-P optimization problem is simplified into a maximization over continuous variables p_k, which can be solved optimally. Furthermore, as a consequence of the sub-carrier assignment scheme, we have that

$$N_1 : N_2 : \cdots : N_J \approx \lambda_1 : \lambda_2 : \cdots : \lambda_J, \tag{B.8}$$

where N_j is the number of sub-carriers reserved to jth user. This approximation gets tighter as $K \to \infty$ and $K \gg J$. This is a reasonable assumption for current wireless systems that use Orthogonal Frequency Division Multiple Access (OFDMA).

The approximation in (B.8) can be done because in practical systems, adherence to the proportionality constraints does not need to be strictly enforced [87]. A soft guarantee of rough proportionality is acceptable as long as the capacity is maximized and the algorithm complexity is low. This approximation can be used to relax constraint B.7 in the SRM-P optimization problem to

$$R_i : R_j = N_i : N_j, \quad \forall i, j = 1 : J, i \neq j. \tag{B.9}$$

With this simplification, the resulting power allocation problem is reduced to a solution of simultaneous linear equations, which can be optimally found with much less computational time (see Algorithm B.4). The power allocation is performed in two steps: power allocation among users (lines 3-13) and power allocation across sub-carriers per user (lines 14-19). The first step solves the set of simultaneous linear equations and assigns the total power P_j for each user in order to maximize the capacity while enforcing the rate proportionality. The second step performs waterfilling and assigns the powers p_k for each user's sub-carriers subject to his total power constraint P_j.

Algorithm B.3 Dynamic Sub-carrier Assignment (DSA) of the SRM-P technique

Initialization

$\mathcal{M} \leftarrow \{1, 2, 3, \cdots J\}$ // Users set
$\mathcal{S} \leftarrow \{1, 2, 3, \cdots K\}$ // Sub-carriers set
for all $j \in \mathcal{M}$ and $k \in \mathcal{S}$ **do**
  $\rho_{j,k} \leftarrow 0$ // Reset connection matrix
  $\mathcal{S}_j \leftarrow \varnothing$ // Reset user sub-carrier subset
  $R_j \leftarrow 0$ // Reset user rate
  $N_j \leftarrow \lfloor \lambda_j K \rfloor$ // Number of sub-carriers reserved to user j
end for
$N_{\text{res}} \leftarrow K - \sum_{j=1}^{J} N_j$ // Number of unallocated sub-carriers

Sub-carrier assignment

for all $j \in \mathcal{M}$ **do**
  Sort $\gamma_{j,k}$ in ascending order
  $k^* \leftarrow \arg\max_k \{\gamma_{j,k}\}, \forall k \in \mathcal{S}$ // Find available sub-carrier with maximum channel quality on user j
  $\rho_{j,k^*} \leftarrow 1$ // Set the connection
  $\mathcal{S}_j \leftarrow \mathcal{S}_j \cup \{k^*\}$ // Update user sub-carrier subset
  $N_j \leftarrow N_j - 1$ // Update number of sub-carriers reserved to user j
  $\mathcal{S} \leftarrow \mathcal{S} \setminus \{k^*\}$ // Update set of available sub-carriers
  $R_j \leftarrow R_j + r_{j,k^*}$ // Update user rate considering equal power allocation
end for
while $|\mathcal{S}| > N_{\text{res}}$ **do** // $|\cdot|$ is the cardinality operator
  $j^* \leftarrow \arg\min_j \{R_j/\lambda_j\}, \forall j \in \mathcal{M}$ // Find available user with minimum proportional rate
  $k^* \leftarrow \arg\max_k \{\gamma_{j^*,k}\}, \forall k \in \mathcal{S}$ // Find available sub-carrier with maximum channel quality on user j^*
  if $N_{j^*} > 0$ **then**
    $\rho_{j^*,k^*} \leftarrow 1$ // Set the connection
    $\mathcal{S}_{j^*} \leftarrow \mathcal{S}_{j^*} \cup \{k^*\}$ // Update user sub-carrier subset
    $N_{j^*} \leftarrow N_{j^*} - 1$ // Update number of sub-carriers reserved to user j^*
    $\mathcal{S} \leftarrow \mathcal{S} \setminus \{k^*\}$ // Update set of available sub-carriers
    $R_{j^*} \leftarrow R_{j^*} + r_{j^*,k^*}$ // Update user rate
  else
    $\mathcal{M} \leftarrow \mathcal{M} \setminus \{j^*\}$ // Update set of available users
  end if
end while
$\mathcal{M} \leftarrow \{1, 2, 3, \cdots J\}$ // Reset users set
for all $k \in \mathcal{S}$ **do** // Set of unallocated sub-carriers
  $j^* \leftarrow \arg\max_j \{\gamma_{j,k}\}$ // Find user with maximum channel quality on sub-carrier k
  $\rho_{j^*,k} \leftarrow 1$ // Set the connection
  $\mathcal{S}_{j^*} \leftarrow \mathcal{S}_{j^*} \cup \{k\}$ // Update user sub-carrier subset
  $R_{j^*} \leftarrow R_{j^*} + r_{j^*,k}$ // Update user rate
  $\mathcal{M} \leftarrow \mathcal{M} \setminus \{j^*\}$ // Update set of available users
end for

Algorithm B.4 Adaptive Power Allocation (APA) of the SRM-P technique

Initialization

$\mathcal{M} \leftarrow \{1, 2, 3, \cdots J\}$ // Users set

$\mathcal{S} \leftarrow \{1, 2, 3, \cdots K\}$ // Sub-carriers set

Power allocation among users

// After sub-carrier assignment is determined by Algorithm B.3, power allocation per user is performed

for all $j \in \mathcal{M}$ **do**

  Sort $\gamma_{j,k}$ in ascending order

  $V_j \leftarrow \sum_{k=2}^{N_j} \frac{\gamma_{j,k}-\gamma_{j,1}}{\gamma_{j,k}\cdot\gamma_{j,1}}$ // N_j is the number of sub-carriers assigned to user j

  $W_j \leftarrow \left(\prod_{k=2}^{N_j} \frac{\gamma_{j,k}}{\gamma_{j,1}}\right)^{\frac{1}{N_j}}$

  $a_{j,j} \leftarrow -\frac{N_1}{N_j} \cdot \frac{\gamma_{j,1}\cdot W_j}{\gamma_{1,1}\cdot W_1}$

  $b_j \leftarrow \frac{N_1}{\gamma_{1,1}} \left(W_j - W_1 + \frac{\gamma_{1,1}\cdot V_1 \cdot W_1}{N_1} + \frac{\gamma_{j,1}\cdot V_j \cdot W_j}{N_j}\right)$

end for

$P_1 \leftarrow \left(P_{\text{total}} - \sum_{j=2}^{J} \frac{b_j}{a_{j,j}}\right) / \left(1 - \sum_{j=2}^{J} \frac{1}{a_{j,j}}\right)$ // Power allocated to the first user

for $j = 2$ to J **do**

  $P_j \leftarrow (b_j - P_1)/a_{j,j}$ // Power allocated to other users

end for

Power allocation across sub-carriers per user

// After calculating the power per user, waterfilling is performed among the sub-carriers of each user

for all $j \in \mathcal{M}$ **do**

  $p_1 \leftarrow \frac{P_j - V_j}{N_j}$ // Power allocated to the 1st sub-carrier of user j

  for $k = 2$ to K **do**

    $p_k \leftarrow p_1 + \frac{\gamma_{j,k}-\gamma_{j,1}}{\gamma_{j,k}\cdot\gamma_{j,1}}$ // Power allocated to other sub-carriers of user j

  end for

end for

B.4. Max-Min Rate with Proportional Rate Constraints (MMR-P)

In order to increase user fairness in an OFDMA system, reference [64] proposed the MMR optimization problem, whose solution maximizes the minimum user's data rate. Maximizing the worst user's capacity, it is assured that all users achieve a similar data rate, and hence maximum fairness among users is provided. The original MMR optimization problem proposed in [64] was not convex due to the constraint of disjoint sub-carriers set, i.e. a sub-carrier could not be shared among the users at the same time. Reference [64] showed that it was possible to formulate a convex problem if this constraint was relaxed. However, the authors concluded that the optimal solution requires an intensive computation due to the recursive nature of solving a convex optimization problem.

This thesis extends the work in [64] and proposes the MMR-P optimization problem. The mathematical formulation of the proposed MMR-P policy has the same constraints of SRM-P, also including the proportional rate constraints given by (B.7). This constraint allows maximum fairness for the general case where there are several service classes with varying priorities. However, the MMR-P problem has a different objective function, which is the maximization of the minimum user's capacity and is mathematically expressed by

$$\max_{\rho_{j,k},p_k} \min_{j} \quad \sum_{k=1}^{K} \rho_{j,k} \cdot \log_2\left(1 + p_k \cdot \gamma_{j,k}\right). \tag{B.10}$$

Like the original MMR policy, the MMR-P optimization problem is not convex and requires a heuristic sub-optimum solution. Based on [64], the proposed solution comprises separate sub-carrier assignment and power allocation. Algorithm B.5 presents the pseudo-code of the MMR-P DSA algorithm, which adds to the original solution presented in [64] the consideration of the proportional rate constraints. Lines 8-14 of Algorithm B.5 show that initially each user can choose his best sub-carrier. On the next step shown by lines 15-22, the users with minimum proportional rate have priority and can choose their best sub-carriers among the remaining sub-carriers. This is a heuristic algorithm with small complexity that provides a sub-optimum solution close to the optimal [64]. Following [64], we also consider Equal Power Allocation (EPA) in the power allocation step, as indicated in Algorithm B.6.

Algorithm B.5 Dynamic Sub-carrier Assignment (DSA) of the MMR-P technique

Initialization

$\mathcal{M} \leftarrow \{1, 2, 3, \cdots J\}$ // Users set
$\mathcal{S} \leftarrow \{1, 2, 3, \cdots K\}$ // Sub-carriers set
for all $j \in \mathcal{M}$ and $k \in \mathcal{S}$ **do**
    $\rho_{j,k} \leftarrow 0$ // Reset connection matrix
    $\mathcal{S}_j \leftarrow \varnothing$ // Reset user sub-carrier subset
    $R_j \leftarrow 0$ // Reset user rate
end for

Sub-carrier assignment

for all $j \in \mathcal{M}$ **do**
    $k^* \leftarrow \arg\max_k \{\gamma_{j,k}\}, \forall k \in \mathcal{S}$ // Find available sub-carrier with maximum channel quality on user j
    $\rho_{j,k^*} \leftarrow 1$ // Set the connection
    $\mathcal{S}_j \leftarrow \mathcal{S}_j \bigcup \{k^*\}$ // Update user sub-carrier subset
    $\mathcal{S} \leftarrow \mathcal{S} \setminus \{k^*\}$ // Update set of available sub-carriers
    $R_j \leftarrow R_j + r_{j,k^*}$ // Update user rate considering equal power allocation
end for
while $\mathcal{S} \neq \varnothing$ **do**
    $j^* \leftarrow \arg\min_j \{R_j/\lambda_j\}$ // Find user with minimum proportional rate
    $k^* \leftarrow \arg\max_k \{\gamma_{j^*,k}\}, \forall k \in \mathcal{S}$ // Find available sub-carrier with maximum channel quality on user j^*
    $\rho_{j^*,k^*} \leftarrow 1$ // Set the connection
    $\mathcal{S}_{j^*} \leftarrow \mathcal{S}_{j^*} \bigcup \{k^*\}$ // Update user sub-carrier subset
    $R_{j^*} \leftarrow R_{j^*} + r_{j^*,k^*}$ // Update user rate
    $\mathcal{S} \leftarrow \mathcal{S} \setminus \{k^*\}$ // Update set of available sub-carriers
end while

Algorithm B.6 Equal Power Allocation (EPA) for the MMR-P technique

Initialization

$\mathcal{S} \leftarrow \{1, 2, 3, \cdots K\}$ // Sub-carriers set

Power allocation

// After sub-carrier assignment is determined by Algorithm B.5, equal power allocation is performed over all sub-carriers
for all $k \in \mathcal{S}$ **do**
    $p_k = P_{\text{total}}/K$
end for

Appendix C

Modified Alpha-Rule and Beta-Rule Frameworks

Two variants of the original utility-based alpha-rule and beta-rule frameworks, which were proposed in chapter 4 and are suitable for Non-Real Time (NRT) and Real Time (RT) services, respectively, are presented in sections C.1 and C.2. These two novel approaches are based on a new range of values of the parameters α and β and a new way that the Dynamic Sub-carrier Assignment (DSA) algorithms balance the efficiency and fairness in their priority functions. Furthermore, we present in sections C.1.2 and C.1.3 two case-studies where the modified alpha-rule framework is applied to solve two different Radio Resource Allocation (RRA) problems: control of the trade-off between resource efficiency and user fairness; and maximization of user satisfaction for several system loads.

C.1. Modified Alpha-Rule Framework

C.1.1. Framework Description

In the original utility-based alpha-rule framework proposed in section 4.4.1 of chapter 4, the parameter α is defined over a large range of values, i.e. $\alpha \in [0, \infty)$. We showed in chapter 4 that considering a maximum value of $\alpha = 10$ in the feedback control loop of the Adaptive Throughput-Based Fairness (ATF) policy was sufficient to provide a controllable trade-off between resource efficiency and user fairness in a scenario with NRT services. However, we decided to investigate a way to accelerate the convergence of the Cell Fairness Index (CFI) by considering a more limited range of values for α in the control loop, such as $\alpha \in [0, 1]$. In order to use this limited range and still be able to have a fully controllable trade-off, a modification on the priority function of the DSA algorithm was needed, as will be explained in the following.

As presented in chapter 4, after using a mathematical development based on Utility Theory, we found out that the DSA algorithm of the original utility-based alpha-rule framework is based on

the following expression:

$$m(k,n) = \arg\max_{j} \left\{ \frac{1}{T_j[n-1]^{\alpha}} \cdot c_{j,k}[n] \right\}, \tag{C.1}$$

which means that the Mobile Terminal (MT) $m(k,n)$ is chosen to transmit on the kth sub-carrier in the nth Transmission Time Interval (TTI) if it satisfies the condition above. We have that $T_j[n-1]$ is the throughput (average data rate) of the jth MT up to the previous TTI and $c_{j,k}[n]$ denotes the instantaneous achievable transmission efficiency of the jth MT on the kth sub-carrier (Shannon capacity given by (A.9) on appendix A) assuming equal power allocation per sub-carrier.

Using the new range of values for the parameter α, we propose to use the following DSA criterion:

$$m(k,n) = \arg\max_{j} \left\{ \frac{c_{j,k}[n]^{1-\alpha}}{T_j[n-1]^{\alpha}} \right\}. \tag{C.2}$$

Although expressions (C.2) and (C.1) have the same functional structure, i.e. they provide a balance between the efficiency indicator (channel quality) and the Quality of Service (QoS)/fairness indicator (throughput), the modified DSA algorithm does not follow rigourously the mathematical formulation advocated by Utility Theory, as the original DSA algorithm does. Expression (C.2) represents a heuristic policy only inspired by the utility-based policy given by (C.1).

As also shown in chapter 4, the Adaptive Power Allocation (APA) algorithm of the original utility-based alpha-rule framework has the solution in the form of a utility-based multi-level waterfilling given by

$$p_k^*[n] = \left[\mu \cdot \frac{1}{T_j[n-1]^{\alpha}} - \frac{\Gamma}{\gamma_{j,k}[n]} \right]^{+}, \quad \forall k \in \mathcal{S}_j \tag{C.3}$$

where $[x]^{+} \triangleq \max(0,x)$, $p_k^*[n]$ is the current optimal power allocated to the kth sub-carrier belonging to the jth MT ($k \in \mathcal{S}_j$), $\Gamma/\gamma_{j,k}[n]$ is the inverse of the effective Channel-to-Noise Ratio (CNR), i.e. channel quality, of the kth sub-carrier assigned to the jth MT at the nth TTI, and μ is a non-negative variable that represents the water-level of the waterfilling problem. The constant Γ is the so-called Signal-to-Noise Ratio (SNR) gap, which indicates the difference between the theoretical limit and the SNR needed to achieve a certain data transmission rate for a practical system.

Due to the structure of the waterfilling problem C.3, it is not possible to find a parametric power allocation expression based on the new range of $\alpha \in [0,1]$ that could provide a balance between the efficiency indicator (channel quality) and the QoS/fairness indicator (throughput). Therefore, we propose to use an Equal Power Allocation (EPA) in the modified alpha-rule framework, taking into account that no major differences between the joint and EPA approaches has been observed in the results reported in chapter 4.

Just like the original utility-based alpha-rule, the modified framework proposed above can be designed to work as different RRA policies depending on the value of the fairness controlling parameter α, achieving different performances in terms of resource efficiency and throughput-based fairness. The main characteristics of the modified alpha-rule and the four particular RRA policies contemplated by this framework are presented in Table C.1. The first three RRA strategies are well-known classic policies, namely Max-Rate (MR), Max-Min Fairness (MMF) and Proportional Fairness (PF), which were already described and evaluated in chapter 4. Although the Modified Adaptive Throughput-Based Fairness (M-ATF) policy uses a different DSA priority function (see expression (C.2)), the feedback control loop of the parameter α defined in the range $[0, 1]$ works in the same way as the original utility-based ATF policy (see section 4.4.1 of chapter 4).

In the following sections we present two applications of the modified alpha-rule framework in different RRA problems in macrocell Orthogonal Frequency Division Multiple Access (OFDMA) networks: management of the trade-off between resource efficiency and user fairness; and maximization of user satisfaction.

Table C.1.: Features of the modified alpha-rule framework

Policies	**Parameter α**	**DSA priority function**	**Characteristics**
MR	0	$c_{j,k}[n]$	High resource efficiency and low throughput-based fairness
PF	0.5	$\sqrt{\frac{c_{j,k}[n]}{T_j[n-1]}}$	Static trade-off between resource efficiency and throughput-based fairness
MMF	1	$\frac{1}{T_j[n-1]}$	Low resource efficiency and high throughput-based fairness
M-ATF	adaptive	$\frac{c_{j,k}[n]^{1-\alpha}}{T_j[n-1]^{\alpha}}$	Dynamic trade-off between resource efficiency and throughput-based fairness

C.1.2. Case-Study 1: Management of the Trade-Off between Resource Efficiency and User Fairness

The main simulation parameters used in the first case-study are presented in Table C.2.

Table C.2.: Simulation parameters for the evaluation of the modified alpha-rule framework (case study 1)

Parameter	Value
Number of cells	1
Maximum BS transmission power	20 W
Cell radius	500 m
MT speed	static
Number of sub-carriers	192
Sub-carrier bandwidth	15 kHz
Path loss	using (A.1)
Log normal shadowing standard dev.	8 dB
Small-scale fading	Typical Urban (TU)
Noise power per sub-carrier	-123.24 dBm
BER requirement	10^{-6}
Modulation schemes	QPSK, 16-QAM, 64-QAM (see section A.5)
Transmission Time Interval (TTI)	0.5 ms
Traffic model	NRT full buffer
Throughput filtering time constant (f_{thru})	50 (see (A.12))
User throughput requirement[a] (T_j^{req})	1.4 Mbps
Minimum α value	0
Maximum α value	1
M-ATF control time window	0.5 ms
M-ATF target fairness index ($\Phi_{\text{target}}^{\text{nrt}}$)	0.9
M-ATF step size (η_{nrt})	0.1
M-ATF filtering time constant	10
Simulation time span	30 s
Number of realizations for each point	10

[a] A MT is considered satisfied if its session throughput up to the present moment is higher than its throughput requirement.

Fig. C.1 shows the CFI calculated by equation (4.23) in chapter 4 for different number of users. The performance of the M-ATF technique is compared to the three classic RRA policies: MMF, PF and MR. In this simulation scenario, the Cell Fairness Target (CFT) is 0.9. It can be observed that M-ATF is successful at achieving its main objective, which is to guarantee a strict fairness distribution among the MTs. This is accomplished due to the feedback control loop that dynamically adapts the parameter α of the modified alpha-rule framework. The classic policies confirm their expected performance, as already seen in chapter 4. MMF provided the highest fairness, very close to the maximum value of 1, while MR was the most unfair strategy with a high variance on the fairness distribution for high cell loads. PF presented a good fairness distribution, close to the performance of MMF.

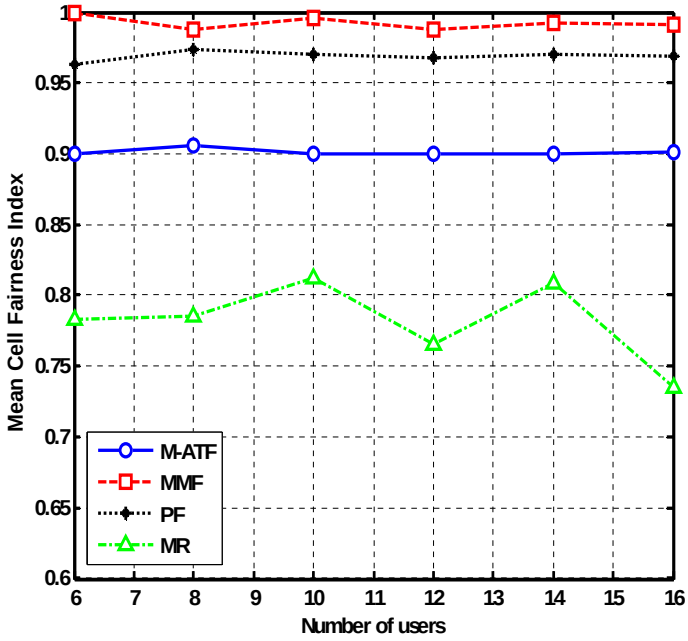


Figure C.1.: Fairness analysis of the modified alpha-rule framework (case-study 1)

The total cell throughput for different cell loads is shown in Fig. C.2. As expected, MR was able to maximize the spectral efficiency, while MMF presented the lowest cell throughput. Since PF is a trade-off between MR and MMF, its performance lied between them. Looking at Fig. C.1, one can notice that the CFT = 0.9 forced the M-ATF technique to behave like an hybrid between the PF and MR strategies. This fact is confirmed in Fig. C.2, where the performance of M-ATF in terms of total cell throughput turned out to be between the performances of PF and MR.

This hybrid behavior between PF and MR presented by the M-ATF technique can also be seen in the user satisfaction, as Fig. C.3 shows. For low and high cell loads, the performance of M-ATF lies between PF and MR. However, for moderate loads, M-ATF provided higher percentage of satisfied users than the classic RRA strategies.

From the simulation results presented in this section, we can conclude that the modified alpha-rule framework is also able to control the trade-off between resource efficiency and throughput-based fairness in a scenario with NRT services in a macrocell OFDMA network.

C.1.3. Case-Study 2: Maximization of User Satisfaction

In this section we propose and evaluate the Adaptive Satisfaction-Based Allocation (ASA) technique. Its objective is to maximize the percentage of satisfied users in the system. The motivation for this proposal was the fact that the classic RRA strategies present different performance results in terms of user satisfaction when the cell load varies. In general, MMF, PF and MR policies

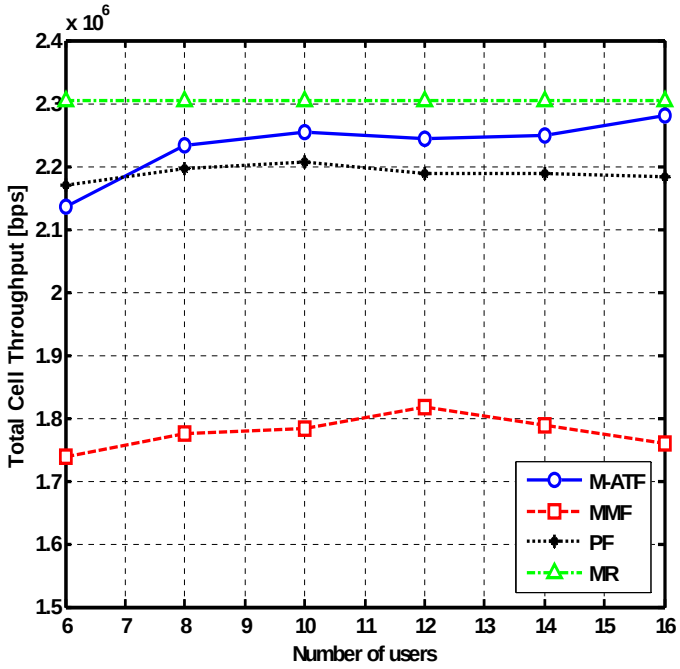


Figure C.2.: Efficiency analysis of the modified alpha-rule framework (case-study 1)

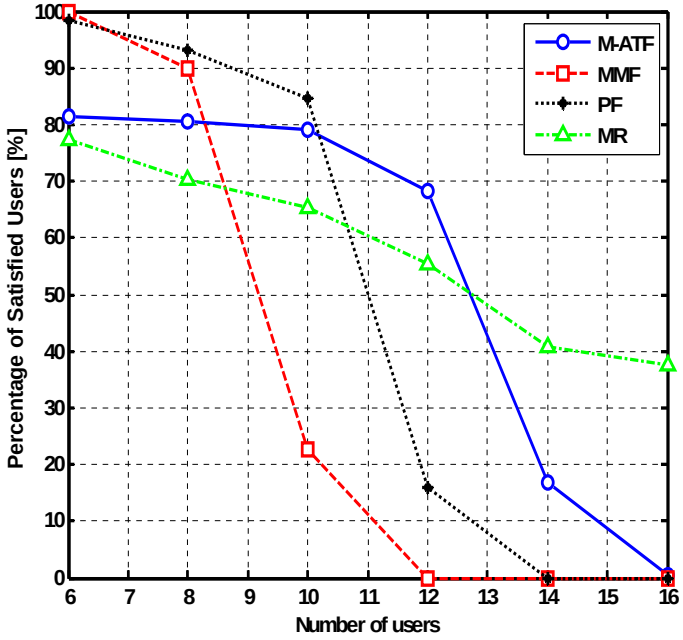


Figure C.3.: Satisfaction analysis of the modified alpha-rule framework (case-study 1)

provide higher user satisfaction at low, moderate and high cell loads, respectively. This fact can be explored and the modified alpha-rule framework can be configured as any particular classic RRA strategy that provides the maximum user satisfaction in that particular time and cell load. In this way, one can make the association between the ASA and Adaptive Modulation and Coding (AMC) techniques. The latter is a physical layer technique widely employed in wireless communication systems, which has the objective of choosing the best Modulation and Coding Scheme (MCS) in order to maximize the system throughput according to the instantaneous channel conditions. In a similar way, by controlling the adaptive parameter α, the ASA technique chooses the most appropriate classic RRA strategy in order to maximize the user satisfaction according to the instantaneous cell load and service quality conditions.

A MT is considered satisfied if its session throughput up to the present moment is higher than its throughput requirement. In each control time window, the ASA technique calculates the percentage of satisfied users $\Psi[n]$ and compares it to two satisfaction transition thresholds. Let us define Ψ_{MMF}^{PF} and Ψ_{PF}^{MR} as the satisfaction transition thresholds between MMF and PF, and between PF and MR, respectively. If $1 \geq \Psi[n] > \Psi_{MMF}^{PF}$, the ASA strategy sets the parameter α to 1 and configures the modified alpha-rule framework to work like the MMF. If $\Psi_{MMF}^{PF} \geq \Psi[n] > \Psi_{PF}^{MR}$, the framework takes the form of PF by setting $\alpha = 0.5$. Otherwise, if $\Psi_{PF}^{MR} \geq \Psi[n] \geq 0$, the ASA technique sets $\alpha = 0$ and the MR criterion takes place. In this way, it is expected that the adaptation of the alpha-rule framework by means of the ASA strategy will take advantage of the best that the classic RRA strategies can offer in terms of user satisfaction for different cell loads. In a real network, the values of the transition parameters Ψ_{MMF}^{PF} and Ψ_{PF}^{MR} must be chosen empirically by the cellular operator based on the observation of the network performance.

In the performance evaluation of the second case-study, we use the same general simulation parameters presented in Table C.2. Furthermore, we replace the particular parameters related to the M-ATF technique with the parameters related with the ASA strategy, as indicated in Table C.3.

Table C.3.: Simulation parameters for the evaluation of the modified alpha-rule framework (case study 2)

Parameter	**Value**
ASA control time window	100 ms
ASA transition threshold (Ψ_{MMF}^{PF})	0.95
ASA transition threshold (Ψ_{PF}^{MR})	0.65

Fig. C.4 presents the main result for the ASA technique: the percentage of satisfied users. Looking at the performance of the classic RRA techniques, one can see that there are ranges of cell load in which each of the classic strategies presents the highest user satisfaction. The ASA technique, by means of the adaptation of the parameter α of the modified alpha-rule framework, is able to select the most appropriate RRA strategy for all ranges of cell loads, i.e. the one which

provides the maximum user satisfaction.

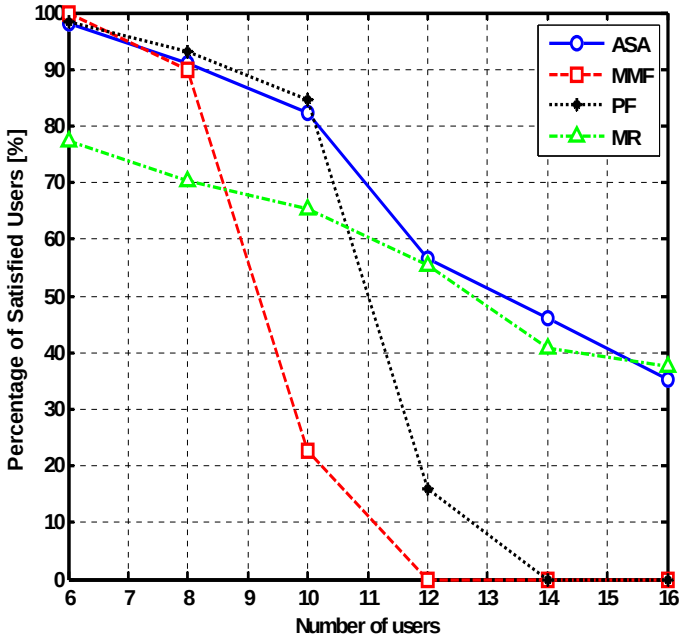


Figure C.4.: Satisfaction analysis of the modified alpha-rule framework (case-study 2)

The comparison of the ASA and the classic RRA techniques regarding the total cell throughput is presented in Fig. C.5. It can be seen that the cell throughput of the ASA technique increases as the cell load also increases. This indicates that for low cell loads, the alpha-rule framework was configured as the MMF strategy, for moderate loads the PF technique was used, and finally for high loads, the framework adopted the form of MR.

This adaptation can also be observed in Fig. C.6, where the system fairness index is presented. Again, one can clearly notice the adaptation pattern of the alpha-rule framework by means of the ASA technique. For low cell loads, since the framework behaves like the MMF strategy, the maximum fairness allocation is achieved. When the load increases, the framework is switched to the PF criterion, which causes a decrease in the fairness index. Finally, in order to achieve a maximum user satisfaction at high loads, the alpha-rule framework is configured as the MR strategy, which makes the system operates with the most unfair resource allocation.

As a conclusion, we can say that the modified alpha-rule framework provides a simple and efficient way to manage the user satisfaction in a macrocell OFDMA network, achieving the highest satisfaction for different system loads.

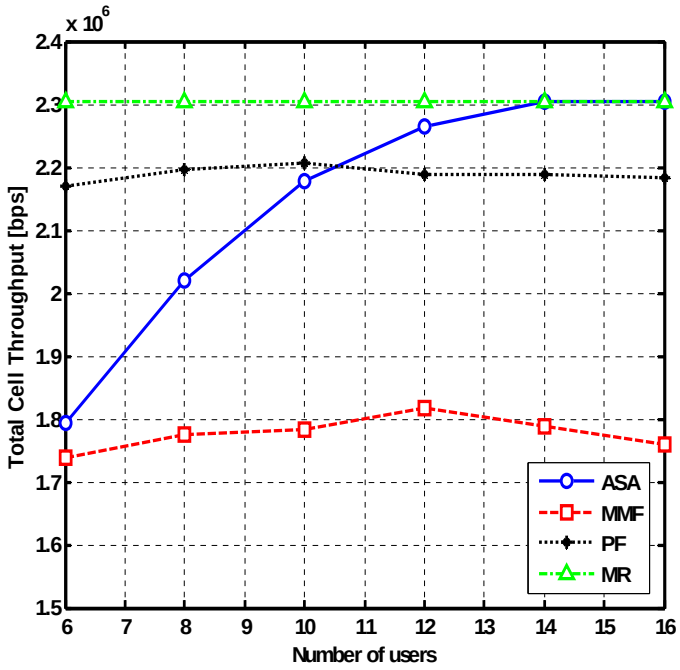


Figure C.5.: Efficiency analysis of the modified alpha-rule framework (case-study 2)

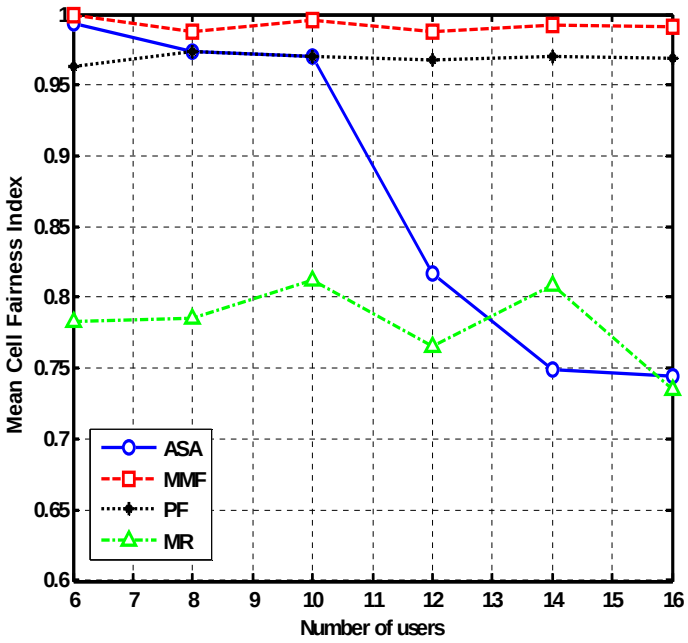


Figure C.6.: Fairness analysis of the modified alpha-rule framework (case-study 2)

C.2. Modified Beta-Rule Framework

The DSA algorithm of the original utility-based beta-rule described in section 4.3.2 of chapter 4 is ruled by the following criterion:

$$m(k,n) = \arg\max_{j} \left\{ \frac{\left(d_j^{\text{hol}}[n]\right)^{\beta}}{T_j[n-1]} \cdot c_{j,k}[n] \right\}. \tag{C.4}$$

The MT $m(k,n)$ is chosen to transmit on the kth sub-carrier at the nth TTI if it satisfies the condition given by (C.4) above. We have that $d_j^{\text{hol}}[n]$ is the current Head-Of-Line (HOL) delay of the jth MT, $T_j[n-1]$ is the throughput (average data rate) of the jth MT up to the previous TTI and $c_{j,k}[n]$ denotes the instantaneous achievable transmission efficiency of the jth MT on the kth sub-carrier (Shannon capacity given by (A.9) on appendix A) assuming equal power allocation per sub-carrier.

Such as in the case of the utility-based alpha-rule, in the original formulation of the utility-based beta-rule, the parameter β is defined on the range $[0,\infty)$. Again it was also shown in the performance evaluation of the Adaptive Delay-Based Fairness (ADF) technique presented in chapter 4 that a maximum value of $\beta = 10$ was enough to provide a controllable trade-off between resource efficiency and delay-based fairness in a scenario with RT services. Now, in this appendix we propose a modification of the beta-rule that gives rise to a Modified Adaptive Delay-Based Fairness (M-ADF) technique. This new technique uses a feedback control loop similar to the one used by the original ADF strategy (see expression (4.30) on page 101), but considering a new range of $\beta \in [0,1]$.

The DSA algorithm of the M-ADF states that the MT $m(k,n)$ is chosen to transmit on the kth sub-carrier at the nth TTI if it satisfies the condition given by (C.5) below:

$$m(k,n) = \arg\max_{j} \left\{ \left(d_j^{\text{hol}}[n]\right)^{\beta} \cdot \left(\frac{c_{j,k}[n]}{T_j[n-1]} \right)^{1-\beta} \right\}, \tag{C.5}$$

Notice that varying the value of the parameter β in (C.4), i.e. $\beta \in [0,\infty)$, or varying β in (C.5), i.e. $\beta \in [0,1]$, provides the same trade-off balancing between system capacity and delay-based fairness.

The M-ADF may have the advantage of a faster convergence of the delay-based CFI. However, it has the drawback that the APA algorithm cannot be parameterized for the new range of β values for similar reasons already explained for the alpha-rule case. Below we show again the waterfilling expression of the APA algorithm of the original utility-based beta-rule:

$$p_k^*[n] = \left[\mu \cdot \frac{\left(d_j^{\text{hol}}[n]\right)^{\beta}}{T_j[n-1]} - \frac{\Gamma}{\gamma_{j,k}[n]} \right]^{+}, \quad \forall k \in \mathcal{S}_j \tag{C.6}$$

Observe that it is not possible to have a free balance between resource efficiency, whose indicator is the ratio between the channel transmission efficiency and the user throughput, and user fairness, whose indicator is the HOL delay. For that reason, the M-ADF technique is composed of a DSA algorithm based on (C.5) followed by an equal power allocation among the sub-carriers. Notice that it has been shown in chapter 4 that no major performance differences exist between EPA and the APA algorithm used in the joint approach.

Just like the original utility-based beta-rule, different performances in terms of resource efficiency and delay-based fairness can be achieved depending on the value of the fairness controlling parameter β in the modified framework. Varying β in the range $[0, 1]$, we have that the new beta-rule framework presented above can be designed to work as different classic RRA policies suitable for RT services, such as PF, Modified Largest Weighted Delay First (M-LWDF) and First-In-First-Out (FIFO). Moreover, if we let the parameter β to be adaptive, we have the M-ADF technique. Table 4.2 summarizes the main characteristics of the modified beta-rule and the four particular RRA policies contemplated by this framework.

Table C.4.: Features of the modified beta-rule framework

Policies	**Parameter β**	**DSA priority function**	**Characteristics**
PF	0	$\frac{c_{j,k}[n]}{T_j[n-1]}$	High resource efficiency and low delay-based fairness
M-LWDF	0.5	$\sqrt{d_j^{\text{hol}}[n] \cdot \frac{c_{j,k}[n]}{T_j[n-1]}}$	Static trade-off between resource efficiency and delay-based fairness
FIFO	1	$d_j^{\text{hol}}[n]$	Low resource efficiency and high delay-based fairness
M-ADF	adaptive	$\left(d_j^{\text{hol}}[n]\right)^{\beta} \cdot \left(\frac{c_{j,k}[n]}{T_j[n-1]}\right)^{1-\beta}$	Dynamic trade-off between resource efficiency and delay-based fairness

Appendix D

Branch and Bound Technique

D.1. Overview of the BnB Technique

Branch and Bound (BnB) is a general algorithmic tool for finding optimum solutions of various optimization problems, especially in discrete and combinatorial optimization [26–29]. It can be used to solve minimization or maximization problems. Without loss of generality, let us consider that the problem is to maximize an objective function $f(x)$ of variables $(x_1, \ldots, x_n)$ over a region of feasible solutions $\mathcal{D}$, as indicated below:

$$\max_{x \in \mathcal{D}} f(x) \tag{D.1}$$

Let us define a set of potential solutions $\mathcal{P}$, which contains $\mathcal{D}$, and where $f(x)$ is still well defined. Fig. D.1 illustrates an example where $\mathcal{D}$ and $\mathcal{P}$ are intervals of real numbers. A bounding function $g(x)$ defined on $\mathcal{D}$ (or $\mathcal{P}$) with the property that $g(x) \geq f(x)$ for all $x \in \mathcal{D}$ (or $\mathcal{P}$) is also considered.

The BnB algorithm searches the complete space of solutions of a given problem in order to find the optimum solution. However, some parts of the solution space are searched only implicitly, which makes this task manageable. The nodes of a dynamically generated search tree represent the subspaces that were not explored yet. A classical BnB is composed of three main steps: selection of the node to process, branching, and bound calculation. These procedures can be briefly explained using Fig. D.2, which illustrates the search into the solution space in sequential phases.

Initially there is only one subset, which is the complete solution space (Fig. D.2(a)). The sequence of phases of the search procedure may vary according to how the next node to be processed is selected. After selecting the node, a branching is performed, i.e. the solution space of the node is subdivided into two or more subspaces (see Fig. D.2(b)). Given a node Q of the tree, the children of Q are subproblems derived from Q through imposing new constraints for each subproblem.

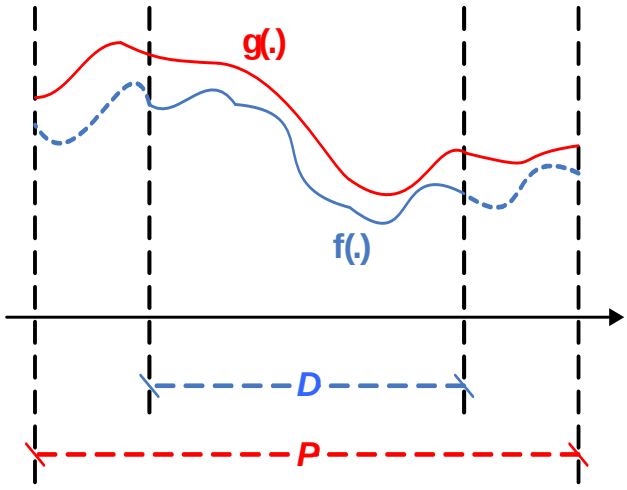
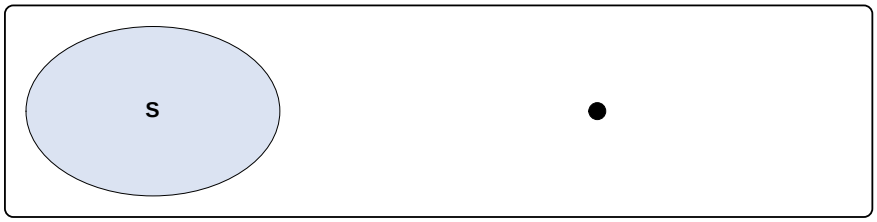
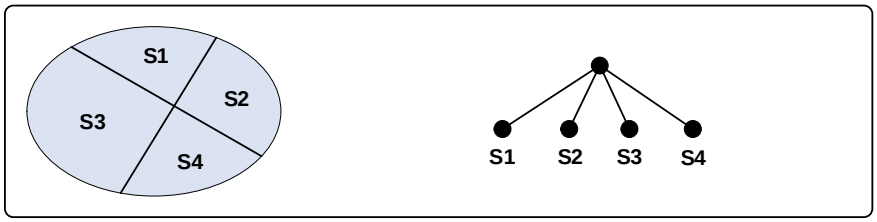


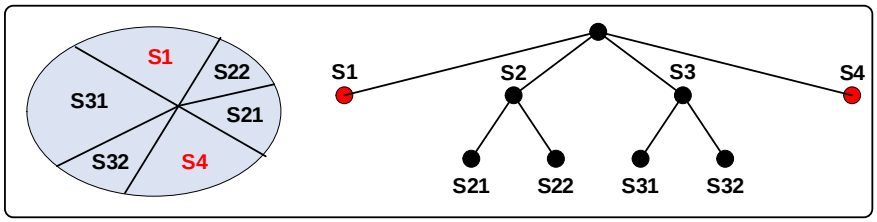
Figure D.1.: The relation between the objective function $f(\cdot)$ and the bounding function $g(\cdot)$ on the sets $\mathcal{D}$ and $\mathcal{P}$ of feasible and potential solutions of a maximization problem, respectively.



(a) Phase 1



(b) Phase 2



(c) Phase 3

Figure D.2.: Example of illustration of the search space of the Branch and Bound technique in three phases

For each node of the tree, it is checked whether the correspondent subspace consists of a single solution. If it is true, it means that this node is a leaf (terminal node) and it represents a feasible solution. In this case, it is compared to the current best solution, which is called incumbent, keeping the best of these. If it is an internal node, the bounding function $g(x)$ (see Fig. D.1) for the corresponding subspace is calculated and compared to the incumbent. If the upper bound is worse than the incumbent, it means that the subspace cannot contain the optimum solution and so the subproblem is discarded (or fathomed) (nodes in red color in Fig. D.2(c)). Otherwise, the node is joined to the pool of live subproblems to be processed in further iterations of the algorithm. The search terminates when there is no unexplored parts of the solution space, and then the optimum solution is the one recorded as the incumbent.

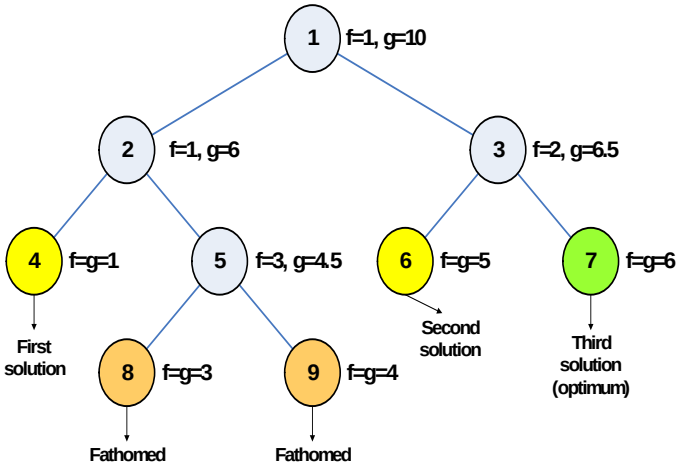
There are different strategies for selecting the next live subproblem to investigate. Fig. D.3 presents three node selection strategies: *Breadth First*, *Depth First* and *Best First*. The numbers in each node corresponds to the sequence in which the nodes are processed.

Firstly, let us observe in these examples the properties of the bounding function and its relation with the objective function throughout the search tree. In the case of a maximization problem, we have that the bounding function $g(x)$ is required to satisfy the following three conditions:

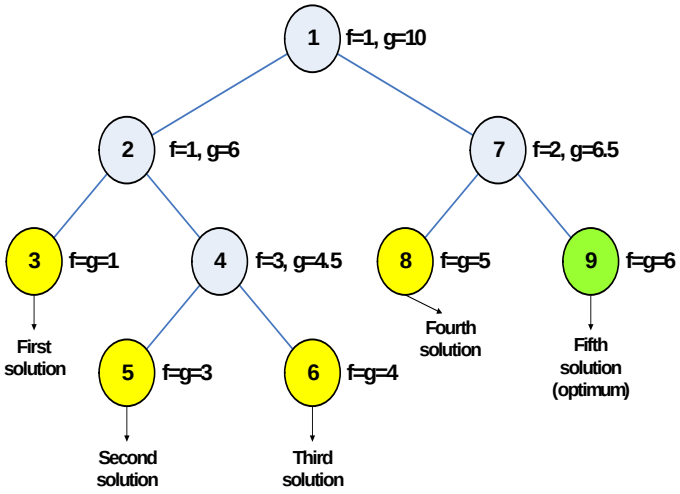
- $g(Q_i) \geq f(Q_i)$ for all nodes Q_i in the tree;
- $g(Q_i) = f(Q_i)$ for all leaves in the tree;
- $g(Q_i) \leq g(Q_j)$ if Q_j is the father of Q_i.

Notice that all these conditions are satisfied in Fig. D.3. In this figure, we also present the solutions found and the nodes that are discarded. Let us call the set of live nodes as *Live*. In the following, we explain the sequence of processing of the nodes for each of the search strategies as well as the branching and bounding procedures:

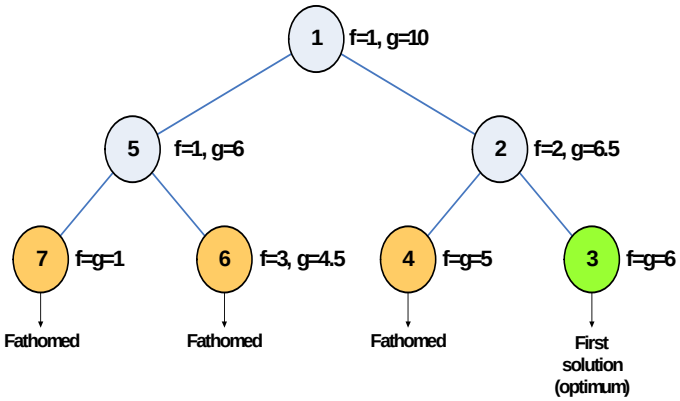
- *Breadth First* (Fig. D.3(a)): all nodes at one level of the search tree are processed before any node at a higher level.
  1. Initially we have $Live = \{1\}$. Select node 1, branch it into nodes 2 and 3 and calculate their bounds. Now we have $Live = \{2, 3\}$. All nodes in the first level were processed, so we can pass to the next level.
  2. Select node 2, branch it into nodes 4 and 5. At this stage, $Live = \{3, 4, 5\}$.
  3. Node 3 is at the same level of node 2, so it is selected and branched into nodes 6 and 7. The bound values of nodes 6 and 7 are also calculated. After that we have $Live = \{4, 5, 6, 7\}$. All nodes in the second level were processed, so we can pass to the next level.
  4. Select node 4. It cannot be branched because it is a terminal node. It is saved as the first solution and its bound value $g = 1$ is recorded as the current incumbent. $Live = \{5, 6, 7\}$.



(a) Breadth first



(b) Depth first



(c) Best first

Figure D.3.: Search strategies of the Branch and Bound technique in a maximization problem

5. Select node 5, branch it into nodes 8 and 9, and calculate their bounds. $Live = \{6, 7, 8, 9\}$.
6. Select node 6. It cannot be branched because it is a terminal node. Since it has a bound value $g = 5$ higher than the current incumbent, it is saved as the current best solution. The current incumbent is 5 now. $Live = \{7, 8, 9\}$.
7. Select node 7. It is recorded as the new best solution because its bound $g = 6$ is the highest. The new incumbent is equal to 6. All nodes in the third level were processed ($Live = \{8, 9\}$), so we can pass to the next level.
8. Select node 8. It cannot be branched because it is a terminal node. The same for node 9. None of them are better solutions than node 7. End of search because all nodes in the tree were processed ($Live = \varnothing$).

- *Depth First* (Fig. D.3(b)): the live node with the largest level in the search tree is chosen for exploration.
  1. Initially, there is only the root node, so $Live = \{1\}$. Select node 1, branch it into nodes 2 and 7 and calculate their bounds. After that, $Live = \{2, 7\}$.
  2. Among the live nodes, select the one with the largest level. In this case, node 2 or 7 could be selected. For instance, in the example we select node 2, branch it into nodes 3 and 4, and calculate the bounds. $Live = \{7, 3, 4\}$.
  3. Node 3 or 4 could be selected because they have the largest level. For example, select node 3. It cannot be branched because it is a terminal node. Save it as the first solution and its bound value, which is $g = 1$, is saved as the current incumbent. $Live = \{7, 4\}$.
  4. Select node 4, branch it into nodes 5 and 6 and calculate the bounds. $Live = \{7, 5, 6\}$.
  5. Among the live nodes, nodes 5 and 6 have the largest levels. For instance, select node 5. It cannot be branched because it is a terminal node. Since it has a bound $g = 3$ higher than the current incumbent, it is recorded as the current best solution and the current incumbent is 3 now. The same occurs for node 6, which will be the next solution and its bound value, which is $g = 4$, will be the next incumbent. After that, we have $Live = \{7\}$.
  6. Node 7 is the only node that is alive at this moment. Select and branch it into nodes 8 and 9, and calculate their respective bounds. $Live = \{8, 9\}$.
  7. Node 8 or 9 could be selected because they are in the same level. Select node 8. It cannot be branched because it is a terminal node and its bound $g = 5$ is higher than the current incumbent. So it is saved as the new solution and the current incumbent will be 5. But the same occurs with node 9, which is also a terminal node and has a bound value $g = 6$ higher than the current incumbent, so it is saved as the best solution. End of search because all nodes in the tree were processed ($Live = \varnothing$).

- *Best First* (Fig. D.3(c)): selects among the live subproblems the one with the highest bound.

    1. Initially we have $Live = \{1\}$. Select node 1, branch it into nodes 5 and 2 and calculate their bounds. After that we have $Live = \{5, 2\}$.

    2. Among the live nodes, select the one with the largest bound (node 2 with $g = 6.5$). Branch it into nodes 4 and 3 and calculate their bounds. $Live = \{5, 4, 3\}$.

    3. Node 3 is selected because it has the largest bound among the live nodes ($g = 6$). It cannot be branched because it is a terminal node. It is saved as the first solution and first incumbent will be 6. Since node 4 is a brother of node 3, its bound is immediately compared with the current incumbent. Since it has a bound $g = 5$ that is lower than the incumbent, it cannot contain the optimum solution and is fathomed. After that, we have $Live = \{5\}$.

    4. Node 5 is the only node alive. So select and branch it into nodes 7 and 6 and calculate their bounds. None of them are better than the current best solution, so both are fathomed. End of search because all nodes in the tree were processed ($Live = \varnothing$).

From the procedures explained above, we can observe that in the *Breadth First* search strategy, the number of nodes at each level of the search tree can grow exponentially with the level. Due to this reason, the use of this search strategy for larger problems can be unfeasible.

This problem is not present in the *Depth First* search strategy. In this case, the memory requirement in terms of number of subproblems to store at the same time is upper bounded by the number of levels in the search tree multiplied by the maximum number of children of any node, which is usually a quite manageable number. The drawback is that if the incumbent is far from the optimum solution, large amounts of unnecessary bounding computations may take place.

The best alternative is the *Best First* search. Since the optimum solution is found earlier, no superfluous bound calculations take place after that and the memory required to store the live nodes is also decreased.

D.2. Application of the BnB Technique to Find the Groups of Mutual Interfering FAPs in a Femtocell Tier

As commented in section 5.6.1 of chapter 5, an important pre-processing task of the Dynamic Frequency Planning (DFP) algorithms is the determination of the groups of mutual interfering Femtocell Access Points (FAPs), which are called G_m groups, associated to each cluster of Femtocell Access Points (FAPs).

Let us assume the cluster topology matrix $\mathbf{A}$ given by (D.2), which corresponds to the cluster of

7 FAPs exemplified in Fig. D.4 (see section 5.6.1 of chapter 5 for more details).

$$\mathbf{A} = \begin{pmatrix} 0 & 1 & 1 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \tag{D.2}$$

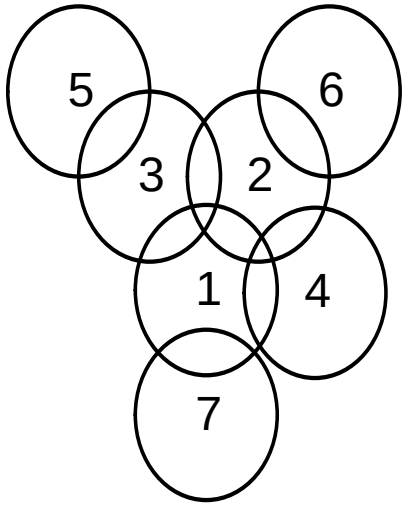


Figure D.4.: Example of a cluster of FAPs - $N = 7$

One can notice that there are a total of M=5 mutual interfering groups G_m, which are given by $G_1 = \{1, 2, 3\}$, $G_2 = \{1, 2, 4\}$, $G_3 = \{1, 7\}$, $G_4 = \{2, 6\}$ and $G_5 = \{3, 5\}$. These groups could be determined by a careful observation of the topology matrix and the use of a proper algorithmic logic. However, the algorithmic structure of the BnB technique could be used to easily solve this problem. For this specific problem, we do not need to find the optimum solution of an optimization problem. We only need to map the problem into the structure of a search tree. In order to do that, only the search strategy and the branching procedure of the BnB technique are necessary. Fig. D.5 illustrates the mapping of our problem into the structure of a search tree. This tree represents the cluster topology with regard to FAP 1. Therefore, with this tree we are able to find only the G_m groups that FAP 1 takes part. In order to find all G_m groups of a cluster, a different search tree must be built for every FAP in that cluster. Since bound calculations are not necessary, the *Breadth First* or the *Depth First* search strategies can be used (see section D.1 for more details).

In order to find the G_m groups of FAP 1, the following steps are executed (see Fig. D.5):

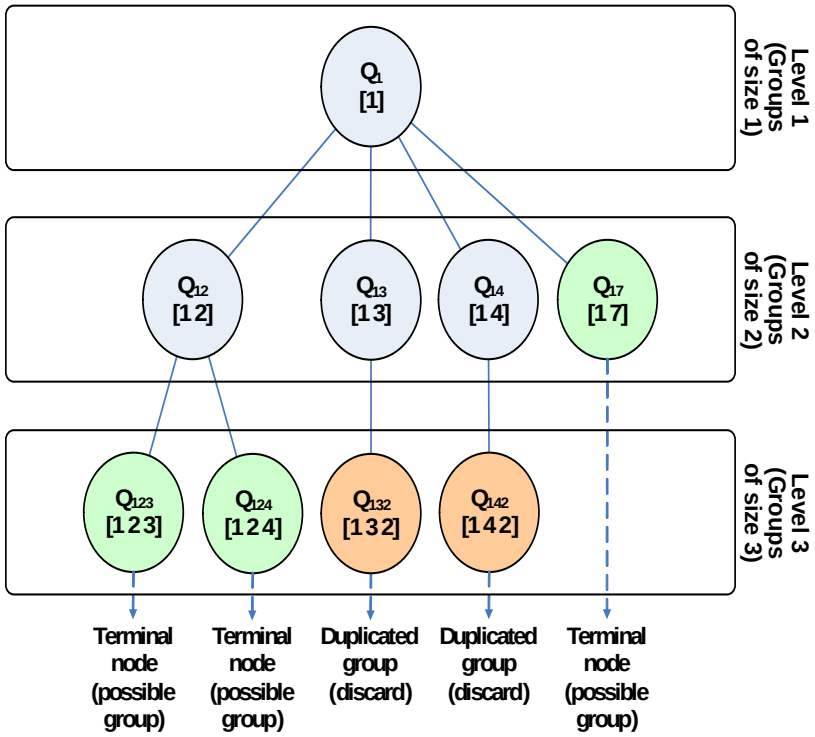


Figure D.5.: Example of a dynamic search tree for finding the G_m groups that FAP 1 takes part

1. Select the father node Q_1, which represents FAP 1, and branch it into the children nodes Q_{12}, Q_{13}, Q_{14} and Q_{17}. These children nodes are determined by imposing interference topology constraints. Each children node corresponds to a different neighboring FAP that interferes with FAP 1 (see Fig. D.4).

2. Select the father node Q_{12}, which represents the group formed by FAPs 1 and 2, and branch it into the children nodes Q_{123} and Q_{124}. Each children node corresponds to a different neighboring FAP that interferes simultaneously with FAPs 1 and 2. Since Q_{123} and Q_{124} are terminal nodes (leaves of the tree), they are feasible solutions and are possible G_m groups.

3. Select the father node Q_{13}, which represents the group formed by FAPs 1 and 3, and branch it into the child node Q_{132}. This child node says that FAP 2 interferes simultaneously of FAPs 1 and 3. Notice that this group was already determined, i.e. it is a duplicated version of the group [1 2 3]. So, this duplicated group is discarded. The same occurs for the child node Q_{142}.

4. Select the father node Q_{17}, which represents the group formed by FAPs 1 and 7. Since there are not other FAPs that interfere simultaneously with FAPs 1 and 7, it means that this node cannot be branched, and so it is a terminal node (leaf). As a terminal node, it represents a feasible solution and a possible G_m group.

5. The search is terminated because there are not unexplored parts of the tree remaining. As a result, the procedure indicates that FAP 1 takes part on the following G_m groups: $G_1 = \{1, 2, 3\}$, $G_2 = \{1, 2, 4\}$ and $G_3 = \{1, 7\}$.

Bibliography

[1] ITU, "Detailed specifications of the terrestrial radio interfaces of International Mobile Telecommunications-2000 (IMT-2000)," Tech. Rep. ITU-R M.1457-9, 2010.

[2] H. Holma and A. Toskala, eds., *WCDMA for UMTS: Radio access for third generation mobile communications.* John Wiley & Sons, 3rd ed., 2004.

[3] 3GPP, "Physical layer - general description," Tech. Rep. TS 25.201 v10.0.0, Release 10, 3rd. Generation Partnership Project (3GPP), 2011.

[4] 3GPP2, "Physical layer standard for cdma2000 spread spectrum systems - revision D," Tech. Rep. 3GPP2 C.S0002-D v1.0, 3rd. Generation Partnership Project 2 (3GPP2), 2004.

[5] IEEE, "IEEE standard for local and metropolitan area networks - part 16: Air interface for fixed and mobile broadband wireless access systems," Tech. Rep. IEEE Std 802.16e-2005, Institute of Electrical and Electronics Engineers (IEEE), 2006.

[6] IEEE, "IEEE standard for local and metropolitan area networks - part 16: Air interface for fixed broadband wireless access systems," Tech. Rep. IEEE Std 802.16-2001, Institute of Electrical and Electronics Engineers (IEEE), 2002.

[7] ITU, "Framework and overall objectives of the future development of IMT-2000 and systems beyond IMT-2000," Tech. Rep. ITU-R M.1645, 2003.

[8] 3GPP, "Requirements for further advancements for Evolved Universal Terrestrial Radio Access (E-UTRA) (LTE-Advanced)," Tech. Rep. TR 36.913 v10.0.0, Release 10, 2011.

[9] IEEE, "IEEE standard for local and metropolitan area networks - part 16: Air interface for broadband wireless access systems - advanced air interface," Tech. Rep. IEEE Std 802.16m, Institute of Electrical and Electronics Engineers (IEEE), 2011.

[10] R. van Nee and R. Prasad, *OFDM for Wireless Multimedia Communications.* Artech House, 2000.

[11] V. Chandrasekhar and J. G. Andrews, "Femtocell networks: A survey," *IEEE Communications Magazine*, vol. 46, pp. 59–67, Sep. 2008.

[12] J. Zhang and G. de la Roche, *Femtocells: Technologies and Deployment.* John Wiley & Sons, 2010.

[13] 3GPP, "Evolved universal terrestrial radio access (E-UTRA) and evolved universal terrestrial radio access (E-UTRAN); overall description; stage 2," Tech. Rep. TS 36.300 v10.3.0, Release 10, 3rd. Generation Partnership Project (3GPP), 2011.

[14] 3GPP, "Evolved universal terrestrial radio access (E-UTRA); FDD home enode B (HeNB) radio frequency (RF) requirements analysis," Tech. Rep. TR 25.921 v9.0.0, Release 9, 3rd. Generation Partnership Project (3GPP), 2010.

[15] M. C. Jeruchim, P. Balaban, and K. S. Shanmugan, *Simulation of Communication Systems - Modeling, Methodology, and Techniques*. Information Technology: Transmission, Processing, and Storage, Kluwer Academic/Plenum Publishers, 2nd ed., 2000.

[16] A. M. Law and W. D. Kelton, *Simulation Modeling and Analysis*. Industrial Engineering and Management Science Series, McGraw-Hill, 3rd ed., 2000.

[17] J. Gross and M. Bohge, "Dynamic mechanisms in OFDM wireless systems: A survey on mathematical and system engineering contributions," Tech. Rep. Tech. rep. TKN-06-001; Telecommunication Networks Group; Technical University of Berlin, 2006.

[18] D. Astely, E. Dahlman, F. Pal, R. Ludwig, M. Meyer, S. Parkvall, P. Skillermark, and N. Wiberg, "A future radio-access framework," *IEEE Journal on Selected Areas in Communications*, vol. 24, no. 3, pp. 693–706, 2006.

[19] M. Bohge, J. Gross, A. Wolisz, and M. Meyer, "Dynamic resource allocation in OFDM systems: an overview of cross-layer optimization principles and techniques," *IEEE Network*, vol. 21, no. 1, pp. 53–59, 2007.

[20] R. Fantacci, D. Marabissi, D. Tarchi, and I. Habib, "Adaptive modulation and coding techniques for OFDMA systems," *IEEE Transactions on Wireless Communications*, vol. 8, pp. 4876–4883, Sep. 2009.

[21] S. H. Ali and V. C. M. Leung, "Dynamic frequency allocation in fractional frequency reused OFDMA networks," *IEEE Transactions on Wireless Communications*, vol. 8, pp. 4286–4295, Aug. 2009.

[22] D. Kivanc and L. Hui, "Subcarrier allocation and power control for OFDMA," in *Proc. 34th Asilomar Conference on Signals, Systems and Computers*, vol. 1, pp. 147–151, 2000.

[23] G. Li and H. Liu, "Downlink dynamic resource allocation for multi-cell OFDMA system," in *Proc. IEEE 58th Vehicular Technology Conference - VTC Fall*, vol. 3, pp. 1698–1702, 2003.

[24] P. C. Fishburn, "Utility theory," *Management Science - Theory Series*, vol. 14, pp. 335–378, Jan. 1968.

[25] D. W. North, "A tutorial introduction to decision theory," *IEEE Transactions on Systems Science and Cybernetics*, vol. ssc-4, pp. 200–210, Sep. 1968.

[26] A. H. Land and A. G. Doig, "An automatic method of solving discrete programming problems," *Econometrica*, vol. 28, no. 3, pp. 497–520, 1960.

[27] L. G. Mitten, "Branch-and-bound methods: General formulation and properties," *Operations Research*, vol. 18, no. 1, pp. 24–34, 1970.

[28] J. Clausen, "Branch and bound algorithms - principles and examples," tech. rep., Departament of Computer Science, University of Copenhagen, 1999.

[29] S. Boyd and J. Mattingley, "Branch and bound methods," tech. rep., Stanford University, 2007.

[30] Z. Mao, X. M. Wang, and J. Lin, "Fast optimal radio resource allocation in OFDMA system based on branch-and-bound method," in *Proc. IEEE Pacific Rim Conference on Communications, Computers and signal Processing - PACRIM*, pp. 348–351, 2005.

[31] I. Kalet, "The multitone channel," *IEEE Transactions on Communications*, vol. 37, no. 2, pp. 119–124, 1989.

[32] H. S. Kim, J. S. Kwak, J. M. Choi, and J. H. Lee, "Efficient subcarrier and bit allocation algorithm for OFDMA system with adaptive modulation," in *Proc. IEEE 59th Vehicular Technology Conference - VTC Spring*, vol. 3, pp. 1816–1820, 2004.

[33] C. H. Yih and E. Geranotis, "Centralized power control algorithms for OFDM cellular networks," in *Proc. IEEE Military Communications Conference*, pp. 1250–1255, 2003.

[34] Y. Wang, F. Chen, and G. Wei, "Adaptive subcarrier and bit allocation for multiuser OFDM system based on genetic algorithm," in *Proc. International Conference on Communications, Circuits and Systems*, vol. 1, pp. 242–246, 2005.

[35] L. T. H. Lee, C. Chung-Ju, C. Yih-Shen, and S. Shen, "A utility-approached radio resource allocation algorithm for downlink in OFDMA cellular systems," in *Proc. IEEE 61st Vehicular Technology Conference - VTC Spring*, vol. 3, pp. 1798–1802, 2005.

[36] H. W. Kuhn, "The Hungarian method for the assignment problem," *Naval Research Logistics Quaterly*, no. 2, pp. 83–97, 1955.

[37] C. Y. Wong, C. Y. Tsui, R. S. Cheng, and K. B. Letaief, "A real-time sub-carrier allocation scheme for multiple access downlink OFDM transmission," in *Proc. IEEE 50th Vehicular Technology Conference - VTC Fall*, vol. 2, pp. 1124–1128, 1999.

[38] M. Ergen, S. Coleri, and P. Varaiya, "QoS aware adaptive resource allocation techniques for fair scheduling in OFDMA based broadband wireless access systems," *IEEE Transactions on Broadcasting*, vol. 49, no. 4, pp. 362–370, 2003.

[39] S. Pietrzyk and G. J. M. Janssen, "Multiuser subcarrier allocation for QoS provision in the OFDMA systems," in *Proc. IEEE 56th Vehicular Technology Conference - VTC Fall*, vol. 2, pp. 1077–1081, 2002.

[40] Z. Han, Z. Ji, and K. J. R. Liu, "Fair multiuser channel allocation for OFDMA networks using Nash bargaining solutions and coalitions," *IEEE Transactions on Communications*, vol. 53, no. 8, pp. 1366–1376, 2005.

[41] D. Niyato and E. Hossain, "Adaptive fair subcarrier/rate allocation in multirate OFDMA networks: Radio link level queuing performance analysis," *IEEE Transactions on Vehicular Technology*, vol. 55, no. 6, pp. 1897–1907, 2006.

[42] I. Kim, H. L. Lee, B. Kim, and Y. H. Lee, "On the use of linear programming for dynamic subchannel and bit allocation in multiuser OFDM," in *Proc. IEEE Global Telecommunications Conference - GLOBECOM*, vol. 6, pp. 3648–3652, 2001.

[43] Z. Han, F. R. Farrokhi, Z. Ji, and K. J. R. Liu, "Capacity optimization using subspace method over multicell OFDMA networks," in *Proc. IEEE Wireless Communications and Networking Conference - WCNC*, vol. 4, pp. 2393–2398, 2004.

[44] Z. Ji and K. J. R. Liu, "Dynamic spectrum sharing: A game theoretical overview," *IEEE Communications Magazine*, vol. 45, pp. 88–94, May 2007.

[45] Z. Han, Z. Ji, and K. J. R. Liu, "Non-cooperative resource competition game by virtual referee in multi-cell OFDMA networks," *IEEE Journal on Selected Areas in Communications*, vol. 25, pp. 1079–1090, Aug. 2007.

[46] 3GPP, "UTRAN architecture for 3G home node b (HNB); stage 2," Tech. Rep. TS 25.467 v10.1.0, Release 10, 3rd. Generation Partnership Project (3GPP), 2011.

[47] W. Forum, "Network architecture: Architecture, detailed protocols and procedures - femtocells core specification," Tech. Rep. WMF-T33-118-R016v01, WiMAX Forum, 2010.

[48] S. Parkvall, A. Furuskär, and E. Dahlman, "Evolution of LTE toward IMT-advanced," *IEEE Communications Magazine*, vol. 49, pp. 84–91, Feb. 2011.

[49] S. Pietrzyk and G. J. M. Janssen, "Subcarrier allocation and power control for QoS provision in the presence of CCI for the downlink of cellular OFDMA systems," in *Proc. IEEE 57th Vehicular Technology Conference - VTC Spring*, vol. 4, pp. 2221–2225, 2003.

[50] C. Lengoumbi, P. Godlewski, and P. Martins, "Dynamic subcarrier reuse with rate guaranty in a downlink multicell OFDMA system," in *Proc. IEEE 17th International Symposium on Personal, Indoor and Mobile Radio Communications - PIMRC*, pp. 1–5, 2006.

[51] A. Abrardo, A. Alessio, P. Detti, and M. Moretti, "Centralized radio resource allocation for OFDMA cellular systems," in *Proc. IEEE International Conference on Communications - ICC*, pp. 5738–5743, Jun. 2007.

[52] J. Li, H. Kim, Y. Lee, and Y. Kim, "A novel broadband wireless OFDMA scheme for downlink in cellular communications," in *Proc. IEEE Wireless Communications and Networking Conference - WCNC*, vol. 3, pp. 1907–1911, 2003.

[53] N. Damji and T. Le-Ngoc, "Adaptive downlink multi-carrier resource allocation for real-time multimedia traffic in cellular systems," in *Communications, 2004 IEEE International Conference on*, vol. 7, pp. 4258–4262, 2004.

[54] M. Moretti and A. Todini, "A resource allocator for the uplink of multi-cell OFDMA systems," *IEEE Transactions on Wireless Communications*, vol. 6, pp. 2807–2812, Aug. 2007.

[55] S. Pietrzyk and G. J. M. Janssen, "Radio resource allocation for cellular networks based on OFDMA with QoS guarantees," in *Proc. IEEE Global Telecommunications Conference - GLOBECOM*, vol. 4, pp. 2694–2699, 2004.

[56] M. Sawahashi, Y. Kishiyama, A. Morimoto, D. Nishikawa, and M. Tanno, "Coordinated multipoint transmission/reception techniques for LTE-Advanced," *IEEE Wireless Communications*, vol. 17, pp. 26–34, Jun. 2010.

[57] C. Y. Wong, R. S. Cheng, K. B. Lataief, and R. D. Murch, "Multiuser OFDM with adaptive subcarrier, bit, and power allocation," *IEEE Journal on Selected Areas in Communications*, vol. 17, no. 10, pp. 1747–1758, 1999.

[58] Y. J. Zhang and K. B. Letaief, "Adaptive resource allocation and scheduling for multiuser packet-based OFDM networks," in *Proc. IEEE International Conference on Communications - ICC*, vol. 5, pp. 2949–2953, 2004.

[59] S. Pfletschinger, G. Mnz, and J. Speidel, "Efficient subcarrier allocation for multiple access in OFDM systems," in *7th International OFDM-Workshop*, 2002.

[60] D. Kivanc, L. Guoqing, and L. Hui, "Computationally efficient bandwidth allocation and power control for OFDMA," *IEEE Transactions on Wireless Communications*, vol. 2, no. 6, pp. 1150–1158, 2003.

[61] J. Jang and K. B. Lee, "Transmit power adaptation for multiuser OFDM systems," *IEEE Journal on Selected Areas in Communications*, vol. 21, no. 2, pp. 171–178, 2003.

[62] H. Yin and H. Liu, "An efficient multiuser loading algorithm for OFDM-based broadband wireless systems," in *Proc. IEEE Global Telecommunications Conference - GLOBECOM*, vol. 1, pp. 103–107, 2000.

[63] Z. Shen, J. G. Andrews, and B. L. Evans, "Optimal power allocation in multiuser OFDM systems," in *Proc. IEEE Global Telecommunications Conference - GLOBECOM*, vol. 1, pp. 337–341, 2003.

[64] W. Rhee and J. M. Cioffi, "Increase in capacity of multiuser OFDM system using dynamic subchannel allocation," in *Proc. IEEE 51st Vehicular Technology Conference - VTC Spring*, vol. 2, pp. 1085–1089, 2000.

[65] C. Mohanram and S. Bhashyam, "A sub-optimal joint subcarrier and power allocation algorithm for multiuser OFDM," *IEEE Communications Letters*, vol. 9, no. 8, pp. 685–687, 2005.

[66] G. Song, *Cross-Layer Resource Allocation and Scheduling in Wireless Multicarrier Networks*. PhD thesis, Georgia Institute of Technology, Georgia, USA, 2005.

[67] Z. Jiang, Y. Ge, and Y. G. Li, "Max-utility wireless resource management for best-effort traffic," *IEEE Transactions on Wireless Communications*, vol. 4, pp. 100–111, Jan. 2005.

[68] B. Soret, M. C. Aguayo-Torres, J. T. Entrambasaguas, and J. F. Paris, "Utility based adaptive resource allocation for heterogeneous QoS requirements," in *Proc. 3rd International Symposium on Wireless Communication Systems - ISWCS*, pp. 35–39, Sep. 2006.

[69] M. Katoozian, K. Navaie, and H. Yanikomeroglu, "Utility-based adaptive radio resource allocation in OFDM wireless networks with traffic prioritization," *IEEE Transactions on Wireless Communications*, vol. 8, pp. 66–71, Jan. 2009.

[70] G. de la Roche, A. Valcarce, D. López-Pérez, and J. Zhang, "Access control mechanisms for femtocells," *IEEE Communications Magazine*, vol. 48, pp. 33–39, Jan. 2010.

[71] D. López-Pérez, A. Valcarce, G. de la Roche, and J. Zhang, "OFDMA femtocells: A roadmap on interference avoidance," *IEEE Communications Magazine*, vol. 47, pp. 41–48, Sep. 2009.

[72] H. Claussen, "Performance of macro- and co-channel femtocells in a hierarchical cell structure," in *Proc. IEEE 18th International Symposium on Personal, Indoor and Mobile Radio Communications - PIMRC*, pp. 1–5, Sep. 2007.

[73] V. Chandrasekhar and J. G. Andrews, "Power control in two-tier femtocell networks," *IEEE Transactions on Wireless Communications*, vol. 8, pp. 4316–4328, Aug. 2009.

[74] Z. Bharucha, A. Saul, G. Auer, and H. Haas, "Dynamic resource partitioning for downlink femto-to-macro-cell interference avoidance," *EURASIP Journal on Wireless Communications and Networking*, vol. 2010, 2010.

[75] L. Giupponi and C. Ibars, "Distributed interference control in OFDMA-based femtocells," in *Proc. IEEE 21st International Symposium on Personal Indoor and Mobile Radio Communications - PIMRC*, pp. 1201–1206, Sep. 2010.

[76] M. Bennis and M. Debbah, "On spectrum sharing with underlaid femtocell networks," in *Proc. IEEE 21st International Symposium on Personal, Indoor and Mobile Radio Communications Workshops - PIMRC*, pp. 185–190, Sep. 2010.

[77] 3GPP, "Home node B radio frequency (RF) requirements (FDD)," Tech. Rep. TR 25.967, Release 9, 2009.

[78] V. Chandrasekhar and J. G. Andrews, "Spectrum allocation in tiered cellular networks," *IEEE Transactions on Wireless Communications*, vol. 57, pp. 3059–3068, Oct. 2009.

[79] D. López-Pérez, A. Ladanyi, and J. Zhang, "OFDMA femtocells: A self-organizing approach for frequency assignment," in *Proc. IEEE 20th Int. Symp. on Personal, Indoor and Mobile Radio Communications - PIMRC*, pp. 2202–2207, Sep. 2009.

[80] D. López-Pérez, G. de la Roche, A. Valcarce, A. J uttner, and J. Zhang, "Interference avoidance and dynamic frequency planning for WiMAX femtocells networks," in *Proc. 11th IEEE Singapore International Conference on Communication Systems - ICCS*, pp. 1579–1584, Nov. 2008.

[81] L. G. U. Garcia, K. I. Pedersen, and P. E. Mogensen, "Autonomous component carrier selection: interference management in local area environments for LTE-Advanced," *IEEE Communications Magazine*, vol. 47, pp. 110–116, Sep. 2009.

[82] X. Liu, E. Chong, and N. Shroff, "Opportunistic transmission scheduling with resource-sharing constraints in wireless networks," *IEEE Journal on Selected Areas in Communications*, vol. 19, pp. 2053–2064, Oct. 2001.

[83] D. Pong and T. Moors, "Fairness and capacity trade-off in IEEE 802.11 WLANs," in *Proc. 29th Annual IEEE International Conference on Local Computer Networks*, pp. 310–317, Nov. 2004.

[84] D.-M. Chiu and R. Jain, "Analysis of the increase and decrease algorithms for congestion avoidance in computer networks," *Computer Networks and ISDN Systems*, vol. 17, no. 1, pp. 1–14, 1989.

[85] Z. Shen, J. G. Andrews, and B. L. Evans, "Adaptive resource allocation in multiuser OFDM systems with proportional rate constraints," *IEEE Transactions on Wireless Communications*, vol. 4, no. 6, pp. 2726–2737, 2005.

[86] R. Jain, D. Chiu, and W. Hawe, "A quantitative measure of fairness and discrimination for resource allocation in shared computer systems," Tech. Rep. TR-301, DEC Research, Sep. 1984.

[87] I. C. Wong, Z. Shen, B. L. Evans, and J. G. Andrews, "A low complexity algorithm for proportional resource allocation in OFDMA systems," in *Proc. IEEE Workshop on Signal Processing Systems*, pp. 1–6, 2004.

[88] J. Hui and Y. Zhou, "Enhanced rate adaptive resource allocation scheme in downlink OFDMA system," in *Proc. IEEE 63rd Vehicular Technology Conference - VTC Spring*, vol. 5, pp. 2464–2468, 2006.

[89] G. Yu, Z. Zhang, Y. Chen, P. Cheng, and P. Qiu, "Subcarrier and bit allocation for OFDMA systems with proportional fairness," in *Proc. IEEE Wireless Communications and Networking Conference - WCNC*, vol. 3, pp. 1717–1722, 2006.

[90] P. Cheng, G. Yu, Z. Zhang, and P. Qiu, "A cross-layer fair resource allocation algorithm for OFDMA systems," in *Proc. International Conference on Communications, Circuits and Systems*, vol. 2, pp. 1342–1346, 2006.

[91] A. Falahati and M. R. Ardestani, "An improved low-complexity resource allocation algorithm for OFDMA systems with proportional data rate constraint," in *Proc. 9th International Conference on Advanced Communication Technology*, vol. 1, pp. 606–611, 2007.

[92] S. Han, S. Kim, E. Oh, and D. Hong, "Adaptive resource allocation with rate proportionality tracking in OFDMA systems," in *Proc. IEEE 65th Vehicular Technology Conference - VTC Spring*, pp. 3031–3035, 2007.

[93] A. Furuskär, *Radio Resource Sharing and Bearer Service Allocation for Multi-Bearer Service, Multi-Access Wireless Networks: Methods to Improve Capacity.* PhD thesis, Royal Institute of Technology, Stockholm, Sweden, 2003.

[94] Y. Guan-ding, Z. Zhao-yang, Q. Pei-liang, and C. Peng, "Fair resource scheduling algorithm for wireless OFDM systems," in *Proc. International Conference on Communications, Circuits and Systems*, vol. 1, pp. 374–377, May 2005.

[95] F. Brah, J. Louveaux, and L. Vandendorpe, "Optimal resource allocation for MIMO OFDM-CDM systems under power, QoS and fairness constraints," in *Proc. Symposium on Communications and Vehicular Technology*, pp. 39–43, Nov. 2006.

[96] T.-D. Nguyen and Y. Han, "Schemes for maximal throughput and fairness in downlink OFDMA systems," in *Proc. First International Conference on Communications and Electronics - ICCE*, pp. 188–192, Oct. 2006.

[97] Y. Hu and L. Li, "A low-complexity resource allocation algorithm based on proportional fairness for multiuser OFDM systems," in *Proc. 4th International Conference on Wireless Communications, Networking and Mobile Computing - WiCOM*, pp. 1–4, Oct. 2008.

[98] P. Ameigeiras, J. Wigard, and P. Mogensen, "Performance of packet scheduling methods with different degree of fairness in HSDPA," in *Proc. IEEE 60th Vehicular Technology Conference - VTC-Fall*, vol. 2, pp. 860–864, Sep. 2004.

[99] G. Aniba and S. Aissa, "Adaptive proportional fairness for packet scheduling in HSDPA," in *Proc. IEEE Global Telecommunications Conference - GLOBECOM*, vol. 6, pp. 4033–4037, Nov. 2004.

[100] L. Yang, M. Kang, and M.-S. Alouini, "On the capacity-fairness tradeoff in multiuser diversity systems," *IEEE Transactions on Vehicular Technology*, vol. 56, pp. 1901–1907, Jul. 2007.

[101] H. Hou, W. Zhou, S. Zhou, and J. Zhu, "Cross-layer resource allocation for heterogeneous traffic in multiuser OFDM based on a new QoS fairness criterion," in *Proc. IEEE 66th Vehicular Technology Conference - VTC-Fall*, pp. 1593–1597, Sep. 2007.

[102] S. Doirieux, B. Baynat, and T. Begin, "On finding the right balance between fairness and efficiency in WiMAX scheduling through analytical modeling," in *Proc. IEEE International Symposium on Modeling, Analysis Simulation of Computer and Telecommunication Systems - MASCOTS*, pp. 1–10, Sep. 2009.

[103] M. Lee and S. K. Oh, "A simple scheduling algorithm capable of controlling throughput-fairness tradeoff performance," in *Proc. IEEE 70th Vehicular Technology Conference - VTC-Fall*, pp. 1–4, Sep. 2009.

[104] J. Mo and J. Walrand, "Fair end-to-end window-based congestion control," *IEEE/ACM Transactions on Networking*, vol. 8, pp. 556–567, Oct. 2000.

[105] M. A. Khan, R. Vesilo, I. B. Collings, and L. M. Davis, "Alpha-rule scheduling for MIMO broadcast wireless channels with linear receivers," in *Proc. Australian Communications Theory Workshop - AusCTW*, pp. 110–115, Feb. 2009.

[106] F. Kelly, "Charging and rate control for elastic traffic," *European Transactions on Communications*, vol. 8, pp. 33–37, 1997.

[107] A. Haider and R. Harris, "A novel proportional fair scheduling algorithm for HSDPA in UMTS networks," in *Proc. 2nd International Conference on Wireless Broadband and Ultra Wideband Communications - AusWireless*, pp. 1–7, Aug. 2007.

[108] J. Shi and A. Hu, "Maximum utility-based resource allocation algorithm in the IEEE 802.16 OFDMA system," in *Proc. IEEE International Conference on Communications - ICC*, pp. 311–316, May 2008.

[109] A. Sang, X. Wang, M. Madihian, and R. D. Gitlin, "A flexible downlink scheduling scheme in cellular packet data systems," *IEEE Transactions on Wireless Communications*, vol. 5, pp. 568–577, Mar. 2006.

[110] S. Shakkottai and R. Srikant, "Network optimization and control," *Foundations and Trends in Networking*, vol. 2, no. 3, pp. 271–379, 2007.

[111] M. Uchida and J. Kurose, "An information-theoretic characterization of weighted alpha-proportional fairness," in *Proc. IEEE 28th Conference on Computer Communications - INFOCOM*, pp. 1053–1061, Apr. 2009.

[112] G. Song and Y. G. Li, "Cross-layer optimization for OFDM wireless networks - part I: Theoretical framework," *IEEE Transactions on Wireless Communications*, vol. 4, pp. 614–624, Mar. 2005.

[113] G. Song and Y. G. Li, "Cross-layer optimization for OFDM wireless networks - part II: Algorithm development," *IEEE Transactions on Wireless Communications*, vol. 4, pp. 625–634, Mar. 2005.

[114] R. Agrawal, R. Berry, J. Huang, and V. Subramanian, "Optimal scheduling for OFDMA systems," in *40th Asilomar Conference on Signals, Systems and Computers - ACSSC*, pp. 1347–1351, Oct. 2006.

[115] P. Viswanath, D. N. C. Tse, and R. Laroia, "Opportunistic beamforming using dumb antennas," *IEEE Transactions on Information Theory*, vol. 48, pp. 1277–1294, Jun. 2002.

[116] R. Srinivasan and J. S. Baras, "An analysis of delay-constrained opportunistic scheduling for cellular wireless systems," Tech. Rep. TR 2004-2, Center for Satellite and Hybrid Communication Networks - CSHCN, 2004.

[117] M. Andrews, K. Kumaran, K. Ramanan, A. Stolyar, P. Whiting, and R. Vijayakumar, "Providing quality of service over a shared wireless link," *IEEE Communications Magazine*, vol. 32, pp. 150–154, Jun. 2001.

[118] S. Shakkottai and A. L. Stolyar, "Scheduling algorithms for a mixture of real-time and non-real-time data in HDR," in *Proc. of 17th International Teletraffic Congress (ITC*, pp. 793–804, 2001.

[119] Y. Zhang and S. C. Liew, "Link-adaptive largest-weighted-throughput packet scheduling for real-time traffics in wireless OFDM networks," in *Proc. IEEE Global Telecommunications Conference - GLOBECOM*, vol. 5, pp. 2490–2494, Nov. 2005.

[120] C. Gueguen and S. Baey, "Scheduling in OFDM wireless networks without tradeoff between fairness and throughput," in *Proc. IEEE 68th Vehicular Technology Conference - VTC-Fall*, pp. 1–5, Sep. 2008.

[121] A. R. Braga, E. B. Rodrigues, and F. R. P. Cavalcanti, "Packet scheduling for VoIP over HSDPA in mixed traffic scenarios," in *Proc. IEEE 17th International Symposium on Personal, Indoor and Mobile Radio Communications - PIMRC*, pp. 1–5, Sep. 2006.

[122] S. Ryu, B. Ryu, H. Seo, and M. Shin, "Urgency and efficiency based wireless downlink packet scheduling algorithm in OFDMA system," in *Proc. IEEE 61st Vehicular Technology Conference - VTC Spring*, vol. 3, pp. 1456–1462, May 2005.

[123] H. Lei, L. Zhang, X. Zhang, and D. Yang, "A packet scheduling algorithm using utility function for mixed services in the downlink of OFDMA systems," in *Proc. IEEE 66th Vehicular Technology Conference - VTC Fall*, pp. 1664–1668, Sep. 2007.

[124] G. Song and Y. G. Li, "Utility-based resource allocation and scheduling in OFDM-based wireless broadband networks," *IEEE Communications Magazine*, vol. 43, pp. 127–134, Dec. 2005.

[125] L. M. C. Hoo, B. Halder, J. Tellado, and J. M. Cioffi, "Multiuser transmit optimisation for multicarrier broadcast channels: asymptotic FDMA capacity region and algorithms," *IEEE Transactions on Communications*, vol. 52, pp. 922–930, Jun. 2004.

[126] P. A. Hosein, "QoS control for WCDMA high speed packet data," in *Proc. 4th International Workshop on Mobile and Wireless Communications Network*, pp. 169–173, 2002.

[127] D. P. Palomar and J. R. Fonollosa, "Practical algorithms for a family of waterfilling solutions," *IEEE Transactions on Signal Processing*, vol. 53, pp. 686–695, Feb. 2005.

[128] T. Lee, J. Yoon, S. Lee, and J. Shin, "Resource allocation analysis in OFDMA femtocells using fractional frequency reuse," in *Proc. IEEE 21st International Symposium on Personal Indoor and Mobile Radio Communications - PIMRC*, pp. 1224–1229, Sep. 2010.

[129] J. Espino, J. Markendahl, and A. Bria, "On spectrum allocation and interference management for WCDMA and OFDMA femtocells," in *Proc. IEEE 21st International Symposium on Personal Indoor and Mobile Radio Communications - PIMRC*, pp. 1561–1566, Sep. 2010.

[130] I. Demirdȫgen, I. Güvenç, and H. Arslan, "Capacity of closed-access femtocell networks with dynamic spectrum reuse," in *Proc. IEEE 21st International Symposium on Personal Indoor and Mobile Radio Communications - PIMRC*, pp. 1315–1320, Sep. 2010.

[131] S. Al-Rubaye, A. Al-Dulaimi, and J. Cosmas, "Cognitive femtocell," *IEEE Vehicular Technology Magazine*, vol. 6, pp. 44–51, Mar. 2011.

[132] T. Zahn, G. O'Shea, and A. Rowstron, "An empirical study of flooding in mesh networks," Tech. Rep. MSR-TR-2009-37, Microsoft Research, Cambridge, UK, 2009.

[133] H. Lim and C. Kim, "Flooding in wireless ad hoc networks," *Elsevier Computer Communications*, no. 24, pp. 353–363, 2001.

[134] G. Vivier, A. Agustin, J. Vidal, O. Muoz, and S. Barbarossa, "Femtocell-based network enhancement by interference management and coordination of information for seamless connectivity; deliv. 2.1: Scenario, requirements and first business model anaysis," Tech. Rep. ICT-248891, STP FREEDOM, Jun. 2010.

[135] UMTS, "Selection procedures for the choice of radio transmission technologies of the UMTS," Tech. Rep. TR 101 112 V3.2.0 - UMTS 30.03, Universal Mobile Telecommunications System (UMTS), Sophia Antipolis, France, Apr. 1998.

[136] T. S. Rappaport, *Wireless Communications: Principles and Practice*. Prentice Hall, 2nd ed., 2002.

[137] W. C. Jakes, *Microwave Mobile Communications*. John Wiley & Sons / The Institute of Electrical and Electronics Engineers (IEEE), 1994.

[138] 3GPP, "Deployment aspects," Tech. Rep. TR 25.943, Release 8, 3rd Generation Partnership Project, 2008.