Resource Allocation and Flexible Scheduling in Wireless Networks

Mohammad Shaqfeh

A thesis submitted for the degree of Doctor of Philosophy.
The University of Edinburgh.
July 2008

Abstract

Many scheduling schemes have been proposed in the literature to control how different users access the channel in a wireless network. Channel-aware schedulers exploit the measurements of instantaneous channel conditions of the different users to obtain throughput gains by proper allocation of the wireless resources based on the channel state information. However, to the best of my knowledge, it is still not satisfactorily discussed in literature how we can define efficiency measures for the scheduling schemes. There are open questions such as: How can we judge how good a scheduling scheme is? And is it possible to find generic schedulers that are always optimum regardless of the specific operating point of the network defined by the different users' allocated rates and quality-of-service constraints. This thesis discusses the topic of organising the multiuser operation in centralised wireless networks.

The design of channel-aware scheduling algorithms for wireless fading channels involves two main objectives: efficient allocation of the scarce wireless resources and achieving suitable fairness criteria and quality-of-service requirements of the different applications. It is demonstrated in this thesis that both objectives can be achieved at the same time by scheduling concepts that are based on multiuser information theory.

A new generic mathematical framework is proposed to evaluate the performance limits of channel-aware scheduling algorithms for delay-tolerant applications, and to compare them. The efficiency of the schedulers in allocating the system resources is compared against the theoretically achievable optimum for the operating point chosen. For use in a case study, a variety of scheduling schemes are described in a novel unified way, which allows for the direct application of the mathematical framework introduced. The practically relevant case, in which system constraints enforce the use of an orthogonal channel-access scheme and constant transmission power, are considered. As an illustrative numerical example the two-user case is analysed, although, qualitatively, the results carry over to the M-user case.

Furthermore, a practical scheduler structure, which is always efficient and at the same time can be flexibly controlled by the network operator according to fairness constraints, is suggested.

Declaration of originality

I hereby declare that the research recorded in this thesis and the thesis itself was composed and originated entirely by myself in the School of Engineering and Electronics at The University of Edinburgh.

Acknowledgements

First of all, I would like to extend my thanks to the University of Edinburgh for giving me the opportunity to do my Ph.D. research degree at a high standard. Special thanks are sent to the Institute for Digital Communications for the interesting 3-year period which I spent as a member of their group.

In particular, I'd like to express my appreciation with respect and profound gratitude for my supervisor Dr. Norbert Goertz for his faith and confidence on me which has pushed me to try to do well. Thanks for guidance, instructions, help and willingness to answer my questions with patience and understanding.

I would like also to thank Prof. Steve McLaughlin, who is my second supervisor, and Dr. John Thompson. The support of MobileVCE is highly appreciated, and many thanks are sent to my colleagues in the Core 4 research program of MobileVCE.

Thanks to anyone who has ever taught me anything and to all those who have contributed to this work in one way or another.

Finally, I should not forget my grandmother because she deserves special admiration and love. My thanks are also sent to my most beloved brother, kindest sisters and all my true friends as well as my dear relatives. Above all, this achievement is gratefully dedicated to my parents, my wife and my little son whose prayers, love, enthusiasm, support and encouragement have always led me to success.

Contents

List of figures

List of tables

Acronyms and abbreviations

AMC	Adaptive Modulation and Coding
AP	Access Point
AWGN	Additive White Gaussian Noise
BC	Broadcast channel
BER	Bit Error Rate
BS	Base Station
CDF	Cumulative Distribution Function
CDMA	Code Division Multiple Access
CP	Constant Power per channel block
CSI	Channel State Information
DSA	Dynamic Sub-carrier Allocation
FDMA	Frequency Division Multiple Access
ICC	International Conference on Communications
IEEE	Institute of Electrical and Electronics Engineers
MAC	Multiple Access Channel *or* Medium Access Control
MIMO	Multiple Input Multiple Output
OFDM	Orthogonal Frequency Division Multiplexing
OFDMA	Orthogonal Frequency Division Multiple Access
OPT	Optimal (no auxiliary constraints)
OS	Orthogonal Signalling
PDF	Probability Density Function
PHY	Physical
PF	Proportional Fair
QoS	Quality of Service
SC	Superposition Coding
SIC	Successive Interference Cancellation
SISO	Single Input Single Output
SNR	Signal to Noise Ratio

SPECTS	International Symposium on Performance Evaluation of Computer and Telecommunication Systems
SU	Single User selection per channel block
TDMA	Time Division Multiple Access
WLAN	Wireless Local Area Network
WWRF	Wireless World Research Forum

Nomenclature

Boldface is used to indicate vectors.

ar_i	Long term average channel access ratio allocated to user i
$ar_i^{(l)}$	Computed access ratio of user i in iteration l of the off-line algorithm
$\widehat{ar_i}[k]$	Measured access ratio of user i at a time window past time index k
B	Number of frequency (flat faded) slots within the total bandwidth
$B_i[k]$	Number of frequency slots in which user i is scheduled in a given time
$f_{H_i}(\,)$	Probability density function of the channel statistics of user i
$F_{H_i}(\,)$	Cumulative density function of the channel statistics of user i
$F_{R_i}(\,)$	Cumulative density function of the user i feasible transmission rates
h	Channel quality (power gain: ratio between received and transmitted power)
$h_i[k]$	Channel quality of user i at channel block k
$\bar{h}_i$	Average channel quality of user i
k	Channel block index *or* time index (in Chapter 5)
m	Index of scheduled user
M	Total number of users in the system
$n_i[k]$	Gaussian noise with zero mean of i-th receiver at channel block k
$\bar{P}$	Long term average power constraint (in Watts/Hz)
P_i	Long term average power allocated to user i
$P_i[k]$	Allocated power to user i in channel block k
$P_{\text{sum}}[k]$	Sum of the allocated power to all users at channel block k
P_T	Transmitted power (For simplicity, P means P_T)
P_R	Received power
Pr	Probability
R_i	Long term average rate (capacity) of user i (in bits/sec/Hz)
$\bar{R}_i$	Average feasible rate of user i (For simplicity, R_i means $\bar{R}_i$)
$\tilde{\bar{R}}_i$	Average rate of user i over the channel blocks in which he is scheduled
$R_i[k]$	Feasible rate of user i in channel block k
R_i^{OPT}	Average rate of user i under no auxiliary constraints
R_i^{CP}	Average rate of user i under constant power per channel block constraint
R_i^{SU}	Average rate of user i under single user selection per channel block constraint

$R_i^{\text{CP-SU}}$	Average rate of user i under both CP and SU auxiliary constraints
R_{sum}	Long term average sum rate of all users (sum capacity)
$T_i[k]$	Average throughput of user i in a past window prior to channel block k
$x_i[k]$	Transmitted signal from user i at channel block k
$x[k]$	Transmitted signal from the base station at channel block k
$y_i[k]$	Received signal by user i at channel block k
$y[k]$	Received signal at the base station at channel block k
$u_i(x)$	Marginal utility function (rate revenue minus power cost)
t_c	Time constant
Δ	Step size
μ_i	Weighting factor of user i (used in scheduler optimisation)
λ	Lagrangian multiplier (power price parameter)
v_i	rate offset of user i (used in scheduler optimisation)
$\tau_i[k]$	Sharing ratio of user i in channel block k when orthogonal signalling is applied
x^+	$\max(x, 0)$
$E[\,]$	Expected value
log	Logarithm to base 2
ln	Natural logarithm

Chapter 1
Introduction

This introductory chapter provides a motivation to the work presented in the thesis by highlighting the importance of the topic and the main challenges in the field. The objectives of this research are presented and the outline of the thesis structure is given. Furthermore, the main contributions of this work are summarised alongside a list of publications related to this Ph.D. research project.

1.1 Resource Allocation in Wireless Networks

With the rapid growth of wireless technology in the past few years and the increasing demand to support a mix of real-time applications and data traffic as well as to achieve higher rates (bits/sec) at high quality-of-service (QoS) requirements – matching those that can be obtained in wired networks – a lot of design issues and challenges have become open research fields which needed to be addressed adequately. *Improving the efficiency of wireless networks* is a very important objective for wireless networks' operators (i.e. service providers) in order to be able to support higher service rates at lower costs and thus to maximise their financial profits. Furthermore, the growing popularity of both fixed and mobile wireless applications such as in cellular phone systems or in wireless local area networks (WLAN) motivates researchers in the field to study the possible means to improve the performance of these systems in order to satisfy the users with the offered quality and cost of service.

A fundamental problem in networking is the allocation of the limited resources among the users of the network taking the service requirements of the supported applications over the network into consideration. It is a fact that optimising the performance of resource allocation schemes plays an important role in the improvement of the overall efficiency of the system. However, there is a big diversity in the approaches used in the literature to design resource allocation and multiuser scheduling schemes for wireless systems. During the last decade, the design of multiuser scheduling schemes for wireless networks has been extensively studied. Algorithms that were state-of-the-art ten years ago [1], [2], [3], [4], [5], [6] are now outperformed by

more recent schemes. In general, scheduling algorithms have two main objectives: (i) efficient allocation of the scarce wireless resources to achieve throughput gains and (ii) to define suitable solutions for the fundamental trade-offs existing in the multiuser communication problem, that is the trade-off between sum-throughput and individual throughputs and the trade-off between throughput and delay constraints. Since the network operator may have different objectives, depending on the applications supported by the network and the way "fairness" is defined, one scheduling algorithm can be most suitable for a certain scenario with its specific definition of fairness while it performs poorly elsewhere. Although not uncommon in the literature, it does not make real sense to compare, e.g., the total throughput achieved by two schedulers, when one is designed to maximise throughput while the other one is designed to provide every user with at least a minimum rate whenever this is at all possible; both designs correspond to different operating points. This raises the question how "efficiency" of a scheduler can be measured and compared in general terms. To the best of my knowledge, this question has not been answered satisfactorily in the literature, and one of the main goals of this work is to contribute to close this gap.

One important question is to define and characterise the intersection between the design of proper resource allocation schemes and improving the overall efficiency of wireless systems. Unlike the classical point-to-point communication problem, the appropriate definition of overall efficiency of multiuser systems is ambiguous. For example, sum throughput (bits/sec) can not be used as a measure of the overall efficiency of a multiuser system. Having satisfying clear answers to the appropriate definition of overall efficiency of wireless systems will help to define the best criteria to be used in the optimisation of the resource allocation schemes.

The tasks of the scheduler in a wireless network include the decision of who should access the network when (i.e. user selection to access the channel). However, in order to get considerable gains in the overall throughputs achieved by the system, the user selection task should be integrated with the air interface techniques. Hence, the selection of the scheduled user(s) should be accompanied by the rate allocation of the scheduled user(s). To support different rates, the transmitted power as well as the modulation and coding schemes should be flexibly controlled according to the scheduler decisions.

Although many scheduling schemes which are capable of supporting quality of service (QoS) differentiation and guarantees have been developed for wire-line networks, those are not suitable for wireless networks. Examples of wireline scheduling protocols [7], [8] include the

schemes for fair queueing [9], virtual clock [10], self-clocked fair queueing [11], and earliest-due-date [12] ones. The wireless link creates unique problems and challenges which need to be solved adequately.

1.1.1 Challenges created by wireless channels

Wireless channels [13], [14], [15] have the following characteristics and challenges which make the use of those scheduling schemes that are applied in wired networks inappropriate:

- *The air-link resources in the wireless networks are scarce.* The system must operate within the allowed air-link frequency bandwidth for the network. Furthermore, paying to get license to use more frequency bandwidth is very expensive. This limitation of the air link resources raises the importance of increasing the efficiency of the resource allocation schemes and seeking to operate at the capacity limits. Such a problem does not create a fundamental challenge in wired networks since increasing the capacity of the network can be supported by adding more cables. Since the signal transmission in wired medium is guided, and the interference between cables is negligible, the addition of more cables with higher capacities does not create any technical problem. This is obviously not the case in wireless networks.

- *The wireless medium is a shared medium* which complicates the multiuser operation relative to the wired medium case. Communication to and from each user in wireless multiuser networks have to be carried out in such a way that all other users can communicate as well, i.e., the physical resources have to be shared efficiently. This greatly complicates the communication problem compared with point-to-point scenarios considered by classical information theory. Although in some wired networks, the case of sharing the transmission between users through a cable is relevant, sharing over the wireless medium has more challenges due to its variation over time, frequency and space, as explained below.

- *Wireless channels suffer from fading, shadowing and interference and hence they vary over time, frequency and space* [13]. There are small-scale (or short term) fast variations due to fading resulting from the multi-path nature of the wireless link, the mobility of the users' terminals and possible bursty interference. In addition, there are large-scale (or long term) slow variations depending on the terminal location and interference levels.

Each one of these two types of channel variations creates new challenges:

- Due to the fast channel variations over time and frequency, it becomes inefficient to use conventional channel access schemes such as frequency division multiple access (FDMA) or time division multiple access (TDMA) accompanied with Round-Robin schedulers in which the users access the network on a regular basis regardless of their instantaneous channel conditions. Applying such schemes over wireless channels will result in higher probability of the situation of transmitting to the users while their channels are in a bad state. Thus, the scarce wireless resources will be wasted over bad quality channels. Obviously, such a case does not appear over a wired link since the channel quality is almost constant over time.
- The location-dependent long-term variations in the channel conditions create a fairness problem. This is because the users' terminals will have different long-term average channel qualities. As a consequence, the users with worse channels will need to get higher percentage of the resources and to access the channel more frequently in order to get the same throughput (bits/sec) achieved by the users with better channel conditions. However, this will severely degrade the throughput that can be achieved by the good channel users as well as the overall sum throughput of the network. The definition of fairness is ambiguous in this case as there is a contradiction between achieving throughput fairness between the users and maintaining resource sharing fairness between them. If the resources are divided fairly between the users, then every user will have throughput rates according to his average channel quality. This approach prevents degrading the performance of good-channel users to help bad-channel users. Indeed, the fairness criteria depend on the specific application with its QoS constraints. Some applications require constant service rates while other applications require as high service rates as possible.

1.1.2 Other challenges for the resource allocation in wireless networks

In addition to the challenges created by the characteristics of the wireless channel, there are some other challenges which should be taken into consideration when designing resource allocation (i.e. multiuser scheduling) schemes for wireless systems.

1.1.2.1 Conflicting requirements in the system

The contradiction between maximising the total throughput over the network and achieving fairness between users is just one example of the challenge of the conflicts between the requirements which the network operator would aim to achieve. Thus, many trade-offs are needed to deal with the contradicting objectives. Another example is the contradiction between the throughput and the QoS constraints. Having strict delay constraints and very low tolerable error rate requirement will necessarily lead to degradation of the achievable throughput. Thus, a careful decision regarding the constraints to be maintained is needed.

1.1.2.2 New emerging technologies and techniques

In the last few years, there have been a lot of new promising techniques that would allow to improve the performance of the wireless systems. Using multiple antennas at the transmitters or the receivers is one example of the emerging technologies which could help to increase the capacity over a wireless link [14]. Another example is using relays (i.e. multi-hop link) between the user terminal and the base station (BS) or access point (AP) of the network [16],[17]. There have been also many advances in the design of the transceivers and their capabilities to apply multi-carrier channel-access schemes and adaptive modulation techniques as well as to have the ability to control the transmitted power. Of course using such emerging technologies is advantageous. However, it is needed to properly study how to organise the multiuser operation depending on the supported technologies in the system. The optimal scheduling scheme is highly dependent on the specific system constraints. Thus, whenever a new technique is applied, the topic of designing optimal scheduling schemes should be re-visited. However, the fundamentals will be the same, and that is what is studied in this work.

1.1.2.3 Flexibility in controlling the operating point of the system

The design of schedulers for wireless networks should provide the network operator with the required flexibility in controlling the operating point of the system (i.e. the amount of resources and the rate allocated to each user as well as the appropriate QoS constraints based on the served applications). Furthermore, the service provider may decide to differentiate between the users based on, e.g., a pricing policy.

1.1.3 Approaches to improve the performance of wireless networks

1.1.3.1 Exploiting the multiuser diversity

Multiuser diversity [14] is an important approach to get considerable gains of achievable throughputs in wireless networks with many users having independently fading channels. The concept is to track the channel conditions of the users and to utilise the wireless resources in a good way by selecting to transmit to or from a user's terminal when its channel is at good condition. Thus, wasting resources by transmitting on bad channels conditions is avoided. This is sometimes called opportunistic communication; i.e. utilising a channel during the opportunities of good condition. The multiuser diversity gain is obtained because the independent fading characteristic of the users' channels raises the probability of having a user with very good channel condition at each time instance. This probability increases as the number of users increases. The multiuser diversity gain is hence obtained due to the smart scheduling of the users by exploiting their fading channels' conditions. However, it should be kept in mind that obtaining this gain requires perfect tracking of the channels and the ability to adapt the transmission rates according to the channel conditions.

The multiuser diversity was motivated by the work of Knopp and Humblet [18]. In the theoretic approach presented in their paper, it was shown that in order to maximise the information capacity of the uplink in single-cell multiuser communications with frequency-flat fading, one user only should be allowed to transmit at any given time. This is the user with the best channel condition. The user keeps transmitting using the whole bandwidth as long as he has the best channel among the users. Although that work was discussing the uplink communications in the network, the statement was extended to the downlink case in [19].

Although transmitting using the best channel maximises the system overall throughput, there are two problems which should be solved. These are maintaining fairness among users and meeting delay constraints of the served applications. In [20] a solution to tackle the problem of fairness while achieving multiuser diversity gains was suggested. The suggested scheduler in that paper was called the proportional fair scheduler. However, the scheduler has the drawback that it does not provide the network operator with the flexibility in controlling the operating point of the system.

1.1.3.2 More advanced adaptive air-interface (physical layer)

In order to achieve the aim of approaching capacity limits over wireless multiuser channels, an advanced physical layer is required. This includes three main requirements:

- Channel state measurement and prediction [21], [22]: In order to be able to exploit the multiuser diversity in wireless networks, it is needed to track the channel conditions continuously and then to forward the channels conditions measurements to the base station of the network where the scheduling decisions are taken. This technique has already started to be implemented in wireless networks (see for example [23]).
- Dynamic sub-carrier allocation (DSA): In order to obtain maximum possible multiuser diversity gains, the multiuser diversity should be exploited in the frequency domain (due to frequency-selective channels in the total transmission bandwidth) in addition to the time domain. The orthogonal frequency division multiple-access (OFDMA) scheme enables the exploitation of multiuser diversity in the frequency domain *and* in the time domain [24]. Therefore, OFDMA systems achieve better performance than spread-spectrum (CDMA-based) systems that use one spreading code across the whole spectrum. Furthermore, the system should have the capability of allocating the sub-carriers to all users dynamically based on channel conditions.
- Adaptive modulation and coding (AMC) [25], [26]: In order to achieve capacity limits, the optimal scheduling decisions should be accompanied with the selection of the modulation [13] and coding [27] schemes which are capable of achieving the capacity limits over each time and frequency slot. Since, the instantaneous capacity limits are changing over time and frequency, the scheduling decision should be integrated with the appropriate adaptive modulation and coding schemes.

1.1.3.3 Cross layer design

The layered network architectures which have served well for wired networks are not directly suitable for wireless networks. This is because the layered architecture divides the overall networking task into independent layers. Every layer has the task of providing some services or functionalities independently from other layers. It is obvious that not all the tasks which are independent in a wired medium are so in a wireless medium. The wireless medium has its unique problems and modalities which are different than in the wired medium. Thus, it is

argued that when designing multiuser wireless networks, exploiting the dependence between the layers can lead to potential performance gains. This is the main motivation behind the cross-layer design approaches [28], [29], [30], [31], [32].

From this perspective, the scheduling task should not be considered as a functionality of the medium access control (MAC) layer carried out independently of other network stack layers. Multiple access and medium access strategies should no longer be considered as independent tasks. The scheduler should be integrated with the physical (PHY) layer (air-interface) modulation and coding schemes as well as channel access scheme. Furthermore, the scheduling decisions require channel-state-information (CSI) which can be obtained from the physical layer as well as QoS constraints which should be known from higher layers. It is also the case that the scheduling task and the admission control task can be integrated together.

1.2 Thesis Objectives

The main goal of this research work is to *give important guidelines for the design of efficient flexible scheduling schemes for centralised wireless networks.* The objectives of this research include:

- to discuss the appropriate approach to formulate the problem of optimising the performance of scheduling schemes in wireless networks.
- to give the optimal scheduling schemes for a variety of systems' implementation constraints (such as orthogonal multiple-access and constant transmission power).
- to systematically analyse and compare channel-aware scheduling schemes known in the literature.
- to develop algorithms to control the optimal schedulers flexibly according to the network operator's objectives.

1.2.1 General assumptions

In centralised wireless networks, which are under consideration in this work, the scheduling decisions for both the uplink and the downlink are taken at the wireless access point (AP) or the base station (BS) of the network. This unit is provided with a rich set of information such

as the traffic load and the different QoS requirements of the served traffic classes as well as the instantaneous channel conditions of the wireless links to active users. In this work, it is assumed that the channel variations are not too fast so that the effect of the channel-measurement delay [33] is negligible and the channel coefficients can be estimated at the receiver and be communicated to the transmitter with sufficient accuracy at low overhead. This also means that coherent detection can be performed at the receiver, i.e., with no phase error and the "I" and the "Q" components are both scaled by the magnitude of the channel's fading coefficient which is the square root of the channel power gain.

Furthermore, it is assumed in the analysis that the schedulers are applied in generic systems such as in OFDMA systems. The scheduling decisions, i.e. the decisions of the users who are allowed to access the channel, are taken for each *channel block* over which the channel fading coefficients can be considered to be constant (i.e. flat faded) as shown in Figure 1.1. Each channel block may consist of many *slots* in time and many subcarriers over which the channel fading coefficients do not change as the slots belong to one channel block.

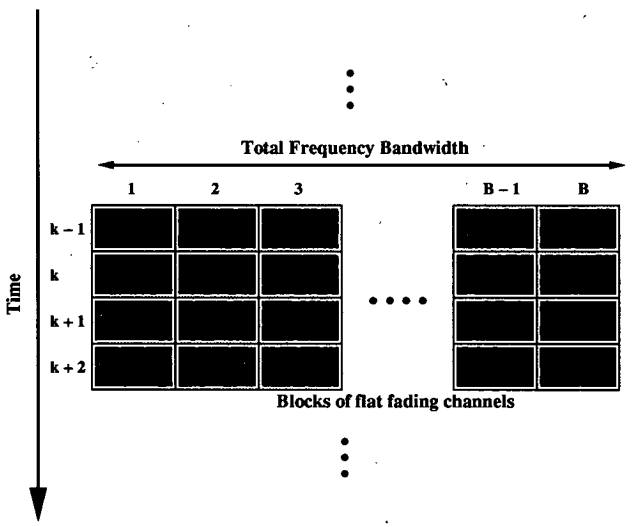


Figure 1.1: Dividing time-frequency domain into blocks of flat fading channel conditions

The systems under consideration are assumed to support dynamic sub-carrier allocation (DSA) and adaptive modulation and coding (AMC), and hence they can **approach the capacity limits over each flat-faded channel block**. Of course, the design and implementation of physical layer schemes capable of achieving the capacity limits over the flat-faded channel blocks is a challenging task and needs extensive study. However, the results known from the information theory literature show that with the use of the block fading channel model, the two tasks of

(i) allocating resources to the users over the channel blocks (i.e. multiuser scheduling), and (ii) achieving the capacity over each of the channel blocks, are both needed to achieve the multiuser channel capacity limits. However, **these two tasks can be done separately** since the physical layer is assumed to be adaptive and flexible and thus capable of achieving the capacity regardless of the instantaneous rates allocated to each user over the channel blocks. Refer to Chapter 3 for further details. Thus, it is sensible to use the assumption of being able to achieve the capacity limits over the flat-faded channel blocks while studying the optimality of multiuser schemes. On the other hand, if all the physical layer design parameters such as modulation constellation, code rate, etc are assumed to be controlled jointly by the scheduler, the problem will become very complicated. Furthermore, this assumption is actually unnecessary.

In this work, the physical layer implementation issues are not taken into consideration. This is a big research task which can be done separately of this work. [14] provides a good review on physical layer schemes achieving capacity over fading channels. We carried out some initial work on implementation issues of coding schemes (refer to Section 1.4.1). However, in order to concentrate on the challenging topic of multiuser scheduling, no further research was done towards the implementation issues.

Some of the scheduling algorithms analysed in this work were originally proposed and applied for single-carrier systems (e.g. the proportional fair scheduler [23]). However, these scheduling algorithms are, on a sub-carrier level, also applicable to OFDMA systems.

For simplicity, single-transmit single-receive antenna (SISO) systems are taken into consideration, but with a variety of possible air-interface constraints such as power control and access scheme per channel block. However, the analysis framework can be extended to multiple-antenna (MIMO) systems, but the analysis is more complicated and this would distract from the main issues (scheduling and resource-allocation) discussed in this research project.

In order to visualise the performance and efficiency of the scheduling algorithms considered, the two-user case is often considered in this thesis because of the simplicity of its rate-region plots. However, the main results and conclusions of this work are, of course, also applicable to the general M-user case, and qualitatively the results carry over from the two-user case.

1.3 Thesis Structure

In Chapter 2 the optimality criteria for resource allocation and multiuser scheduling schemes in centralised wireless networks is discussed. This chapter forms as the basis for the results in the following chapters of the thesis. A discussion of the appropriate definition of overall system efficiency is given, followed by a universal framework to define *optimality* or *efficiency* of scheduling schemes.

The optimal resource allocation and multiuser scheduling schemes under different system constraints are presented in Chapter 3 based on known results from the information theory literature in addition to many novel contributions. The results are provided for both the downlink (i.e. transmission from the AP towards the users' terminals) and the uplink (i.e. transmission from the users' terminals towards the AP) cases. The system constraints which are taken into considerations include the specific channel-access schemes and power control schemes used in the system. Comparisons of the performance of the system when auxiliary constraints are applied are provided in numerical examples.

A comparison-study of well-known scheduling polices that are applicable for delay-tolerant applications in centralised wireless networks is given in Chapter 4. This comparison-study applied to schedulers known from the literature is helpful in order to make the theory presented in Chapter 2 clear. Furthermore, the results of the comparison study highlight the importance of the suggested concepts in this thesis, as it is demonstrated that some well-known schedulers are actually inefficient in the sense that their performance is bounded far away from the appropriate point on the boundary of the capacity/rate region. A new mathematical framework is presented in order to be able to evaluate the performance of the schedulers. Algorithms that consider strict delay-constraints are not included in the analysis. However, the main concepts of the thesis are also applicable in this case.

In Chapter 5, a flexible and efficient multiuser scheduling scheme is suggested. The flexibility of the presented scheduler is in the ability to control the percentage of the bandwidth resources given to each user according to the network operator's own criteria. Two algorithms – an offline and an online solution – to meet the flexible resource-sharing constraints are suggested. The ability of the suggested scheduler to track the channel variations and to maintain the resource-sharing constraints is verified in the simulation results. Furthermore, the suggested scheduler is also analysed based on the achievable multiuser diversity gains.

Finally, the main messages of this work are summarised in Chapter 6. It highlights the important guidelines which are needed for the design of optimal resource allocation schemes for wireless networks. Some suggestions and recommendations to continue this important research are presented as well.

1.4 Contributions and Publications

The main *contributions and novelties* of this work include:

- To the best of my knowledge, the suggested framework to define optimality of multiuser scheduling schemes has not been satisfactorily highlighted and presented in the literature, as there are still many new proposals in the field which are actually inefficient.
- The literature review of the optimal resource allocation schemes known from the information theory.
- Modifications to well-known results from the information theory literature such as the characterisation of the boundary of the ergodic capacity region of Gaussian multi-access fading channels [34], [35].
- Novel contributions to multiuser information theory including the characterisations of the multiuser capacity region under different systems' constraints. The optimal resource allocation under some auxiliary constraints are also novel. Furthermore, there are new approaches and examples used to properly present the theory.
- A new approach to survey the important multiuser scheduling polices known from the literature. This includes the extension of some polices and presenting them in generic forms in which any possible operating point of the system can be achieved.
- A unified mathematical framework to compute the capacity region of scheduling polices for delay-tolerant applications. The framework is applied in numerical examples.
- The design of a flexible scheduler achieving resource-sharing constraints. The scheduler is simulated and analysed to compute the achievable multiuser diversity gains of the scheduler.

The results of this research project have been published in several important international conferences or submitted to highly-reputed journals. Here is a list of the papers related to this work:

- [36] was presented in the IEEE ICC 2008 in Beijing. It was mainly based on the results which are presented in Chapter 5 of the thesis. A journal version of that paper including the analysis of the multiuser diversity gains is under preparation.
- [37] was presented in the SPECTS conference in Edinburgh, 2008. The paper was based on the comparison-study of known scheduling polices, which is presented in Chapter 4 of the thesis.
- [38] was presented in the WWRF meeting in Ottawa, 2008. The paper gives an overview of the topic by discussing the efficiency measure of the multiuser network operation, and giving some examples of scheduling policies that achieve the optimal operating points as well as an introduction to the resource-sharing fair scheduler presented in [36].
- [35] has been accepted for publication in IEEE Transactions on information theory. In this paper, a modification is proposed for the formula known from the literature that characterises the boundary of the capacity region of Gaussian multiaccess fading channels. The modified version takes into account potentially negative arguments of the cumulative density function that would affect the accuracy of the numerical capacity results.
- [39] has been submitted to IEEE Transactions on Communications. This paper provided a more detailed version of the work presented in [37]. It is mainly based on Chapter 2 and Chapter 4 as well as some parts of Chapter 3.
- [40] is under preparation to be submitted to a journal on information theory. In this paper the ergodic capacity region of block-fading Gaussian multiuser channels under auxiliary constraints is presented. The results are based on Chapter 3 of the thesis.

1.4.1 Prior work on encoding of LDPC codes

My Ph.D. project has formed part of the Mobile VCE (www.mobilevce.com) Core 4 research program. I have been part of the *Delivery Efficiency* Work Area and my work package is *E3 - Joint Link and System Optimisation.*

I started my Ph.D. in October 2005 and my task with Mobile VCE in January 2006. Prior to starting my project with Mobile VCE, I was investigating encoding of LDPC codes and I had novel contribution in that field. My work [41] was accepted and presented in the IEEE ICC 2007 in Glasgow. In [41], an algorithm for efficient encoding of LDPC codes is presented that does not impose any restrictions on the construction of the parity-check matrices. The algorithm modifies the parity check matrix, without changing the subspace spanned by its rows, by removing linear dependent rows and adding a small number of new rows such that the graph-based message-passing *en*coder will not get stuck in a stopping set. The added rows are designed by a new algorithm which is based on the notion of the "key set". The encoder exploits the sparseness of the parity-check matrix, and the encoding complexity grows almost linear with the blocksize, because the number of added rows, which may not be sparse, is relatively small.

Although my research on LDPC codes was completed during my Ph.D. work in The University of Edinburgh, I am not including it in the thesis as it is a different field than the main topic presented in the thesis.

Chapter 2
Optimality of Multi-user Scheduling Schemes

A fundamental characteristic of multiuser communication over wireless fading channels is the existence of many trade-offs between the system capacity, quality of service (QoS), the cost to achieve capacity in terms of physical air-link resources, etc. To assess the overall efficiency of a wireless network, it is needed to take into consideration all the potentially conflicting requirements. In the literature, there are many performance metrics that are used to measure the performance of a wireless system such as the total throughout in (bits/sec), the average spectral efficiency in (bits/sec/Hz), the outage/ blocking probability, the bit (frame) error rate, fairness between users, or QoS measures such as average packet delay or queue size, etc. Due to the conflicts between these measures, it becomes unclear how to define and measure the overall efficiency of the system. As a consequence, there are many approaches applied in the literature to design multiuser scheduling schemes or to judge how good a scheduler is.

For example, in [42] it was suggested to formulate the optimisation problem of the multiuser scheduler by maximising the average total (i.e. sum of all users) system performance (i.e. rate) with constraints on the "time-fraction" (channel-access rate) of the users, which are pre-assigned based on fairness criteria. Another way to optimise the multiuser scheduler was suggested in [43]. The concept was to track the channels' statistics and to schedule the user whose feasible rate (i.e. based on channel quality) in a given time slot is high enough, but least likely to take even larger values in other (e.g. future) time slots. There are more examples on the way the scheduling problem is formulated in the literature. It may appear that all formulations of the optimisation problem are appropriate. However, as discussed in Chapter 4, both concepts in [42] and [43] are not the best way to design scheduling schemes for wireless networks.

A discussion of the most appropriate approach to formulate the problem of resource allocation for wireless networks is presented in this chapter.

2.1 System Model and Existing Trade-offs

A wireless network, like any other system, can be represented by the simple generic model shown in Figure 2.1. The inputs of the system, in this specific example of a wireless network, are the physical resources used for transmitting the information from the transmitter(s) to the receiver(s). They include mainly the frequency bandwidth, the time and the power[1]. The output of the system is the overall average information transmission rate through the network in (bits/sec). The system can be controlled by adjusting the QoS constraints in terms of tolerable packet delay and error rate, and by managing the multiuser operation in terms of the percentage of the overall service rate assigned to each user.

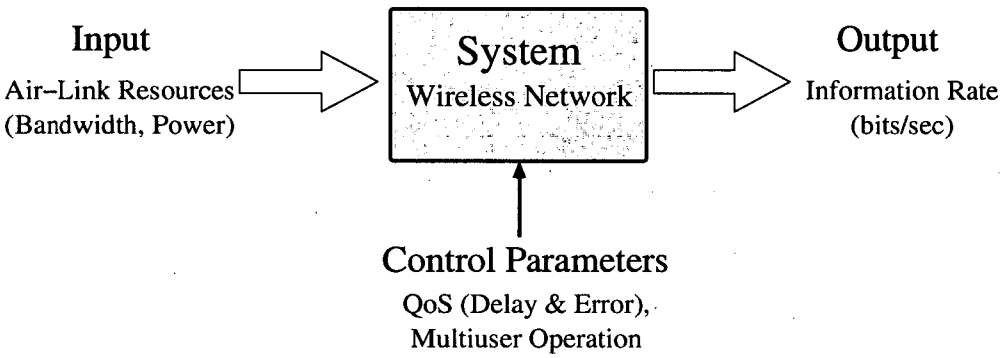


Figure 2.1: Basic system model which is applicable for a wireless network.

The operation of the system can be properly defined by characterising the relation between the input and the output, and the effects of tuning the system operation by adjusting the control parameters. There are four main trade-offs existing in the operation of a wireless network, and those are shown in Figure 2.2.

2.1.1 Main trade-offs in the operation of a wireless system

2.1.1.1 System capacity vs. QoS constraints

In general, to maximise the system capacity, the instantaneous transmission rates should be adjusted based on the measured channel conditions. More resources are allocated over the instances when the channel is good so that, in the long term, the average rate is increased. However, this means that maximum tolerable packet delay constraints are likely to be violated. Similarly, by maintaining the delay constraints, the system achievable capacity will be

[1]Space is another physical resource in MIMO systems

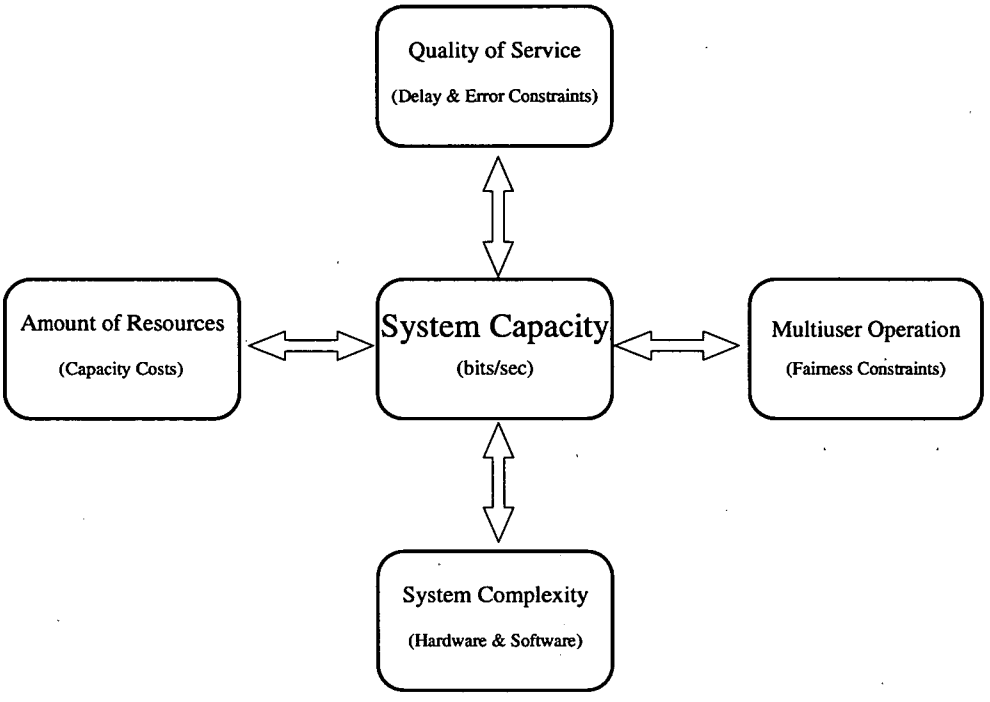


Figure 2.2: Main trade-offs existing in a wireless system.

degraded. For example, if the system should maintain a constant minimum rate all the time with no tolerable delay, the power should be controlled in every time slot to maintain the rate regardless of the channel condition. Thus, more resources are needed during the bad-channel conditions in order to compensate for the loss in the channel quality. This is known in the literature as channel inversion [44]. It is obvious that channel inversion to maintain QoS works differently to the strategy which should be applied in order to maximise the system long-term average capacity. In [44], the achievable capacity under channel inversion is analysed and compared with the ergodic capacity in numerical examples with different fading channel statistics. It is demonstrated that the performance is degraded by applying strict delay and minimum rate constraints.

On the other hand, some of the applications served over wireless networks such as voice telephone calls require that these QoS constraints be applied, and hence these constraints can not be ignored. A good approach to prevent big degradation in the capacity of the system is to relax the QoS constraints a bit by, e.g., allowing for some outage in the service during severely bad channel conditions. Indeed, this is an important research field, which is out of the scope of this work, to fully characterise and define the loss of the achievable capacity when specific QoS

constraints are applied. Strict delay constraint affects the capacity of the system even over non fading channels because short codewords are needed in this situation. With shorter codewords, an additional power margin more than the actually needed will be allocated to maintain the required minimum bit or frame error rate (BER).

Another example of the QoS constraints is the tolerated error in the information transmission. A suitable diversity-multiplexing trade-off [45], [46] should be used based on tolerable error constraints of the served applications. A diversity gain means more reliability in the transmission and thus less error, while a multiplexing gain means more amount of traffic (i.e. increased capacity). Full characterisation of the relation between delay and error constraints and the system capacity is a challenging important research field.

2.1.1.2 System capacity vs. fairness between users

As discussed in Chapter 1, the contradiction between the capacity of the system and fairness between the users of the network appears due to the variation in the location-based long-term average channel qualities of the users' channels. As a consequence, transmitting to users with worse channels costs more of the physical resources. However, the network operator will not be interested in serving only the users with good channel conditions. Further discussions on the fairness problem are available in Section 2.3.

2.1.1.3 System capacity vs. system complexity

Improving the system capacity requires more complex system structures. Thus, it becomes sensible to accept some reduction in the achievable capacity by using less complex systems. For example, to achieve capacity limits of the downlink in SISO systems adaptive power control and superposition coding with successive interference cancellation (SIC) at the receivers is needed [47]. Alternatively, orthogonal signalling and constant transmission power can be used instead to reduce the complexity of the system as long as this will not lead to severe degradation in the capacity limits of the system. More analysis on the trade-off between capacity and system constraints are available in Chapter 3.

2.1.1.4 System capacity vs. capacity costs

A fundamental characteristic of the communication problem is that the cost of increasing the spectral efficiency (bits/sec/Hz), in terms of transmitted power (i.e. the cost is presented in bits/Joul), becomes higher as the spectral efficiency over the link increases. This is a result of the fact that the relation between the capacity and the resources is logarithmic.

$$R = \log\left(1 + \frac{h\,P}{\sigma^2}\right) \tag{2.1}$$

where the capacity R is in (bits/sec/Hz) and the transmitted power P is in (Watts/Hz). The logarithm is to the base 2. σ^2 is the noise variance and h is the channel quality (i.e. power gain: ratio between received and transmitted power).

Figure 2.3 shows the relation between the capacity and the capacity cost in the AWGN channel with noise variance equal to 1.

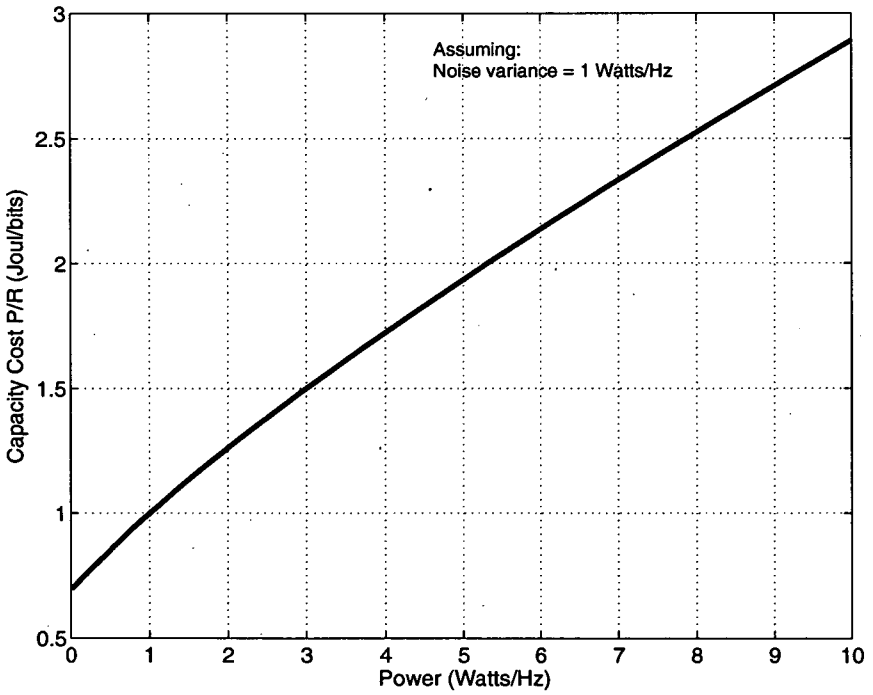


Figure 2.3: The tradeoff between capacity and capacity Cost.

It is clear that increasing the spectral efficiency too much is a waste of the energy resources

(i.e. power). On the other hand, operating at too low spectral efficiency is a waste of the scare bandwidth resources. Thus, a trade-off is needed to decide a suitable level of spectral efficiency based on the channels' conditions.

2.2 Overall System Efficiency

The difficulty in defining a measure of the overall efficiency of the system is due to the existing trade-offs. However, all of the trade-offs in the system are fundamental characteristics of the communication problem over multiuser wireless channels. Thus, the specific solutions to the contradicting requirements should not be used as a measure of how good the system is, but rather a measure of how well the system is operated (i.e. managed). In other words, we can distinguish between how the system is managed, and how good the system itself is. The efficiency of the system can be measured by its ability to achieve the capacity limit at the specific operating point chosen by the network operator. The operating point is defined by the output, input, and control parameters of the system. The possible suitable solutions to the contradicting requirements is a type of system management by selecting the operating point of the system.

For example, the unfair capacity distribution over the network area can not be solved by the real-time operation techniques (i.e. physical layer optimisation and scheduling, etc). The system design can not affect or change this characteristic of the problem which is a direct consequence of the broadcast nature of the wireless medium (unlike a wired medium in which power is directed towards the receiving user only). However, the network operator can choose a suitable solution to this problem by deciding the amount of resources allocated to each user.

In an analogy, we can judge how good an organisation is by two factors; its management and the efficiency of its workers in accomplishing the tasks assigned to them by the managers of the organisation. The management decisions are good if the resources of the organisation are utilised properly in order to maximise the profits of the organisation. Furthermore, the efficiency of a worker in the organisation can not be measured by the decisions of his managers, but rather with his skills and capabilities to complete the tasks that are assigned to him efficiently.

The overall system efficiency is hence defined by two types of efficiency measures; *system management efficiency* and *system operation efficiency.*

2.2.1 System management efficiency

There is no unique measure of the system management efficiency. Many possible metrics can be applied based on, e.g., business models, etc. However, this topic is very important and it requires extensive research in order to be able to design rational methodologies to control the operating point of the system. For example, research is needed to quantitatively characterise and describe the relations between the contradicting requirements. More advanced algorithms are needed to be applied for the network admission control task. The admission control should take into considerations the channel conditions, the requested services, the traffic load, etc. in order to decide the degree of fairness between the admitted users, and the specific QoS constraints to be maintained as well as the spectral efficiency over the wireless links.

2.2.2 System operation efficiency

This type of efficiency can be uniquely defined and measured. For any operating point of the system, there is an upper limit (capacity) of the possible achievable rates using the allocated amount of physical resources. The efficiency of the system can be judged based on how close it performs relative to the capacity limit for the specific operating point. **The evaluation of the efficiency of the scheduling schemes is part of the system operation efficiency and hence can be uniquely measured**. This is a key message of this work that despite all the trade-offs that make the definition of system management efficiency ambiguous, the efficiency of the scheduling schemes can be uniquely measured and analysed. The scheduler should be designed in order to achieve the capacity limits over all possible operation points of the system. Further discussion is in the next section and some numerical examples are provided in Chapter 4 which help to make the concepts clear.

2.3 Optimising Multiuser Scheduling

First, the two main objectives of any scheduling scheme are given, followed by a discussion of how to define the optimality of the scheduling task.

2.3.1 Objectives of multiuser scheduling schemes

In general, scheduling algorithms have two main objectives:

- Efficient allocation of the scarce physical resources: The physical resources include power, frequency bandwidth and time. Efficiency of allocating the resources can be obtained by exploiting the multiuser diversity of the network.

- Achieving suitable fairness criteria and QoS requirements of the different applications: Although always transmitting only over the best channel maximises the system sum-throughput, this strategy results in "unfair" allocation of the wireless resources among the users. The network operator will not be only interested in maximising the sum-throughput over the network, but rather in providing QoS for the served applications and in defining a suitable degree of fairness between the network users.

2.3.2 Pareto-optimal operation

A key point of this work is to show that it is actually possible to compare scheduling algorithms in terms of their efficiency in allocating the wireless resources, i.e. power and bandwidth, across the *whole* range of possible operating points. Although there is a contradiction between the maximum total throughput and fairness and QoS constraints, there is *no* contradiction between the two objectives of (i) efficient resource allocation and (ii) achieving fairness with certain QoS requirements. It must be accepted that there will be a loss in terms of total throughput when users with bad channels have to be served, but this is by no means a weakness of a particular scheduling scheme – it is rather a fundamental trade-off that is described by the theoretical limits of information theory, and the key question is how close to those limits a scheduling scheme performs. Efficient scheduling algorithms are those which operate on or close to the boundary of the capacity region, which is the set of long-term average achievable user rates for given average power constraints. In other words, schedulers should operate at a *Pareto-optimal point*. This means that no user can have a higher long-term average service rate without the need to decrease the long-term average rates for at least one other user.

As an example, Figure 2.4 shows the capacity region of a two-user case. The two-user case is selected in order to visualise the capacity region and the pareto-optimal operating points. Although it is not possible to visualise the capacity region for systems with more than 3 users, the concepts presented here are still valid in such multi-user systems. The capacity region in Figure 2.4 is the convex hull bounded by the shown capacity region boundary. Figure 2.4 shows all possible operating points (i.e. long term average service rates) given a certain amount of physical resources. The optimal operating points are shown in the figure as well.

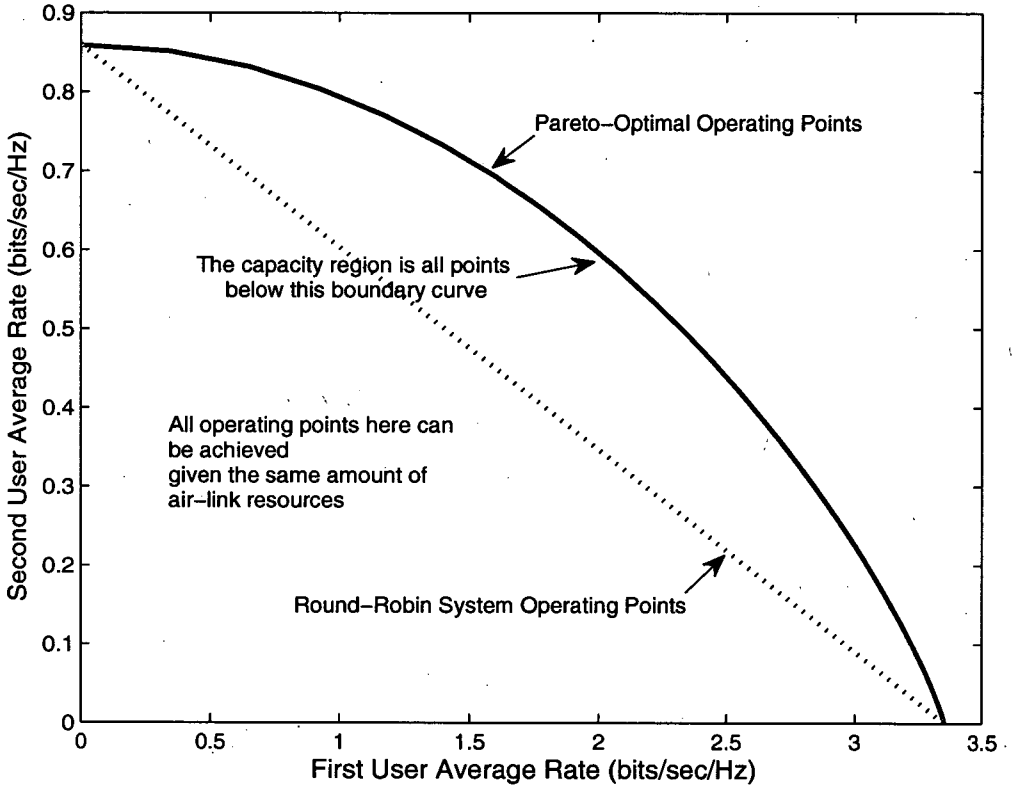


Figure 2.4: Capacity region and optimal operating points for two-user case. In this illustrative figure, the long term average channel quality of first user is assumed to be better than the average channel quality of the second user.

The operating point of the network is affected by many elements of the system such as physical layer interface optimisation including modulation, coding, etc. as well as the percentage of wireless resources given to each user in the system and the channel access scheme adopted in the system. One of the important factors affecting the operating point of the system is the scheduling scheme. For example, without exploiting the channel state information in the scheduling decision (i.e. in Round-Robin systems), the system will operate away from its Pareto-optimal points as shown in Figure 2.4. Furthermore, not all of the channel aware scheduling schemes perform at the optimal points, even though they are all utilising the same amount of wireless resources. Thus, it is very important to use a proper scheduler that results in the best performance of the system in terms of wireless resources allocation.

2.3.3 Formulation of the optimisation problem of multiuser schedulers

The most appropriate formulation of the optimisation problem to design multiuser scheduling schemes for wireless networks is to operate under Pareto-optimal conditions. This concept can be mathematically formulated as described below.

The multiuser capacity region is always convex [48] because if the system can achieve two distinct operating points by the same amount of resources, then all the operating points located on the line connecting these two operating points can be achieved by alternating (i.e. time sharing) of the schemes used to achieve these two operating points [49].

Since the capacity region is convex, then all operating points on the capacity boundary are Pareto-optimal. Furthermore, the boundary of the capacity region can be characterised as the closure of the parametrically defined surface

$$\left\{\mathbf{R}(\boldsymbol{\mu}) : \boldsymbol{\mu} \in \Re_{+}^{M}, \sum_{i} \mu_i = 1\right\} \tag{2.2}$$

where for every weighting vector $\boldsymbol{\mu}$, the rate vector $\mathbf{R}(\boldsymbol{\mu})$ can be obtained by solving the optimisation problem:

$$\max \sum_{i=1}^{M} \mu_i R_i \tag{2.3}$$

where R_i is the long-term average achievable rate of user i, and M is the total number of users in the system.

Problem (2.3) should be solved so that the resources constraints are maintained, and based on the specific systems' constraints. The selection of the weighting factors should be done to select one of the possible Pareto-optimal operating points of the system.

In Chapter 3, the solutions of (2.3) for different systems' constraints are given based on multiuser information theory.

2.4 Conclusions

Many trade-offs exist in the multiuser communication over wireless fading channels between the system capacity (throughput), and the quality-of-service of the served applications, the fairness between the users, the cost of the capacity in terms of physical resources and the system complexity. For every operating point of the network related to one of the possible solutions to the existing trade-offs in the system, an upper limit (i.e. capacity) of the information rates that can be achieved over the wireless links exists. The solution to achieve the capacity limits of any possible operating point involves the application of THE optimal scheduling policy. Thus, the best definition for the optimality of a scheduling scheme is not to maximise the total throughput of the system (which is only one of the possible operating points), but rather to operate at the capacity limits. This is equivalent to operating at Pareto-optimal conditions in which no user can have higher rate without decreasing the rates of other users or increasing the amount of physical resources. Maximising any utility function or maintaining fairness criteria should be done such that a suitable operating point is chosen on the capacity region's boundary. This can be done by adjusting the control parameters of the optimal scheduling policies achieving the capacity limits.

The optimisation problem (to find the optimal resource allocation schemes in order to operate at the capacity region's boundary) is to maximise a weighted sum of the long-term average rates of the users under the main constraint on the long-term average transmitted power as well as possible additional constraints based on the systems' capabilities. There is no contradiction between efficient resource allocation and achieving fairness and QoS requirements, taking into consideration that the efficiency of a scheduler is by operating at Pareto-optimal conditions.

In the following chapter, the optimisation problem to achieve the multiuser capacity limits is solved in order to obtain the optimal scheduling policies under different constraints of the system. Furthermore, comparisons of the performance of the system under different system

constraints are provided in numerical examples. The drawbacks of not applying the optimal scheduling policies are analysed in Chapter 4 using a comparison-study of well-known scheduling policies that are suggested in the literature. The comparison study demonstrates the importance of applying the optimal scheduling policies derived in Chapter 3. Adjusting the optimal scheduling policies to achieve many possible operating points flexibly is discussed in Chapter 5. A scheduler is suggested to achieve resource-sharing fairness constraints, which is one of the possible ways to select the operating points of the system.

Chapter 3
Information-Theoretic Optimal Multi-user Scheduling and Resource Allocation

In this chapter, multiuser information theory is applied to solve the optimisation problem (2.3) for different constraints on the physical layer schemes. The solutions involve (i) characterising the boundary of the capacity region and (ii) describing the resource allocation schemes to achieve any point on the capacity region's boundary. Some of the results are already known from the literature. However, there are many new contributions and novelties presented in this thesis.

In order to make this chapter easy to follow, only sketches of the proofs of the derived equations are provided in the main body of the chapter, and the more detailed proofs are presented in Appendix A.

3.1 Literature Review

The problem of efficient resource allocation for centralised single-cell wireless networks has been investigated in information theory. The *block-fading channel* (e.g., [50]) is used in information theory to model time and frequency selective fading channels. The fading channels are divided into a family of parallel Gaussian flat-faded channels, each corresponds to a fading state. These flat-faded channels are called blocks. A channel block (refer to Figure 1.1) could last for several time slots as long as the channel quality is almost constant (dependent on fading statistics). In general, the capacity of block-fading multiuser channels with channel-state-information (CSI) at both the transmitter(s) and the receiver(s) can be achieved by (i) optimal power allocation over the channel blocks and (ii) optimal resource (rate and power) allocation over the users in each of the channel blocks. OFDMA systems with Dynamic Subcarrier Allocation (DSA), Adaptive Modulation and Coding (AMC) and optimal scheduling

over the subcarriers allows for the realisation of information-theoretic solutions for efficient resource allocation over wireless fading channels.

Different notions of "capacity" have been defined in the literature. The *ergodic (Shannon) capacity region* was investigated in [49], [34]. In this case, the problem is formulated as to maximise a weighted sum of the long-term average rates with the only constraint on the long-term average power. Thus, it is applicable when the served applications in the network are delay-tolerant. The *delay-limited capacity region* and the *outage capacity* can be applied to analyse the network operation with delay-sensitive applications ([51], [52], [53], [54]) as additional constraints regarding the delay or outage probability are added to the problem.

The *"broadcast channel"* (BC) is the term used in information theory to define the downlink (one-to-many) case [55]. The characterisation of the ergodic capacity region of BC channels [47], [49] is formulated for a given long-term average power constraint of the transmitter. The term *"multi-access channel"* (MAC) is used for the uplink (many-to-one) case [55]. In [18], [34] the characterisation of the ergodic capacity region of MAC channels is provided. The problem is formulated by assuming individual power constraints for each transmitter. In practical scheduling, determining the individual average power of the transmitters is a degree of freedom for the scheduler. Thus, the problem can also be formulated by replacing the individual constraints by one sum-power constraint covering all transmitters [56]. Further details are provided in Section 3.3. For the case of a single power constraint on the total power transmitted over the network, the duality of the MAC and BC channels is discussed in [57]. BC and MAC channels are dual if they have the same channel vector $\boldsymbol{h} \doteq \{h_1, h_2, ...\}$ (i.e. h_i of receiver i in the BC equals h_i of transmitter i in the MAC). In [57], it is shown that the capacity region of the MAC channels with sum-power constraint is identical to the capacity region of the dual BC channels, and there exists a striking similarity between the optimal resource allocation schemes for the two cases. Furthermore, the optimal power and rate allocation policies for the MAC under a sum-power constraint and BC channels are exactly identical when orthogonal multiple-access schemes (like in OFDMA) are used. This is why the scheduling policies under orthogonal signalling constraints are applicable to both the uplink and the downlink.

The optimal power allocation scheme over (block-)fading Gaussian broadcast and multi-access channels is given by the water-filling approach: more power is allocated when the channel is better and, depending on the desired operating point on the capacity region's boundary surface, some users are assigned higher average power to meet their rate demands.

The optimal resource allocation over a (flat-faded) channel block involves applying the optimal channel-access scheme, which is code division multiple access (in MAC) [34] or superposition coding (in BC) [49] with successive interference cancellation (SIC) at the receivers. Furthermore, the number of users scheduled in a channel block varies depending on the channel conditions.

The optimal solutions are in most cases difficult if not impractical to implement. Thus, sub-optimal solutions with close-to-optimum performance that lend themselves to an easy implementation are more favourable from a practical perspective. That's why, one of the objectives of this work is to study the optimal resource allocation schemes under practical constraints on the physical layer schemes adopted by the system, and to characterise the capacity limits in these cases in order to compare the system performance when these constraints are applied and to analyse the loss in terms of user rates with respect to the optimal cases.

3.2 Channel Model

The wireless fading channels (both time and frequency selective) are modelled as a family of parallel "constant" Gaussian channels. Each of the parallel Gaussian flat-faded channel blocks corresponds to a fading state.

The M-user Gaussian block-fading broadcast channel (BC) consists of a single transmitter and M receivers. In channel block k, the transmitter broadcasts a signal $x[k]$, and the received signals are

$$y_i[k] = \sqrt{h_i[k]}x[k] + n_i[k], \quad i = 1, \cdots, M \tag{3.1}$$

where $h_i[k] > 0$ is the constant channel quality (i.e. power gain) between the transmitter and the i-th receiver at channel-block k. In this work, it is assumed, without loss of generality, that the channel gain h is real and represents the power gain of the link. h does not have an imaginary part since perfect phase information at the receivers is assumed. $n_i[k]$ is Gaussian noise with zero mean of that receiver. The noises $n_i[k]$ are statistically independent, and are assumed to have a common variance σ^2. Although in practice the noise variance may differ between the users' terminals, we can still assume a common noise variance. This assumption helps to simplify the derived equations in this work. The variations in noise levels can be treated by scaling the channel quality vector $\mathbf{h}$ according to noise levels since the achievable rates over a channel depends on Signal-to-Noise ratio (SNR) rather than the absolute values of the signal

power and the noise power. For example if the noise variance of user i equals $\sigma_i^2 = \alpha\sigma^2$, then the channel quality of user i should be scaled as $h_{i\_scaled} = \frac{h_i}{\alpha}$.

The M-user Gaussian block-fading multi-access channel (MAC) consists of a single receiver and M transmitters. At channel block k, each transmitter i transmits a signal $x_i[k]$, and the receiver receives the composite signal

$$y[k] = \sum_{i=1}^{M} \sqrt{h_i[k]} x_i[k] + n[k] \tag{3.2}$$

where $h_i[k] > 0$ is the constant channel quality between the i-th transmitter and the receiver at channel-block k.

The fading processes of all users are independent of each other, are stationary and have continuous probability density functions, $f_{H_i}(h)\ \forall i$. The cumulative density functions of the fading processes are denoted by $F_{H_i}(x) \doteq \int_0^x f_{H_i}(h')dh'$. In the numerical examples provided in the thesis, the magnitudes $a_i = \sqrt{h_i}$ of the users' channel coefficients are assumed to have Rayleigh or Rice distributions [13, pp. 45–48], [15, pp. 78–79]. As the channel power gain, h_i, is the square of the channel-coefficient's magnitude, we have to use the variable-substitution $h_i = a_i^2$ in the original Rayleigh/Rice PDFs. As $q(a_i) = a_i^2$ is monotonically increasing for $a_i > 0$, the standard rule $f_{H_i}(h_i) = \frac{f_{A_i}(a_i)}{q'(a_i)}\Big|_{a_i = q^{-1}(h_i)}$ with $q'(a_i) = 2a_i$ and $q^{-1}(h_i) = \sqrt{h_i}$ is used to obtain the PDF of the channel power gain h_i from the PDF of the channel coefficient's magnitude a_i. The following results are obtained for the Rayleigh-case:

$$\text{PDF}: \quad f_{H_i}(h_i) = \frac{1}{\bar{h}_i} \exp\left(-\frac{h_i}{\bar{h}_i}\right) \tag{3.3}$$

$$\text{CDF}: \quad F_{H_i}(h_i) = 1 - \exp\left(-\frac{h_i}{\bar{h}_i}\right) \tag{3.4}$$

with $\bar{h}_i$ the long-term average channel power gain.

For the Rice-case:

$$\text{PDF}: \quad f_{H_i}(h_i) = \frac{\kappa+1}{\bar{h}_i} \exp\left(-\kappa - \frac{\kappa+1}{\bar{h}_i} h_i\right) I_0\left(2\sqrt{\frac{(\kappa+1)}{\bar{h}_i}\kappa h_i}\right) \tag{3.5}$$

$$\text{CDF}: \quad F_{H_i}(h_i) = \int_0^{\frac{\kappa+1}{\bar{h}_i}h_i} e^{-(\kappa+x)}\, I_0\left(2\sqrt{\kappa\, x}\right) dx \tag{3.6}$$

with $I_0(.)$ the zero-th order modified Bessel function[1] of the first kind, and κ is the fading parameter that is defined as the ratio $\kappa \doteq h_{i,\text{LOS}}/h_{i,\text{NLOS}}$ of the average power gains h_{LOS} on the line-of-sight path and h_{NLOS} on the non-line-of-sight path: $\kappa = 0$ means we get Rayleigh fading, and $\kappa \to \infty$ means "no fading". The long-term average channel power gain is again denoted by $\bar{h}_i$. Note that $h_i \geq 0$ in (3.3), (3.4), (3.5) and (3.6).

The notations $P_i[k]$ and $R_i[k]$ are used to indicate the power[2] and the rate in (bits/sec/Hz) respectively that are allocated to user i in channel block k. The long-term average rate that is allocated to user i is denoted as R_i. The long-term average sum-power constraint is denoted as $\bar{P}$. In the numerical examples presented in this thesis, the rates R_i are presented in (bits/sec/Hz) because the throughput in (bits/sec) can not be used alone – without the knowledge of the bandwidth in Hz – to measure the efficiency of the system.

3.3 Problem Formulation

The ergodic capacity limits and the optimal solutions to achieve these limits under practically relevant restrictions are considered for both the downlink (BC) and the uplink (MAC) cases. The ergodic capacity region is defined as the set of all rate vectors $\mathbf{R}$ such that the long-term average power constraint $\bar{P}$ over all channel blocks is not exceeded. The optimum points within the capacity region are those that are located on the boundary surface. Although the ergodic capacity is relevant for the case for delay-tolerant applications over the network, modifications of the optimal scheduling polices to provide some degrees of restrictions over the transmission delay are discussed in Chapter 5.

The boundary of the capacity region can be characterised by (2.2) and (2.3). The way to scan different operating points on the capacity-region boundary is by adjusting the vector $\boldsymbol{\mu} \doteq \{\mu_1, \mu_2, ...\}$ of weighting factors defined in the problem of optimal resource allocation [47]. The ratios[3] of the weighting factors – not their absolute magnitudes – affect the location

[1]This function can be represented [13, p. 44] by the infinite series $I_0(x) = \sum_{k=0}^{\infty} \frac{1}{(2^k\, k!)^2} x^{2k}$, $x \geq 0$. As, for k sufficiently large, the denominator will dominate the result, a limited number of summands will suffice to get accurate results. This means that (3.5) can be evaluated without explicit use of any Bessel function and that (3.6) can be evaluated without any numerical integration.

[2]When the notation P is used, it means the transmit power P_T. The received power is indicated as P_R.

[3]Any constant used to scale *all* μ_i would have no effect with respect to the scheduling decisions, and in water-filling power allocation (such as in (3.16)) the normalisation by the Lagrangian multiplier λ ("power-price parameter") will compensate for any scaling factors in the ratio μ_i/λ. Hence, only the relative contributions of the μ-values determine the operating point on the boundary surface of the capacity region.

of the operating point. That is why the constraint ($\sum_i \mu_i = 1$) is added to (2.2).

For the block fading channel model, problem (2.3) can be written as:

$$\max \frac{1}{K} \sum_{k=1}^{K} \sum_{i=1}^{M} \mu_i R_i[k] \tag{3.7}$$

where k is the index of a channel block, K is the total number of channel blocks, and M is the number of active users. We can assume without loss of generality that all channel blocks have identical frequency bandwidth and time duration.

The problem should be solved under the main constraint on the transmitted power. In the analysis of the *"multi-access channels"* (MAC), a single long-term average sum-transmit-power constraint is used instead of individual power constraints for the users, which were assumed in the original work in literature [34]. This case is also relevant in practice [56]. Furthermore, it gives a more general solution with extra information (can not be obtained from the original work) about the optimal average powers to be allocated to each user to achieve a certain operating point. In [57] the duality of the MAC and BC channels was discussed. It was shown that the capacity region of the MAC channels with sum-power constraint is identical to the capacity region of the dual BC channels. Thus, in all the cases under consideration, the equations characterising the boundary surface of the capacity region are applicable to both the BC and the dual MAC channels. Furthermore, there exists a striking similarity between the optimal resource allocation for both channels.

Thus, problem (3.7) is solved with the power constraint always maintained, in both cases of broadcast and multi-access channels:

$$\frac{1}{K} \sum_{k=1}^{K} \sum_{i=1}^{M} P_i[k] = \bar{P} \tag{3.8}$$

In addition to the primary power constraint, the objectives in this chapter is to solve the problem under more auxiliary constraints. The solution of (3.7) should involve two objectives: (i) to give closed-form expressions that characterise the capacity limits and (ii) to describe the resource allocation schemes (for BC and MAC) to achieve these limits for each of the cases under considerations. The advantage of driving equations[4] to characterise the capacity region limits

[4]All the equations to characterise the capacity region's boundary that are presented in the thesis are novel contributions.

is to be able to compare the different cases and thus to analyse the trade-off between system capacity and system complexity.

The following cases are considered in this chapter:

- **OPT**: Optimal case (i.e. without any auxiliary constraint):

  New analytical results that complement the original work in the literature ([49] for the BC case and [34] for the MAC case are presented in this thesis.

- **OS**: Orthogonal signalling constraint:

  With the orthogonal signalling constraint, the channel can be "shared" by more than one user in the sense that within a single channel block more than one user can be scheduled but with the application of an orthogonal signalling scheme such as FDMA or TDMA. In other words, superposition coding with successive interference cancellation at the receiver is excluded by this constraint. The assumption in a block fading model is that a channel block comprises many time slots or frequency subcarriers, and thus it is possible to let more than one user share a given channel block.

  The number of users sharing a channel block and the access ratios of each scheduled user is subject to optimisation. Thus, with the orthogonal signalling constraint, the scheduler should decide at each channel block k, (i) the channel sharing ratio of each user ($\tau_i[k]$) and (ii) the power of each user ($P_i[k]$). The following equations describe the orthogonal signalling auxiliary constraint:

$$\sum_i \tau_i[k] = 1 \tag{3.9}$$

$$R_i^{\mathrm{OS}}[k] = \tau_i[k] \ \log\left(1 + \frac{h_i[k] \ P_i[k]}{\sigma^2 \ \tau_i[k]}\right) \tag{3.10}$$

  In the TDMA case (3.10) is valid taking into consideration that $Pi[k]$ is the average power of user i in block k. So, if the transmitter has a constant power P_T, then user i (who transmits τ_i of the time) will have an average power of $P_i[k] = \tau_i.P_T$. The analytical characterisation of the capacity region's boundary under the OS constraint presented in this thesis is novel.

- **CP**: Constant transmitted power per channel block:

  "Constant total power per block" is to be interpreted such that in each and every channel block the sum transmitted power for all users is constant. For a broadcast channel this

means that the total power used by the base station for all users is the same in every channel block although the number of users scheduled in every block is variable and subject to optimisation. In the multiple-access case, again the sum of all powers of all users' transmitters is assumed to be constant. The auxiliary constraint in this case is:

$$P[k] = \sum_{i=1}^{M} P_i[k] = \bar{P} \tag{3.11}$$

To the best of my knowledge, all the analytical results for this case that are presented in this thesis are new.

- **SU**: Single user selection per channel block:

  This constraint means that in a single channel block, no more than one user can be scheduled to access the channel. The rate of the scheduled user is:

$$R_i^{\mathrm{SU}}[k] = \log\left(1 + \frac{h_i[k]\ P_i[k]}{\sigma^2}\right) \tag{3.12}$$

  $\mathbf{R}[k]$ has a maximum number of one non-zero element.

- **CP-OS**: Constant sum power *and* orthogonal signalling in every channel block:

  Both constraints are applied in this case. The literature review on this case is presented in this thesis.

- **CP-SU**: Constant sum power *and* single-user selection per channel block:

  Both constraints are applied in this case. New analytical results are presented in this thesis to characterise the boundary of the capacity region.

These cases are compared by numerical results. To visualise the capacity limits, the two-user case is considered, with the assumption of different long-term average channel qualities of the users. Qualitatively, the results carry over to the M-user case. Analysis for a higher number of users is provided as well by selecting a specific operating point (max. sum-throughput) with the assumption of symmetric channels.

In the following sections, the solutions for each of the cases under consideration is given. However, the detailed proofs are not provided. In Appendix A.1, a summary of the main results of paper [47] is given as well as new further discussions which serve as proofs for the results in this chapter, especially for the optimal case and the constant power per block case. Although

some of the results are based on [47], the results are presented here in a different way than the context of the mentioned paper. This includes providing some proofs that were not included in the original work, or deriving close-form equations characterising the capacity region based on the general forms provided in the original paper.

3.4 Solution of Optimal Case (No Auxiliary Constraints)

3.4.1 Optimal resource allocation

As discussed in [47], problem (3.7) can be solved by first applying the Lagrangian characterisation [48] in order to define the problem in an unconstrained format. The resulting optimisation problem is:

$$\max_{\{P[k]\}} \sum_{k=1}^{K} \left(\sum_{i=1}^{M} \mu_i R_i[k] - \lambda \sum_{i=1}^{M} P_i[k] \right) \tag{3.13}$$

This is equivalent to

$$\sum_{k=1}^{K} \max_{P[k]} \left(\sum_{i=1}^{M} \mu_i R_i[k] - \lambda P[k] \right) \tag{3.14}$$

where λ is selected such that

$$\frac{1}{K} \sum_{k=1}^{K} P[k] = \bar{P} \tag{3.15}$$

Thus, the main optimisation problem is decomposed into (i) a family of optimisation problems, one for each channel block, and (ii) an equation to control the power price λ in order to maintain the long-term average power constraint. Following the procedure described in [47] by defining marginal utility functions, and by extending these results to the MAC case, a summary of the solution is provided below:

Power allocation over the channel blocks:

The total power transmitted in a block k is identical in BC [47] and MAC channels:

$$P_{sum}[k] = \max_{i} \left[\sigma^2 \left(\frac{\mu_i}{\lambda} - \frac{1}{h_i[k]} \right)^+ \right] \tag{3.16}$$

The notation $x^+ = \max(x, 0)$ is used.

Resource allocation in each channel block:

The optimal resource allocation over a (flat-faded) channel block involves applying the optimal channel-access scheme, which is code division multiple access (in MAC) or superposition coding (in BC) with successive interference cancellation (SIC) at the receivers. The SIC at the receivers of BC channels is in order of decreasing μ. Each receiver decodes the signals sent to users of higher μ before decoding its own signal. However, in MAC channels, the receiver performs SIC in order of increasing μ [57]. A summary of the greedy algorithm procedure to compute the power allocated to each user in channel block k is provided for both BC [47] and MAC – novel contribution by extending results of BC [47] (equation (3.17)) to MAC with sum power constraint (equation (3.18)) – channels:

Marginal utility functions ("rate revenue minus power cost") are defined for each channel block k:

$$\text{BC: } u_i(z) \equiv \frac{\mu_i}{\frac{1}{h_i[k]} + z} - \lambda, \; z \geq 0 \tag{3.17}$$

$$\text{MAC: } u_i(z) \equiv \frac{\mu_i}{1+z} - \frac{\lambda}{h_i[k]}, \; z \geq 0 \tag{3.18}$$

Then based on the marginal utilities which are dependent on the channel quality vector $\mathbf{h}[k]$, the periods $\mathcal{A}_i$ are obtained:

$$\mathcal{A}_i \equiv \{z \in [0, \infty) : u_i(z) > u_j(z) \;\; \forall j \neq i \text{ and } u_i(z) > 0\}$$

Since $u_i(z)$ is monotonically decreasing and $u_i(z)$, $u_j(z)$ $(i \neq j)$ cross each other at maximum once, the period $\mathcal{A}_i$ is continuous. The power allocation is calculated as:

$$\text{BC: } P_i[k] = \sigma^2 \int_{\mathcal{A}_i} dz \tag{3.19}$$

$$\text{MAC: } P_i[k] = \frac{\sigma^2}{h_i[k]} \int_{\mathcal{A}_i} dz \tag{3.20}$$

3.4.2 Characterisation of the boundary of the capacity region

To derive equations to characterise the boundary surface of the capacity region, extensions to the work in [47] were done to get the following equations to compute $\mathbf{R}(\boldsymbol{\mu})$. With the assumption that the fading processes of all users are stationary with continuous probability density functions

and independent of each other: for each user $i = 1, ..., M$

$$R_i^{\text{OPT}} = \frac{1}{\ln 2} \int_0^{\infty} \frac{1}{1+z} \int_{\frac{\lambda(1+z)}{\mu_i}}^{\infty} f_{H_i}(x) \prod_{j \neq i} F_{H_j}(\alpha^*)\, dx dz \tag{3.21}$$

or equivalently

$$R_i^{\text{OPT}} = \frac{1}{\ln 2} \int_0^{\frac{\mu_i}{\lambda}} \int_{\frac{1}{\frac{\mu_i}{\lambda} - z}}^{\infty} \frac{1}{\frac{1}{x} + z} f_{H_i}(x) \prod_{j \neq i} F_{H_j}(\beta^*)\, dx dz \tag{3.22}$$

where λ in (3.21), (3.22) is computed based on (3.15) which, in our case of independent fading processes, is equivalent to:

$$\sum_i \int_{\frac{\lambda}{\mu_i}}^{\infty} f_{H_i}(x) \prod_{j \neq i} F_{H_j}(\zeta^*) \left[\frac{\mu_i}{\lambda} - \frac{1}{x} \right] dx = \frac{\bar{P}}{\sigma^2} \tag{3.23}$$

There are two other equivalent forms to compute λ:

$$\sum_i \int_0^{\infty} \int_{\frac{\lambda(1+z)}{\mu_i}}^{\infty} \frac{1}{x} f_{H_i}(x) \prod_{j \neq i} F_{H_j}(\alpha^*)\, dx dz = \frac{\bar{P}}{\sigma^2} \tag{3.24}$$

$$\sum_i \int_0^{\frac{\mu_i}{\lambda}} \int_{\frac{1}{\frac{\mu_i}{\lambda} - z}}^{\infty} f_{H_i}(x) \prod_{j \neq i} F_{H_j}(\beta^*)\, dx dz = \frac{\bar{P}}{\sigma^2} \tag{3.25}$$

α, β, ζ are given by:

$$\alpha = \frac{\lambda}{\frac{\lambda}{x} + \frac{\mu_j - \mu_i}{1+z}} \tag{3.26}$$

$$\beta = \frac{\mu_i x}{\mu_j + zx(\mu_j - \mu_i)} \tag{3.27}$$

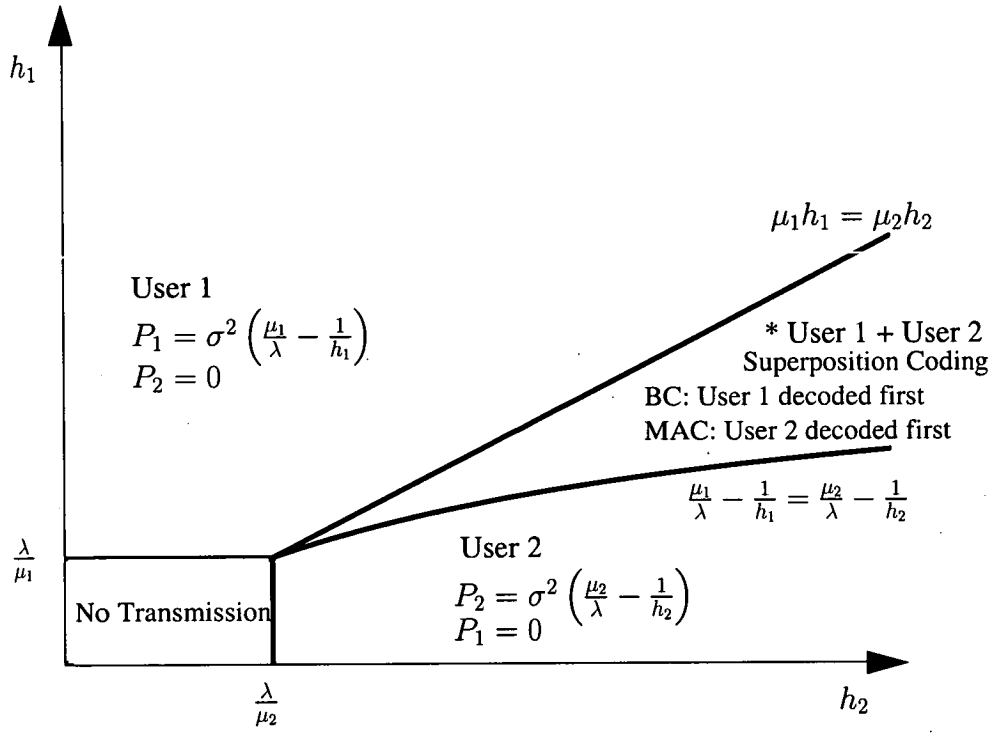
$$\zeta = \frac{1}{\frac{1}{x} + \frac{\mu_j - \mu_i}{\lambda}} \tag{3.28}$$

The notation $[x]^*$ in (3.21), (3.22), (3.23) is defined as:

$$[x]^* \doteq \begin{cases} x & \text{if} \quad x \geq 0 \\ +\infty & \text{if} \quad x < 0 \end{cases} \tag{3.29}$$

3.4.3 Example of two-user Case

Figure 3.1 gives a new kind of illustration of this optimal resource allocation scheme for the two-user case. The figure shows the dependence of the resource-allocation on the channel gains of the users. We can observe that the region of (h_1, h_2)-pairs for which superposition coding is performed is relatively narrow, so in most of the cases at most one user is scheduled by the optimal scheme; therefore, a suboptimal scheme, which always schedules at most one user, may well perform close to the optimum.



$$* \; P_{\text{sum}} = \sigma^2 \left(\frac{\mu_1}{\lambda} - \frac{1}{h_1}\right), \quad P_2 = P_{\text{sum}} - P_1$$

$$\text{BC: } P_1 = \sigma^2 \left(\frac{\mu_1}{\lambda} - \frac{\mu_1}{\mu_1 - \mu_2}\left(\frac{1}{h_1} - \frac{1}{h_2}\right)\right)$$

$$\text{MAC: } P_1 = \sigma^2 \left(\frac{\mu_1 - \mu_2}{\lambda\left(1 - \frac{h_1}{h_2}\right)} - \frac{1}{h_1}\right)$$

Figure 3.1: Optimal resource allocation for the two-user case with $\mu_1 > \mu_2$

An example of applying the greedy algorithm for the optimal resource allocation over BC channels is shown in Figure 3.2 for a two-user case. The marginal utility functions (3.17) are plotted, and the power allocated to each user is based on the periods (i.e. of the variable z) over

which his marginal utility function is positive and larger than the other user's marginal utility function. The illustrative example in Figure 3.2 corresponds to a channel vector (h_1, h_2) for which superposition coding is required. From Figure 3.1, we know that this condition appears when $\mu_1 h_1 < \mu_2 h_2$, and $\frac{\mu_1}{\lambda} - \frac{1}{h_1} > \frac{\mu_2}{\lambda} - \frac{1}{h_2}$, given that $\mu_1 > \mu_2$ is assumed in this specific example.

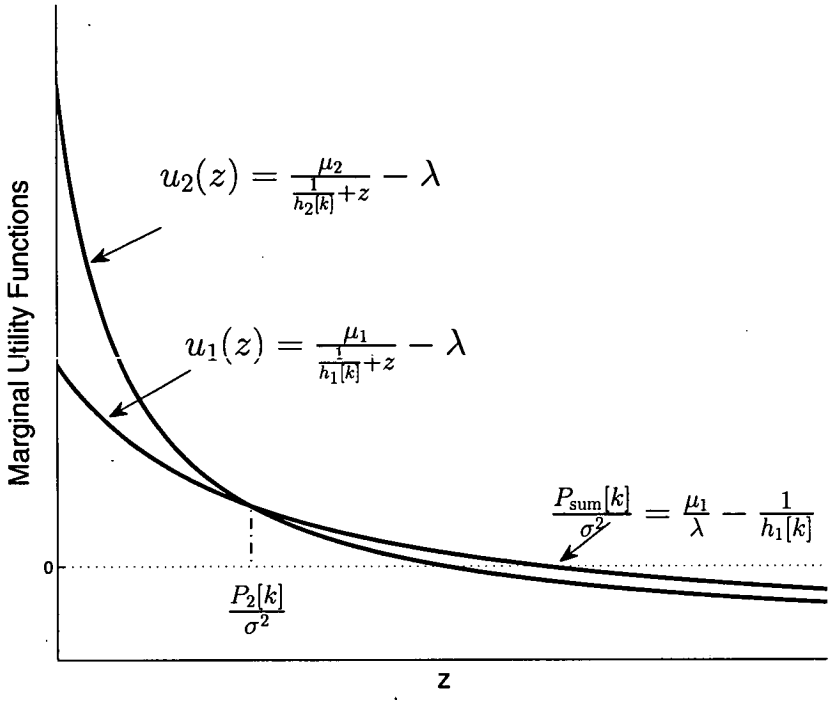


Figure 3.2: The greedy algorithm for the optimal resource allocation for BC channels. In this specific example $\mu_1 > \mu_2$, $\mu_1 h_1 < \mu_2 h_2$, and $\mu_1/\lambda - 1/h_1 > \mu_2/\lambda - 1/h_2$

3.4.4 Complexity of the system

From a practical communications engineering perspective, the optimal solutions are in most cases difficult if not impractical to implement. Thus, sub-optimal solutions which have close-to-optimum performance and, at the same time, lend themselves to an easy implementation are favourable. This "suboptimal option" may be very attractive, as the optimal resource allocation scheme has major *dis*advantages from a practical perspective:

- Superposition coding with Successive Interference Cancellation (SIC) at the receivers can hardly be implemented in practice, because of (i) the complexity involved, (ii) the fact that all receivers would need to know the channel coefficients of other users, (iii) the necessity to inform all users about the order in which successive cancellation has to be conducted including the coding schemes used (signalling overhead), and (iv) different blocksizes used for encoding of different users. The issue with (iv) is that the cancellation of a user's signal is only possible when the whole codeword[5] for this user has been received, although the user to be detected – due to delay constraints – may well have a much shorter (although still long) channel coding blocksize. As this user would have to wait with decoding until the "interfering" user's much longer codeword has been received, even very relaxed delay-constraints are likely to be violated.

- The optimal power allocation (i.e. the water-filling approach) is adaptive based on $\boldsymbol{h}$. As a consequence of this power allocation policy, the total and individual transmission powers will vary hugely. Figure 3.3 shows the statistical distribution of the sum transmit power to achieve an arbitrary selected operating point (this point is indicated by "*" in the capacity region shown in Figure 3.4). It is possible that the transmitted power can, in some cases, be more than twice the average power constraint. This will cause problems when, e.g., the transmitter (i.e. the base station in the broadcast case) has maximum power constraints in order not to cause too much interference in adjacent cells. Furthermore, adaptive power control requires additional computational complexity to maintain the average power constraint, and variable transmission power is also likely to require more expensive radio-frequency circuitry.

Due to the apparent practical disadvantages, the problem is solved again in the following sections with constraints forcing the use of orthogonal signalling (due to the disadvantages of SC with SIC), constant power (due to disadvantages of water-filling approach) or both.

[5] In practice, although not required by information theory of block-fading channels, coding for a user will be spread over as many blocks as possible to obtain long code words that will allow for efficient channel coding.

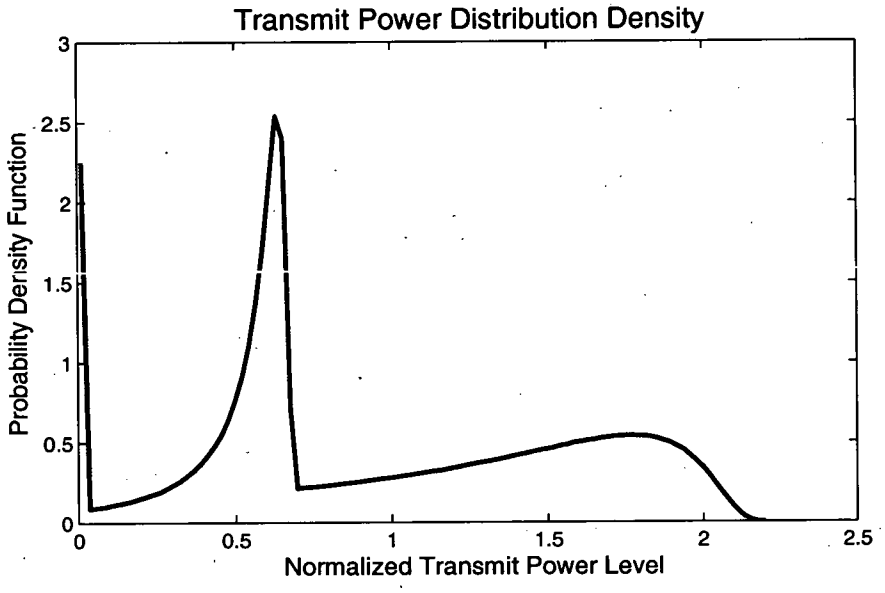


Figure 3.3: Statistical distribution of the transmit power to achieve the point indicated by "*" in the two-user capacity region shown in Figure 3.4. The peak at zero-power is a result of the case when both channel power gains are lower than their thresholds. The peak around a normalised value of 0.6 is related to cases when user 1 is transmitting (receiving). The distribution above this peak is related to situations when user 2 or both users (superposition coding) are scheduled. This difference in allocated power levels is because $\mu_2 > \mu_1$.

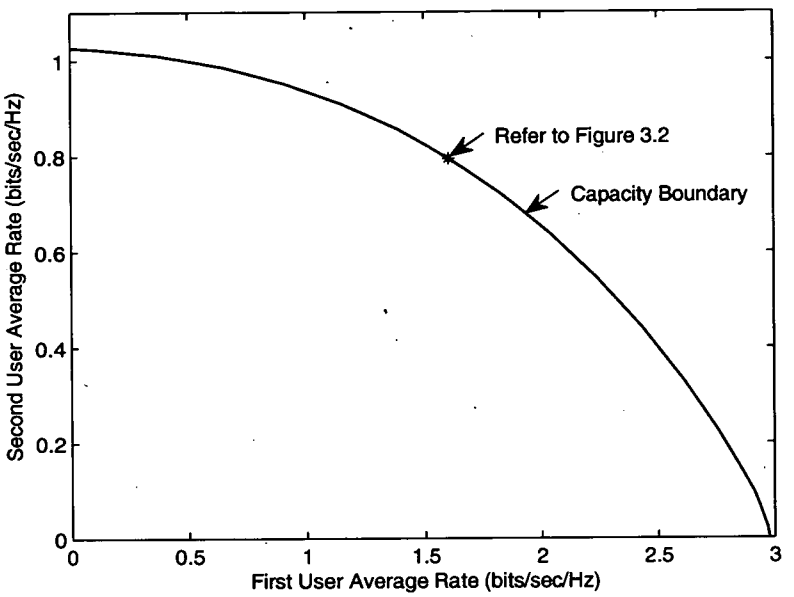


Figure 3.4: The capacity region of two-user case. The channels are assumed to be Rayleigh faded with 10 dB difference of the average power gains. The point indicated by "*" is related to the explanation of transmit power statistics in Figure 3.3.

3.5 Solution with Orthogonal Signalling Constraint

3.5.1 Resource allocation

In this case, after applying a Lagrangian characterisation to the main optimisation problem (3.7) (details are explained in [47]) the resulting equivalent problem is to find

$$\max_{\{\tau[k],\mathbf{P}[k]\}} \sum_i \left(\mu_i \tau_i[k] \log \left(1 + \frac{h_i[k] P_i[k]}{\sigma^2 \tau_i[k]} \right) - \lambda P_i[k] \right) \tag{3.30}$$

over each channel block (index k). The value of λ is controlled to maintain the primary power control constraint (3.8). The auxiliary constraint (3.9) should be maintained in the solution.

The solution of the optimisation problem (3.30) involves deciding the access ratio $\tau_i[k]$ and power $P_i[k]$ allocated to each user in every channel block k. The solution can be obtained based on the work in [49]. In that paper it is shown that if orthogonal signalling is used, then, the solution of the optimisation problem involves the selection of at most one user for every channel block k. In other words, in a channel block, one user will have $\tau_i[k] = 1$, while all other users will have $\tau_i[k] = 0$. Thus, *the orthogonal signalling (OS) constraint is actually*

equivalent to a single-user selection per channel block (SU) constraint.

Thus, although not explicitly stated in [49], we can use the fact that the OS constraint is equivalent to the SU constraint to solve our optimisation problem. The solution of (3.30) over each channel block becomes a user selection strategy, where the user to be scheduled is the one who maximises the selection argument (policy). The only user m scheduled to transmit (MAC) or receive (BC) in block k, and the power allocated to this user are calculated according to[6]:

$$m = \arg\max_i \left(\mu_i R_i[k] - \frac{\lambda P_i[k]}{\sigma^2} \right) \tag{3.31}$$

where $P_i[k]$ is calculated according to:

$$P_i[k] = \sigma^2 \left[\frac{\mu_i}{\lambda} - \frac{1}{h_i[k]} \right]^+ \tag{3.32}$$

Since the rate as a function of channel quality and power is presented by (3.12) in the case of single user selection per channel block, the selection strategy (3.31) can be written as a function of channel quality as well as μ_i and λ by substituting (3.12) and (3.32) in (3.31):

$$m = \arg\max_i \left(\mu_i \log \left(\frac{\mu_i \, h_i[k]}{\lambda} \right) - \mu_i + \frac{\lambda}{h_i[k]} \right) \tag{3.33}$$

3.5.2 Characterisation of the boundary of the capacity region

The boundary surface is characterised by the equation:

$$R_i^{\mathrm{SU}} = \int_{\frac{\lambda}{\mu_i}}^{\infty} f_{H_i}(x) \prod_{j \neq i} F_{H_j}(\gamma) \log \left(\frac{\mu_i x}{\lambda} \right) dx \tag{3.34}$$

where λ in (3.34) is computed according to

$$\sum_i \int_{\frac{\lambda}{\mu_i}}^{\infty} f_{H_i}(x) \prod_{j \neq i} F_{H_j}(\gamma) \left[\frac{\mu_i}{\lambda} - \frac{1}{x} \right] dx = \frac{\bar{P}}{\sigma^2} \tag{3.35}$$

[6]Similar results in an OFDM framework were obtained in [58] and [59].

γ in (3.34), (3.35) is given by:

$$\gamma = \frac{-\lambda}{\mu_j W\left[-\left(\frac{\lambda}{\mu_i x}\right)^{\frac{\mu_i}{\mu_j}} \exp\left(\frac{\mu_i}{\mu_j} - \frac{\lambda}{\mu_j x} - 1\right)\right]} \tag{3.36}$$

with $W()$ the Lambert function [60] (inverse of $f(x) = xe^x$).

The proof is provided in Section 4.2.

3.6 Solution with Constant Power Constraint

3.6.1 Resource allocation

In this case, the main optimisation problem (3.7) becomes equivalent to optimising over each channel block k:

$$\max \sum_{i=1}^{M} \mu_i R_i[k] \text{ subject to } \sum_{i=1}^{M} P_i[k] = \bar{P} \tag{3.37}$$

Thus, the problem is to allocate the resources over the users in each channel block. The power allocated to each user in each channel block can be obtained using the same greedy procedure of the optimal case, but with the replacement of the global power price λ in (3.17), (3.18) by block-dependent power price $\lambda[k]$, which is obtained as:

$$\lambda[k] = \max_i \left(\frac{\mu_i}{\frac{1}{h_i[k]} + \frac{\bar{P}}{\sigma^2}}\right) \tag{3.38}$$

The channel access scheme and the order of SIC is identical to the OPT case.

3.6.2 Characterisation of the boundary of the capacity region

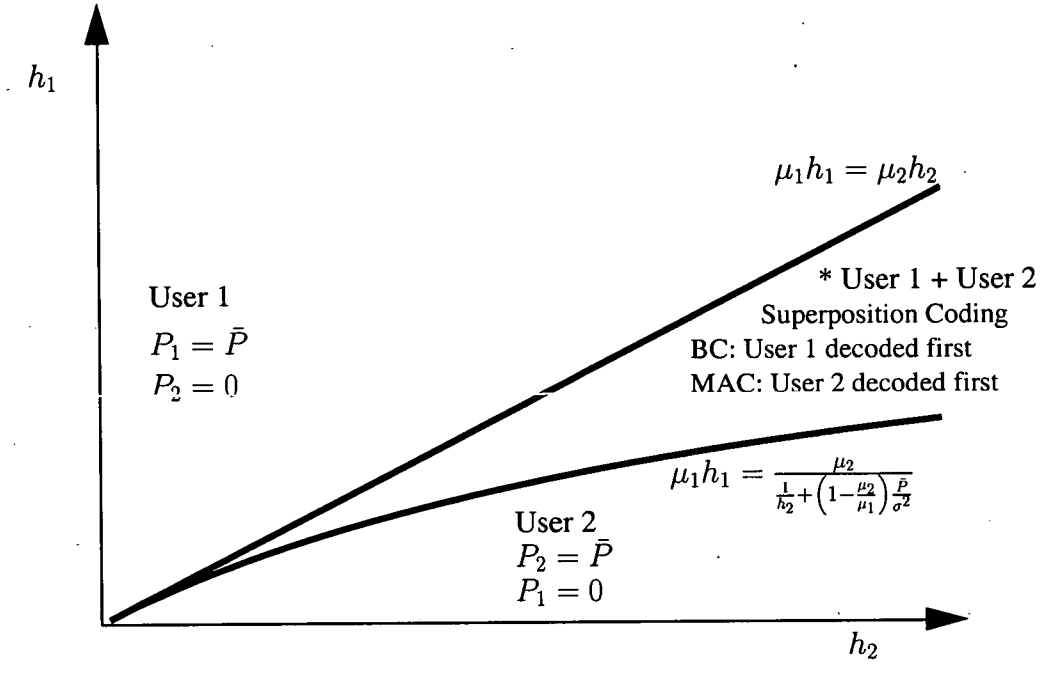
The boundary surface is characterised by the equation:

$$R_i^{\text{CP}} = \frac{1}{\ln 2} \int_0^{\frac{\bar{P}}{\sigma^2}} \int_0^{\infty} \frac{1}{\frac{1}{x} + z} f_{H_i}(x) \prod_{j \neq i} F_{H_j}(\beta^*)\, dx dz \tag{3.39}$$

where β defined in (3.27), and the notation $[x]^*$ in (3.29).

3.6.3 Example of two-user Case

Similar to Figure 3.1, Figure 3.5 gives an illustration of the resource allocation scheme for the two-user case under constant power constraint. The figure shows the dependence of the resource-allocation on the channel gains of the users.



$$*$$

$$P_{\text{sum}} = \bar{P}, \quad P_2 = P_{\text{sum}} - P_1, \quad \lambda = \frac{\mu_1}{\frac{1}{h_1} + \frac{\bar{P}}{\sigma^2}}$$

$$\text{BC: } P_1 = \bar{P} - \sigma^2 \left(\frac{\frac{\mu_2}{h_1} - \frac{\mu_1}{h_2}}{\mu_1 - \mu_2} \right)$$

$$\text{MAC: } P_1 = \frac{(\mu_1 - \mu_2)\bar{P} + \sigma^2 \left(\frac{\mu_1}{h_2} - \frac{\mu_2}{h_1} \right)}{\mu_1 \left(1 - \frac{h_1}{h_2} \right)}$$

Figure 3.5: Resource allocation under constant power constraint for the two-user case with $\mu_1 > \mu_2$

3.7 Solution with Orthogonal Signalling and Constant Power Constraints

When a constant sum-power constraint is used for each channel block, the problem formulation of optimal resource allocation in block k is given by

$$\max_{\{\tau[k],\mathbf{P}[k]\}} \sum_i \mu_i \tau_i[k] \log\left(1+\frac{h_i[k]P_i[k]}{\sigma^2\tau_i[k]}\right) \quad \text{subject to} \quad \sum_i P_i[k] = \bar{P}, \quad \sum_i \tau_i[k] = 1 \tag{3.40}$$

In [49, Section III.B.2, p. 1088–1089)] the solution of this problem is provided[7]. Depending on the channel condition $\boldsymbol{h}[k]$, either one user transmits (receives) or two users share the medium (by orthogonal signalling such as time or frequency division) in one channel state (details of power ($P_i[k]$) and time ratio ($\tau_i[k]$) calculations for this case are given in [49]). The maximum of two users sharing the channel block is regardless of the total number of users, as long as the fading statistics of the users' channels are independent and have continuous probability distribution PDF.

Figure 3.6 illustrates the two-user case. As shown in the figure, the region of shared access is narrow and thus re-formulating the problem with single-user scheduling in each block gives a very similar performance. This is studied in Section 3.8.

3.8 Solution with Single-User Selection and Constant Power Constraints

3.8.1 Resource allocation

In this case, one user m is scheduled to transmit (MAC) or receive (BC) in block k, and the power allocated to this user are calculated according to:

$$m = \arg\max_i \mu_i R_i[k] \quad \text{and} \quad P_i[k] = \bar{P} \tag{3.41}$$

$R_i[k]$ in (3.41) is the Shannon capacity of AWGN channel. Thus, we can write the selection

[7] We need only one part of the general solution provided in [49]: as in this case (i.e. OS-CP) we have constant sum power in every channel state, we do *not* need to optimise this sum power, but we do need the solution for the best possible distribution of the sum power to the users.

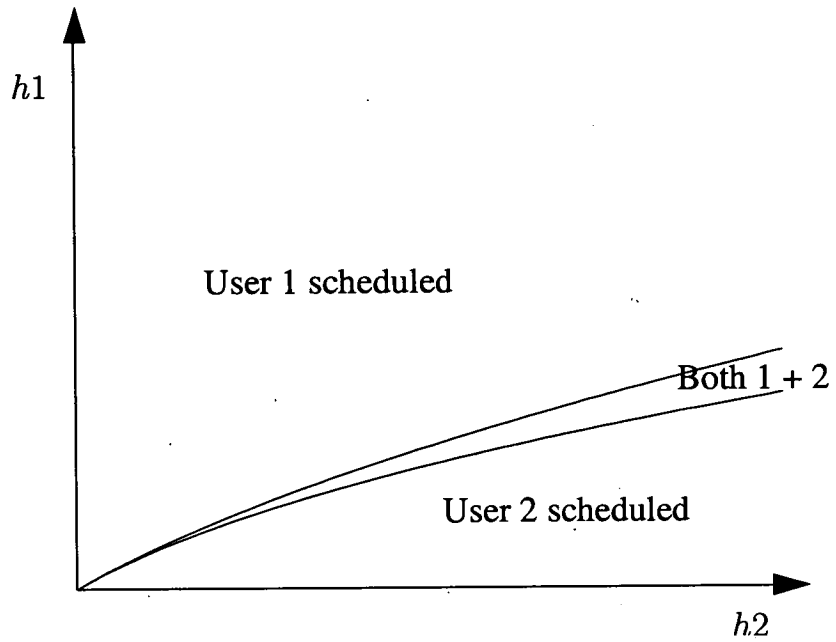


Figure 3.6: Dependence of scheduling decisions on channel conditions in the two-user case with orthogonal-signalling and constant-power constraints ($\mu_1 > \mu_2$).

strategy (3.41) as:

$$m = \arg\max_i \mu_i \log\left(1 + \frac{h_i[k]\bar{P}}{\sigma^2}\right) \tag{3.42}$$

3.8.2 Characterisation of the boundary of the capacity region

The boundary surface is characterised by:

$$R_i^{\text{CP-SU}} = \int_0^\infty f_{H_i}(x) \prod_{j \neq i} F_{H_j}(\eta) \log\left(1 + \frac{x\bar{P}}{\sigma^2}\right) dx \tag{3.43}$$

where η is defined as:

$$\eta = \frac{\left(1 + x\bar{P}\right)^{\frac{\mu_i}{\mu_j}} - 1}{\bar{P}} \tag{3.44}$$

The proof is provided in Section 4.2.

3.9 Numerical Examples

3.9.1 Comparison of two-user case

In this numerical example, the equations to characterise the boundary surface of the capacity region for the different cases under consideration are applied in a scenario of two users. Figure 3.7 shows the capacity regions with the assumption that the users channels are fading independently and with Rayleigh distributions. The first user channel has 10dB better long-term average channel quality over the second user channel. Any specific point in the capacity boundary can be achieved by adjusting the weighting factors μ (i.e. the curves in Figure 3.7 are produced by computing the corresponding equations that characterise the boundary of the capacity region over many selections of the weighting factors μ to scan the whole capacity region). A relevant case is selected in this example in which the network average spectral efficiency can range between 1 and 3 bits/sec/Hz.

The main conclusions that we obtain from this numerical example are:

- Power control (applied in OPT and SU) is more important when the operating point of the system has overall low spectral efficiency (to serve weak-channel users). For high spectral efficiencies, using a constant power per block is justified and has only minor detrimental effects to the capacity of the system.
- For constant transmit power systems, applying superposition coding (such as in CP) provides negligible improvements to the achievable rates. Thus, using single-user selection scheme (i.e. CP-SU) in such systems is justified. On the other hand, for systems applying optimal power control, superposition coding is useful for a range of operating points.

3.9.2 Sum-throughput comparison for symmetric channels

In this example, a comparison of the capacity limits in systems that apply optimal power control and systems that apply constant power per block is provided for various number of users. Since it is not possible to visualise the capacity regions for systems with more than 3 users, a specific operating point within the capacity boundary surface is used for the comparison. The maximum sum-throughput capacity is selected for the comparison and the analysis is done with the assumption of symmetric users channels. Furthermore, with the assumption of Rayleigh fading channels, we can derive close-form expressions for the capacities as a function of the number

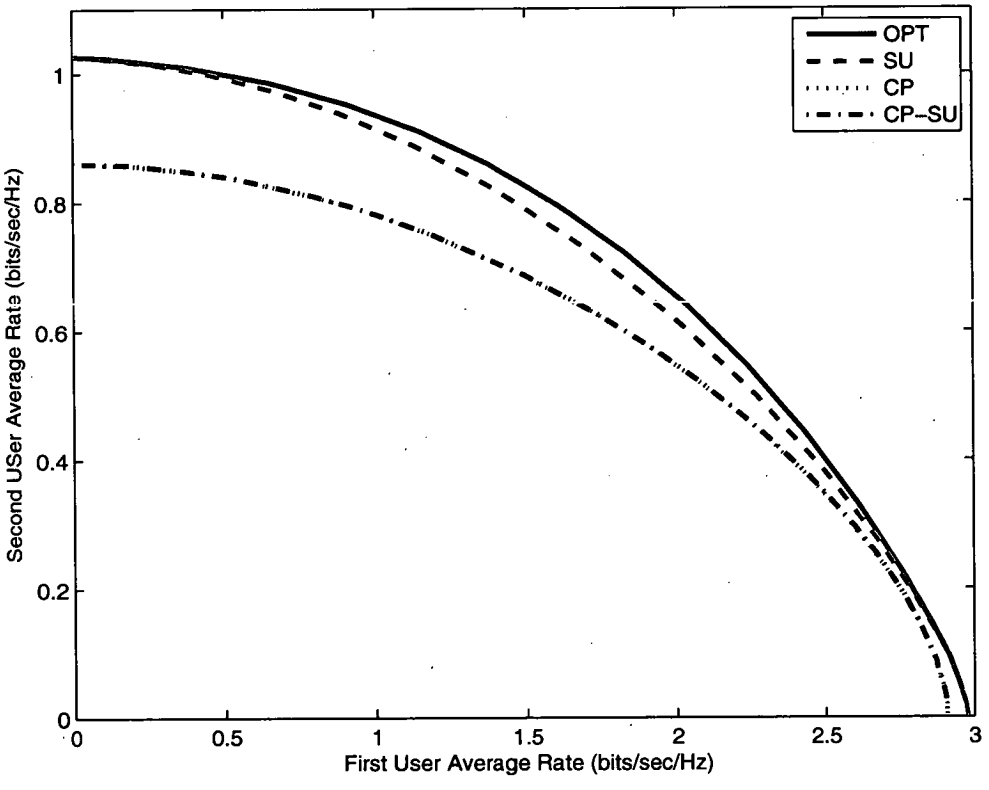


Figure 3.7: Boundaries of the ergodic capacity regions for the two-user case. The users are Rayleigh-faded with 10dB difference in average channel qualities.

of users M. In Appendix A.2 the derivations of the following equations are provided.

For the constant power system, we obtain:

$$R_{\text{sum}} = \frac{1}{\ln 2} \sum_{i=1}^{M} (-1)^{(i-1)} \binom{M}{i} \exp\left(\frac{i\sigma^2}{\bar{h}\bar{P}}\right) E_1\left(\frac{i\sigma^2}{\bar{h}\bar{P}}\right) \tag{3.45}$$

where E_1 is the exponential integral function

$$E_1(x) \equiv \int_x^{\infty} \frac{\exp(-u)}{u} du$$

For the system applying optimal power control, we obtain:

$$R_{\text{sum}} = \frac{1}{\ln 2} \sum_{i=1}^{M} (-1)^{(i-1)} \binom{M}{i} E_1(i\lambda) \tag{3.46}$$

where λ is adjusted so that the power constraint is achieved:

$$\frac{\bar{h}\bar{P}}{\sigma^2} = \sum_{i=1}^{M} (-1)^{(i-1)} \binom{M}{i} \left[\frac{\exp(-i\lambda)}{\lambda} - i\, E_1(i\lambda)\right] \tag{3.47}$$

From the results in Figure 3.8, we can find rough estimates of system spectral efficiencies over which the application of the constant power constraint is justified. As the number of users in the system increases, the rate level (i.e. spectral efficiency), over which the constant power system approaches the optimal power control system, decreases. For example, in a single user system, using constant power while operating above 4 bits/sec/Hz is very close to the optimal case. While a value of 3 bits/sec/Hz is applicable in a two users system, and approximately 1.5 bits/sec/Hz for $M = 10$.

3.10 Conclusions

The optimisation problem (2.3) is solved to obtain the optimal scheduling policies and characterise the multiuser capacity region under different constraints on the physical layer schemes. The block-fading channel is used to model time and frequency selective fading channels. The fading channels are divided into a family of flat-faded channels (blocks). A channel block could last for several time slots and frequency subcarriers as long as the channel quality is almost con-

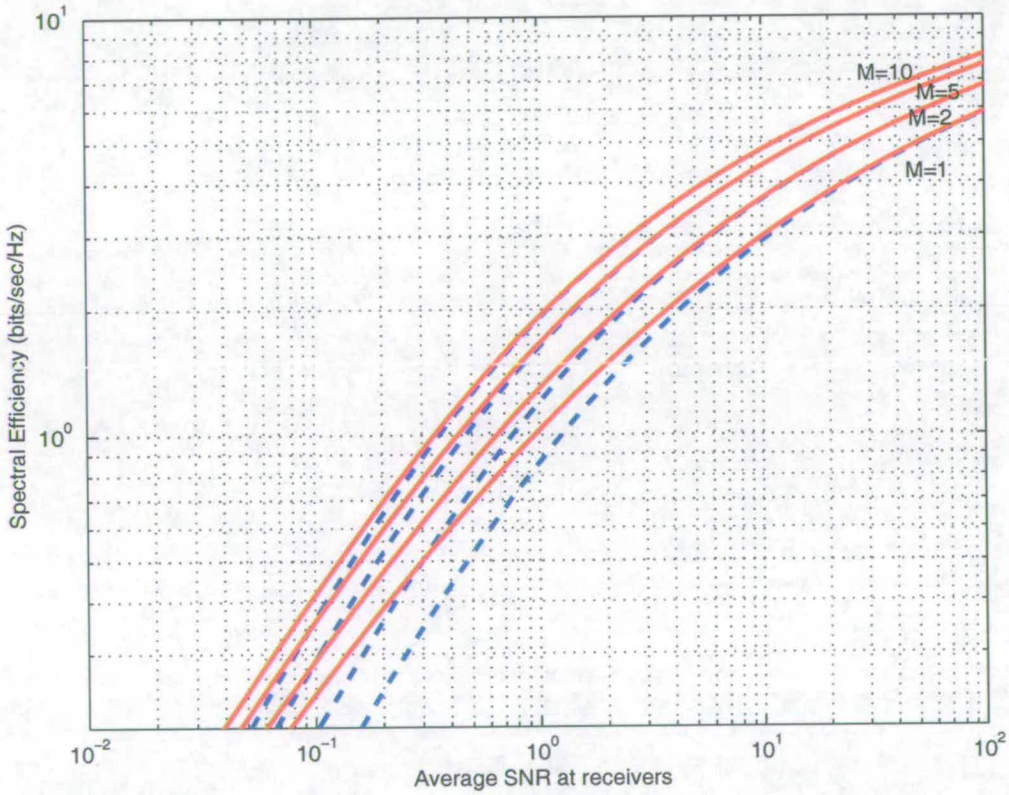


Figure 3.8: Differences in spectral efficiency achieved by systems applying optimal power control (solid lines) and systems applying constant power per channel block (dashed lines). The results are presented for different number of users M and with the assumption of Rayleigh fading channels with identical long-term average channel qualities of the users. The maximum sum-throughput is considered.

stant. The capacity of block-fading multiuser channels with channel-state-information (CSI) at both the transmitter(s) and the receiver(s) can be achieved by (i) optimal power allocation over the channel blocks and (ii) optimal resource (rate and power) allocation over the users in each of the channel blocks. As known from the literature, the optimal resource allocation scheme over flat-faded channel blocks involves applying CDMA (in uplink) or SC (in downlink) with SIC at the receivers. The number of users scheduled in a channel block varies depending on the channel conditions. The optimal power allocation scheme is given by the water-filling approach: more power is allocated when the channel is better, and based on the operating point of the system some users get higher average power. The optimal solutions (SC with SIC, and water-filling power control) are difficult if not impractical to implement. Thus, the problem is solved in this chapter with practical constraints of constant power control and orthogonal signalling.

The optimal scheduling policy depends on the system constraints such as power control, but does not depend on the fading statistics of the users' channels. With a single constraint on the total power transmitted, the optimal resource allocation schemes for the uplink and the downlink are exactly identical when orthogonal signalling is used. Under orthogonal signalling constraints, the optimal scheduling policy is (3.33) with water-filling power control. Under orthogonal signalling and constant power constraints, either one user is scheduled or a maximum of two users share the channel. However, most of the time one user only is scheduled. Thus, if a single user selection per channel block is applied instead of orthogonal signalling, the performance of the system remains almost the same. For constant power and single user selection per block, the best scheduling policy is (3.42). Power control is useful at low spectral efficiency. Applying constant power constraint is justified if the system is operating at high spectral efficiencies. Superposition coding with successive interference cancellation at the receivers can be useful in adaptive power systems, but is not needed in constant power systems.

Chapter 4
Case-Study: Comparison of The Performance of Known Scheduling Policies

The main contribution in this chapter is a new generic mathematical framework which allows one to systematically analyse and compare channel-aware multiuser scheduling algorithms that are applicable for delay-tolerant[1] applications over centralised wireless networks. The analysis is accompanied by a case-study to illustrate the theory. Although the objective here is *not* to provide a comprehensive review on the multiuser scheduling policies that are suggested in the literature, the provided survey includes well-known contributions in the field.

As discussed in Chapter 2, it is actually possible to compare multiuser scheduling algorithms for wireless networks in terms of their efficiency in allocating the physical resources (power and bandwidth), across the whole range of possible operating points. Efficient scheduling algorithms are those which operate on or close to the boundary of the capacity region.

The general assumptions that are presented in Section 1.2.1 are applied in this chapter.

4.1 Survey of Scheduling Policies

In the literature, there exist scheduling schemes which use similar policies, but with different operating points[2] depending on the control parameters of the policy such as weighting factors and rate offsets, which are adjusted according to the fairness criteria and constraints for the given application. Examples of schemes to select an operating point include proportional user-rate ratios (throughput fairness) [61], proportional channel-access ratios (resource-sharing fairness) [42], [36], and utility function maximisation [62].

[1] Algorithms that consider strict delay-constraints are, therefore, not included in the analysis. However, the main concepts presented in Chapter 2 are also applicable in this case.

[2] An operating point in a delay-tolerant system is defined by the vector of long-term average rates of the users.

The aim in this section is to provide a summary, along with a new unified algorithmic description, of scheduling policies proposed in the literature (all are applicable for both the uplink and the downlink) that lend themselves to a comparison in the proposed framework. This also includes some novel extensions of the originally proposed schemes. Of course, the presented choice of schedulers is far from being exhaustive, but I believe this selection is sensible and provides a useful basis for the comparison of other schemes as well.

Moreover, the main contribution in this chapter is a new framework (detailed in Section 4.2) with which we can analyse various types of schedulers; the specific scheduling algorithms considered here are actually just examples used to demonstrate the new analysis. There are many other schedulers that are not included in the comparison such as utility-based schedulers ([62], [63], [64], [24]), modified largest weighted delay first (M-LWDF) scheduler [65], queue proportional scheduler (QPS) [66].

Due to the objectives of this comparison study, scheduling algorithms that explicitly take delay-constraints into account (such as the ones presented in [67], [68]) are not included in the mathematical analysis.

4.1.1 Scheduling policies for constant-power systems

All policies in this section will schedule exactly one user per channel block and, if scheduled, each user's power is the same. In what follows, and in particular in Figures 4.2 and 4.3 which present numerical results (details are discussed in Section 4.3), the different scheduling and power allocation policies are referred to by their *equation numbers* stated below.

4.1.1.1 User selection based on weighted feasible rate

Scheduling policies which schedule user m in channel block k according to a weighted value of the instantaneous feasible[3] rate $R_i[k]$ of the user are given by

$$m = \arg\max_i \mu_i R_i[k] \tag{4.1}$$

[3]The well-known capacity equation for an Additive White Gaussian Noise (AWGN) channel is used to estimate the rate $R_i[k]$ from a given power $P_i[k]$ in this case. Moreover, the channel power gain $h_i[k]$ is assumed to be known at the transmitter, so that it can be exploited for scheduling decisions and/or transmit-power allocation.

The power allocated to the scheduled user is constant, i.e.

$$P_m[k] = \bar{P} \tag{4.2}$$

This power allocation rule is used by all scheduling policies with a constant transmission-power constraint.

Examples of suggested schedulers in the literature that use this scheduling policy include:

- In the maximum sum-throughput scheduler[4], the weighting factors are all equal, i.e., $\mu_i = 1$.
- In the Proportional Fair (PF) Scheduler [20], μ_i is inversely proportional to the average throughput $T_i[k]$ of the user in a past window, i.e., $\mu_i = \frac{1}{T_i[k]}$.
- In [67] proportional fairness is suggested with payloads c_i depending on the specific application ($\mu_i = \frac{c_i}{T_i[k]}$).
- In [69] this policy is suggested in a generic form to maximise throughput relative to pre-specified target ratios.

Although the PF scheduler is included in the numerical results in Section 4.3, it is not included in the new analytical framework in Section 4.2. The main reason is that delay-tolerant applications are considered only: in this case any dynamic adaptation (e.g. as above by $T_i[k]$) of the scheduler parameters μ_i is actually counter-productive with respect to the achievable long-term average rates we are interested in. Further details are discussed in Section 4.1.3.

4.1.1.2 User selection based on weighted channel quality

This policy is given by

$$m = \arg\max_i \mu_i h_i[k] \tag{4.3}$$

where $h_i[k]$ is the instantaneous power gain of user i's channel. In constant transmission power systems, this is equivalent to a policy which schedules the user with highest weighted received

[4] In [18] power control is used to achieve capacity. The same selection policy (4.1) maximises the sum-capacity under a constant transmission-power constraint.

Signal-to-Noise ratio (SNR).

This policy was suggested in the literature as follows:

- In [70] this policy is suggested in two forms: maximum throughput ($\mu_i = 1$) and proportional fairness ($\mu_i = 1/\bar{h}_i$), where $\bar{h}_i$ is the long-term average channel power gain.
- In Chapter 5 of this thesis, the policy is suggested in a generic form to achieve pre-specified resource-sharing ratios.

4.1.1.3 User selection based on feasible rates with rate offset

In [42] it is suggested to maximise the average total system performance while satisfying pre-assigned "time-fraction" (channel-access rate) requirements of the users. The proposal is generic and applicable to any system performance measures. This scheduling concept is included in the comparison with the assumption that throughput is the system-performance measure to maximise. The resulting scheduling policy is given by

$$m = \arg\max_i \left(R_i[k] + v_i\right) \tag{4.4}$$

where $R_i[k]$ is (as in Section 4.1.1.1) the feasible rate for user i in channel block k and v_i is a rate offset which is adjusted such that pre-assigned resource-sharing constraints are achieved.

4.1.1.4 User selection based on the cumulative rate-density function

In [43], [71] scheduling based on the cumulative density function (CDF) of user transmission rates $R_i(k)$ is suggested. The concept is to schedule the user whose rate is high enough, but least likely to take even larger values in other (e.g. future) blocks. This scheduling policy is given by

$$m = \arg\max_i \left(F_{R_i}\left(R_i[k]\right)\right)^{\frac{1}{w_i}} \tag{4.5}$$

where $F_{R_i}(.)$ is the CDF of the user's feasible transmission rates. The parameters w_i are used to scan different possible operating points of the system.

4.1.2 Scheduling policies for variable-power systems

4.1.2.1 User selection based on weighted feasible rate with water-filling power allocation

In [72] "Proportional Fairness" with QoS provision in downlink OFDMA is suggested. The user selection in each sub-carrier is based on PF scheduling ([20], see also Section 4.1.1) and the power $P_m[k]$ is allocated using the "water-filling approach":

$$P_m[k] = \max\left(\sigma^2\left[\frac{1}{\lambda_m} - \frac{1}{h_m[k]}\right], 0\right)$$

with user-individual factors $\lambda_m = \frac{1}{\lambda T_m}$ that depend on the average rate T_m recently achieved for user m in a moving time-window of limited size. The factor λ is adjusted such that a specified average power constraint is met.

A generalised version of this concept is used for the comparison-study: user selection is carried out by the general selection policy (4.1) based on weighted feasible rates, and the power allocation for user m (who is assumed to be scheduled in block k) is given by

$$P_m[k] = \sigma^2\left[\frac{\mu_m}{\lambda} - \frac{1}{h_m[k]}\right]^+ \tag{4.6}$$

where $[x]^+ \doteq \max(x, 0)$. The factor λ is again adjusted according to a long-term average power constraint, and μ_m are weighting factors used to pick a desired operating point. As in Section 4.1.1.1, the same comments apply with respect to a dynamic adaptation of the weighting factors (this is further discussed in Section 4.1.3).

It should be noted that for constant weights μ_i for all users this is the same as the suboptimal Time Division (TD) policy given in [49, Section III-C]. The performance of this scheme is, although very close, not optimal in a strict sense (i.e., the performance point does not lie on the boundary surface of the ergodic capacity region with a time-division constraint).

4.1.2.2 User selection based on weighted channel quality and simplified water-filling power allocation

In [73] it is suggested to use a normalised-SNR-based user selection strategy with water-filling power control along the sub-carriers. However, the water-filling level λ is adjusted irrespectively of the user selection policy. For comparison, a similar method is considered in a gener-

alised form: user selection is based on (4.3) and the power is controlled in the blocks according to

$$P_m[k] = \sigma^2 \left[\frac{1}{\lambda} - \frac{1}{h_m[k]}\right]^+ \tag{4.7}$$

Again λ has to be adjusted such that a long-term average power constraint is met. The difference to (4.6) is that the weighting factors μ_i used in the user selection are not used in (4.7).

4.1.2.3 User selection based on optimal scheduling policy under single-user selection constraint and water-filling power allocation

The optimal resource allocation scheme under single-user selection per channel block is also included in the comparison-study. The scheduling policy is (3.33) and the power is controlled using the water-filling approach (4.6).

4.1.3 Detrimental effect of a dynamic adaptation of the scheduler

In general all scheduling policies achieve their maximum performance when the control parameters (such as weighting factors or rate-offsets) are constant. But with constant control parameters it is impossible to control the delay and, hence, schedulers for delay-constrained applications often have to be dynamically adapted. A popular example is the Proportional Fair (PF) Scheduler [20], which uses (4.1) to take scheduling decisions but with the rate weighting factors adapted according to $\mu_i = \frac{1}{T_i[k]}$, where $T_i[k]$ is the "recently" achieved average rate in a moving time-window. Such (or any other) dynamic adaptation will decrease the achievable long-term average rate: this immediately follows from the convexity of the achievable rate regions. We may think of two points (rate-tuples) on the boundary of the scheduler's rate region that are achieved for two different parameter settings. When the scheduler parameters dynamically change between those two parameter sets, we can get the achievable rates pro-rata by time-averaging the achieved rates in both cases. Hence, we obtain a point on the straight line connecting the two points on the boundary of the rate region. As the region is convex, any point on this line will lie inside the rate region but not on its boundary and, therefore, any dynamic adaptation of the scheduler is inherently sub-optimal. The detrimental effect on the achieved rate will be the larger the larger the rate-differences between the two points are. Figure 4.1 shows a diagram explaining this concept.

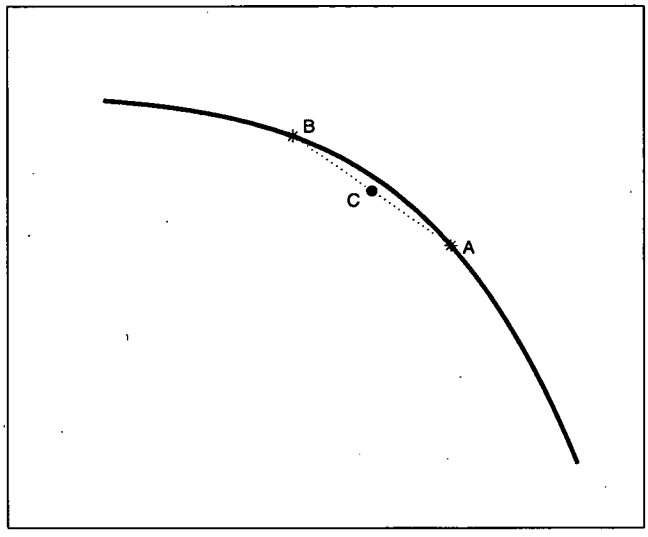


Figure 4.1: The detrimental effect of dynamic adaptation of the operating point of a scheduler. If the scheduler is operating at points A and B with equal probability, then the long-term average rate achieved by the scheduler is at point C which is not on the boundary of the capacity region.

As delay-tolerant applications are considered in the chapter, we will, therefore, not include any scheme in the theoretical analysis that uses dynamic adaptation of the scheduler parameters. However, we will compare the numerical performance achieved by the PF scheduler with the delay-tolerant schemes investigated.

4.2 Mathematical Framework for Performance Evaluation

When a scheduling policy allocates rate to a single user only in each channel block, the maximum possible achievable rate (bits/sec/Hz[5]) of user i who is scheduled in block k equals

$$R_i[k] = \log\left(1 + \frac{h_i[k]P_i[k]}{\sigma^2}\right) \tag{4.8}$$

for additive white Gaussian receiver noise with a variance of σ^2 – of course, (4.8) is the Shannon capacity for the AWGN channel. With AMC a rate close to capacity can be achieved (see, e.g.,

[5]This relates to each Hz of bandwidth on the radio-frequency bandpass channel. Bandwidth is defined as the width of a compact set of positive bandpass frequencies for which the signal spectrum is allowed to be non-zero. The occupied bandwidth in each real (I/Q) sub-channel of an equivalent complex baseband model is half the bandpass radio-frequency bandwidth.

[15]). In practice, wireless systems support a set of discrete rate values rather than a continuous range. However, our objective is not to evaluate the schedulers' achievable rates for some given set of practical modulation and coding schemes. Our goal is rather to evaluate the performances of scheduling schemes as such, without any system constraints that will change from one application to another. Therefore, we can use the ideal formula (4.8) to relate "power" and "rate"; the relative performances[6] of various scheduling schemes will carry over into practice.

In constant power systems, $P_i[k]$ in (4.8) is constant in all blocks in which user i is scheduled, i.e., $P_i[k] = \bar{P} \, \forall \, k$. In systems applying power control, $P_i[k]$ will be a function of $h_i[k]$ (see, e.g., (4.6)).

Assuming that all the blocks have identical bandwidths and time durations, the average achievable rate (bits/sec/Hz) of user i, in the blocks in which user i is scheduled, equals

$$\tilde{\bar{R}}_i = \frac{1}{|\mathcal{S}_i|} \sum_{n \in \mathcal{S}_i} \log \left(1 + \frac{h_i[n] P_i[n]}{\sigma^2} \right) \tag{4.9}$$

with $\mathcal{S}_i$ the set of indices of all channel blocks in which user i is scheduled.

Of course, user i does not transmit in all blocks but rather in a ratio ϱ_i of the total number of blocks. For a very large number of considered blocks this ratio converges against the probability that user i is scheduled and, hence, we set

$$\varrho_i = \Pr\{i \text{ is the scheduled user}\} \tag{4.10}$$

Thus, the achievable long-term average rate (bits/sec/Hz) of user i is

$$\bar{R}_i = \varrho_i \, \tilde{\bar{R}}_i = \frac{\varrho_i}{|\mathcal{S}_i|} \sum_{n \in \mathcal{S}_i} \log \left(1 + \frac{h_i[n] P_i[n]}{\sigma^2} \right) \tag{4.11}$$

The averaging in (4.11) over the realisations $h_i[n]$ from the set $\mathcal{S}_i$ of channel power gains can be replaced by an integration over a probability density function (PDF) of the random variable H_i by exploiting the fact that the random process created by a time series of realisations of H_i

[6]The "absolute" performance of a combination of specific modulation and coding schemes can often be approximated by (4.8) as well. An "acceptable" residual bit or frame error rate will often be achieved by a practical scheme with some (fairly constant) power-offset against the theoretical "zero-error" curve given by (4.8).

from the set $\mathcal{S}_i$ is ergodic[7]:

$$\bar{R}_i = \varrho_i \int_0^\infty \tilde{f}_{H_i}(h_i) \log\left(1 + \frac{h_i\, P_i(h_i)}{\sigma^2}\right) dh_i \tag{4.12}$$

where

$$\tilde{f}_{H_i}(h_i) \doteq f_{H_i}(h_i | i \text{ is the scheduled user}) \tag{4.13}$$

is the conditional PDF of the channel power gain of user i, given that user i is scheduled (to transmit (MAC) or to receive (BC)). The PDF (4.13) is different from that of the actual channel power gain since the user is transmitting with higher probability when the channel gain is larger. Furthermore, note that

$$\tilde{f}_{H_i}(h_i) = \frac{d\tilde{F}_{H_i}(h_i)}{dh_i} \tag{4.14}$$

with $\tilde{F}_{H_i}(h_i)$ the CDF of the channel power gain, h_i, over the blocks in which user i is scheduled (to transmit or to receive). We can write $\tilde{F}_{H_i}(h_i)$ equivalently as follows:

$$\begin{aligned}
\tilde{F}_{H_i}(h_i) &= F_{H_i}(h_i | i \text{ is the scheduled user}) && (4.15)\\
&= \Pr\{H_i \le h_i | i \text{ is the scheduled user}\} && (4.16)\\
&= \frac{\Pr\{H_i \le h_i, i \text{ is the scheduled user}\}}{\Pr\{i \text{ is the scheduled user}\}} && (4.17)\\
&= \frac{\Pr\{H_i \le h_i, i \text{ is the scheduled user}\}}{\varrho_i} && (4.18)
\end{aligned}$$

Thus, we obtain

$$\tilde{f}_{H_i}(h_i) = \frac{d\Pr\{H_i \le h_i, i \text{ is the scheduled user}\}}{\varrho_i\, dh_i} \tag{4.19}$$

Hence, we can rewrite (4.12) as

$$\bar{R}_i = \int_0^\infty v_{H_i}(h_i) \log\left(1 + \frac{h_i\, P_i(h_i)}{\sigma^2}\right) dh_i \tag{4.20}$$

where

$$v_{H_i}(h_i) \doteq \frac{d}{dh_i} \Pr\{H_i \le h_i, i \text{ is the scheduled user}\} \tag{4.21}$$

For adaptive power allocation systems, the power $P_i(h_i)$ in (4.20) contains λ which needs to be

[7]The scheduling decisions depend on the channel power gains of all users (and perhaps on a set of constant weighting factors), and all those channel power gains are assumed to form independent and ergodic random processes. Therefore, the new random process, which is created by considering the power gains only when the user is scheduled, will also be ergodic.

adjusted to maintain the average power constraint $\bar{P}$, i.e., λ is selected such that

$$\sum_i \bar{P}_i = \sum_i \int_0^\infty v_{H_i}(h_i) P_i(h_i) dh_i = \bar{P} \tag{4.22}$$

For systems with a constant-power constraint the situation is much simpler, as, if user i is scheduled in block k, $P_i(h_i[k]) = \bar{P}$ and, hence, $P_i[k] = \bar{P}$.

4.2.1 The functions $g_{ij}(.)$

The next step in order to evaluate (4.20) and (4.22) is to derive the relation of $v_{H_i}(h_i)$ in terms of the known weighting factors μ_m and the known unconditional channel PDFs $f_{H_m}(h_m)$ of all users.

We consider policies that schedule no more than one user in each channel block (i.e., there is no time-sharing between any two users within a channel block with constant channel gain). In order to find the probability that a user is scheduled we define, as a novelty, continuous non-decreasing auxiliary functions $g_{ij}\,(h_i[k])$ which can be used as follows to take scheduling decisions:

$$\text{user } i \text{ is scheduled in block } k \quad \text{if and only if} \quad h_j[k] < g_{ij}\,(h_i[k]) \quad \forall j \neq i \tag{4.23}$$

By (4.23), the functions $g_{ij}\,(h_i[k])$ are as yet only implicitly defined. A description is provided below of how to obtain them; of course, they will depend on the specific scheduling policy chosen. The functions $g_{ij}\,(h_i[k])$ describe the borders of the regions within the channel-gain vector-space $\boldsymbol{h} \doteq \{h_1, h_2, ...\}$ over which the different users are scheduled. In what follows the block index k are dropped, as the functions $g_{ij}(.)$ actually describe how decisions are taken by some policy for a given set of channel coefficients[8] ("channel state").

In general, the scheduling policies we consider have the format

$$m = \arg\max_i y_i(h_i) \tag{4.24}$$

[8] Of course the coefficients depend on the block index k and therefore the value of $g_{ij}()$ will also depend on k. The scheduling policy itself that is described by $g_{ij}()$ is, however, not dependent on the block index k.

where y_i is an increasing function of h_i. Then the only possibility that user i is scheduled is iff[9] the channel power gains $h_j \; \forall j \neq i$ are below certain values which are specified by the $g_{ij}(h_i)$ functions.

For example, if the scheduler is applying policy (4.3), i.e. $y_i(h_i) = \mu_i h_i$, then user i is scheduled iff for every other user $j \neq i$

$$\mu_j h_j < \mu_i h_i \quad \Leftrightarrow \quad h_j < \frac{\mu_i}{\mu_j} h_i \doteq g_{ij}(h_i) \tag{4.25}$$

This defines the function $g_{ij}(h_i)$ for this scheduling policy.

As another example, the scheduler may be applying policy (4.1), i.e. $y_i(h_i) = \mu_i R_i = \mu_i \log(1 + h_i \frac{\bar{P}}{\sigma^2})$. Then user i is scheduled iff for every other user $j \neq i$

$$\begin{aligned}
y_j(h_j) &< y_i(h_i) \quad \Leftrightarrow & (4.26)\\
\mu_j \log\left(1 + h_j \frac{\bar{P}}{\sigma^2}\right) &< \mu_i \log\left(1 + h_i \frac{\bar{P}}{\sigma^2}\right) \quad \Leftrightarrow & (4.27)\\
1 + h_j \frac{\bar{P}}{\sigma^2} &< \exp\left[\frac{\mu_i}{\mu_j} \log\left(1 + h_i \frac{\bar{P}}{\sigma^2}\right)\right] \quad \Leftrightarrow & (4.28)\\
h_j &< \frac{\left(1 + h_i \frac{\bar{P}}{\sigma^2}\right)^{\frac{\mu_i}{\mu_j}} - 1}{\frac{\bar{P}}{\sigma^2}} \doteq g_{ij}(h_i) & (4.29)
\end{aligned}$$

The derivation for scheduling policy (3.33) and power allocation policy (4.6) is given in Appendix B.1. Using the applied procedures in these examples, we can also obtain $g_{ij}(h_i)$ for all other scheduling policies under consideration. The results for $g_{ij}(h_i)$ are summarised in Table 4.1.

Having defined the functions $g_{ij}(h_i)$, we can now go back to calculate $v_{H_i}(h_i)$. We obtain from the definition (4.21) of $v_{H_i}(h_i)$ and from (4.23):

$$\begin{aligned}
v_{H_i}(h_i) &= \frac{d}{dh_i} \Pr\{H_i \leq h_i, i \text{ is the scheduled user}\} & (4.30)\\
&= \frac{d}{dh_i} \Pr\{H_i \leq h_i,\ H_j < g_{ij}(H_i)\, \forall\, j\} & (4.31)\\
&= \frac{d}{dh_i} \int_0^{h_i} \left(f_{H_i}(x) \prod_{j \neq i} \Pr\{H_j < g_{ij}(x)\} \right) dx & (4.32)
\end{aligned}$$

[9] if and only if

m	P_m	$g_{ij}(h_i)$
(4.1)	(4.2)	$\frac{\left(1+h_i\frac{\bar{P}}{\sigma^2}\right)^{\frac{\mu_i}{\mu_j}}-1}{\frac{\bar{P}}{\sigma^2}}$
(4.3)	(4.2)(4.7)	$\frac{\mu_i}{\mu_j}h_i$
(4.4)	(4.2)	$\left(\frac{\exp(v_i-v_j)\left[1+h_i\frac{\bar{P}}{\sigma^2}\right]-1}{\frac{\bar{P}}{\sigma^2}}\right)^+$
(4.5)	(4.2)	$F_{H_j}^{-1}\left([F_{H_i}(h_i)]^{\frac{w_j}{w_i}}\right)$
(4.1)	(4.6)	$\frac{\lambda}{\mu_j}\left(\frac{\mu_i h_i}{\lambda}\right)^{\frac{\mu_i}{\mu_j}} : h_i > \frac{\lambda}{\mu_i}$ and $0 : h_i \leq \frac{\lambda}{\mu_i}$
(3.33)	(4.6)	$\frac{-\lambda}{\mu_j W\left[-\left(\frac{\lambda}{\mu_i h_i}\right)^{\frac{\mu_i}{\mu_j}}\exp\left(\frac{\mu_i}{\mu_j}-\frac{\lambda}{\mu_j h_i}-1\right)\right]}$ $: h_i > \frac{\lambda}{\mu_i}$ and $0 : h_i \leq \frac{\lambda}{\mu_i}$
		$W[y]$ is the Lambert function [60], which is the inverse of $y = W\exp(W)$

Table 4.1: $g_{ij}(h_i)$ for the scheduling policies under consideration. Each scheduling policy is characterised by a selection policy to select the user m in block k, and a power-allocation policy P_m for the scheduled user. In the table we refer to the equation numbers of these policies. The block index k is omitted for brevity.

with (4.32) following from the independence of the channel power gains of the users. From the differentiation of the integral in (4.32) we obtain

$$v_{H_i}(h_i) = f_{H_i}(h_i) \prod_{j \neq i} \Pr\{H_j < g_{ij}(h_i)\} = f_{H_i}(h_i) \prod_{j \neq i} F_{H_j}\left(g_{ij}(h_i)\right) \tag{4.33}$$

with f_{H_i} the (unconditional and stationary) PDF of user i's channel gains and $F_{H_j}(h_j) \doteq \int_0^{h_j} f_{H_j}(x)dx$ the unconditional CDF of the channel power gains for the users j.

Equations (4.20) and (4.22) can now be evaluated by using $v_{H_i}(h_i)$ according to (4.33). Note that (4.33) involves the known simple channel models (unconditional PDFs and CDFs) for the users' channel coefficients. The structural properties of the scheduling policy and the power allocation scheme are completely captured by the newly defined $g_{ij}(h_i)$-functions: by use in (4.20), these $g_{ij}(h_i)$-functions allow for a simple evaluation of the achievable rate of any scheduling policy with single-user selection in each block k (see Table 4.1). By (if necessary numerical) evaluation of the integral (4.22), the $g_{ij}(h_i)$-functions also allow to adjust the control factor λ which is used, e.g., in power control policy (4.6). Note that the evaluation of (4.20) and (4.22) by means of the $g_{ij}(h_i)$-functions is a great simplification in comparison to a time-simulation of the scheduler with its associated power control strategy and time-averaging of the rates over "many" channel realisations to obtain statistically significant values for the rate-averages. Moreover, the $g_{ij}(h_i)$-functions provide a useful tool for an analytical characterisation of scheduling policies; we provide results in Section 4.3 that would, due to the extensive simulation time required, be very hard to obtain by simulation and time-averaging.

4.3 Numerical Examples

Equation (4.33) is applied to evaluate (4.20) and (4.22) under the assumptions that the magnitudes of the users' channel coefficients have Rayleigh (3.3), (3.4) or Rice (3.5), (3.6) distributions. The two-user case is considered in two examples. This is for clarity only, as it is difficult to visualise and compare higher-dimensional rate regions. Of course, the mathematical concepts and the results from Section 4.2 can also be applied to the general M-user case. Moreover, the numerical results for the two-user case presented below provide interesting insights that will carry over to the M-user case.

Figures 4.2 and 4.3 show numerical results. Similar to Table 4.1, the user selection policy and

power allocation policy of each scheduling scheme shown in the figures are indicated by their *equation numbers*.

In Figure 4.2 the achievable rates (in bits/sec/Hz) are depicted using the policies which schedule a single-user per channel block with constant transmit power. The assumption here is that the first user has a Rice fading channel ($\kappa = 10$), with a long-term average channel power gain that is 10dB higher than that of the second user, who has a Rayleigh fading channel. Figure 4.2 shows that the weighted feasible-rates policy (4.1) is the best (as expected by information theory, see Chapter 3). However, the weighted channel-gains policy (4.3) works almost as good for all operating points and, thus, is attractive for generic schedulers, as (4.3) does not need to assume a particular relation (such as the AWGN capacity equation) between "power" and "rate". Another advantage of this policy is that, unlike the weighted feasible-rate policy, it has a continuous probability distribution and thus, with probability of one, a single user will maximise the scheduler metric.

Policy (4.4) coincides with the constrained capacity boundary given by policy (4.1) (which has the best possible performance for constant power) at the maximum sum-throughput point[10]. The latter is achieved by policy (4.1) when its weighting factors are all equal, and it is achieved by policy (4.4) when the rate offsets are all "zero". For all other points policy (4.4) has degraded performance in comparison to policy (4.1), and this degradation is larger when the system operates at low spectral efficiency (low sum-rate); the degradation at high spectral efficiency is however very small. Policy (4.5) is similar to policy (4.4) in that it has degraded performance compared to (4.1), but unlike (4.4) its best performance (again coinciding with policy (4.1)) is not at the maximum sum-throughput but rather on another operating point which depends on the fading channel models; that is why we observe different results in Figures 4.2, 4.3 for both policies.

In Figure 4.3 we investigate scheduling policies involving power control. We assume Rayleigh fading channels for both users, with 10dB higher average channel power gain for the first user. For comparison, the constant-power policies investigated in Figure 4.2 already[11] are also included. The boundary of the ergodic capacity region for schemes using superposition coding and optimal power control are also shown for comparison.

[10]Policies (4.4) and (4.1) also coincide for the the trivial case that either user 1 or user 2 are scheduled all the time, but this is no longer a "multiuser case".

[11]There is no duplication in the results, as the channels are different from those in Figure 4.2.

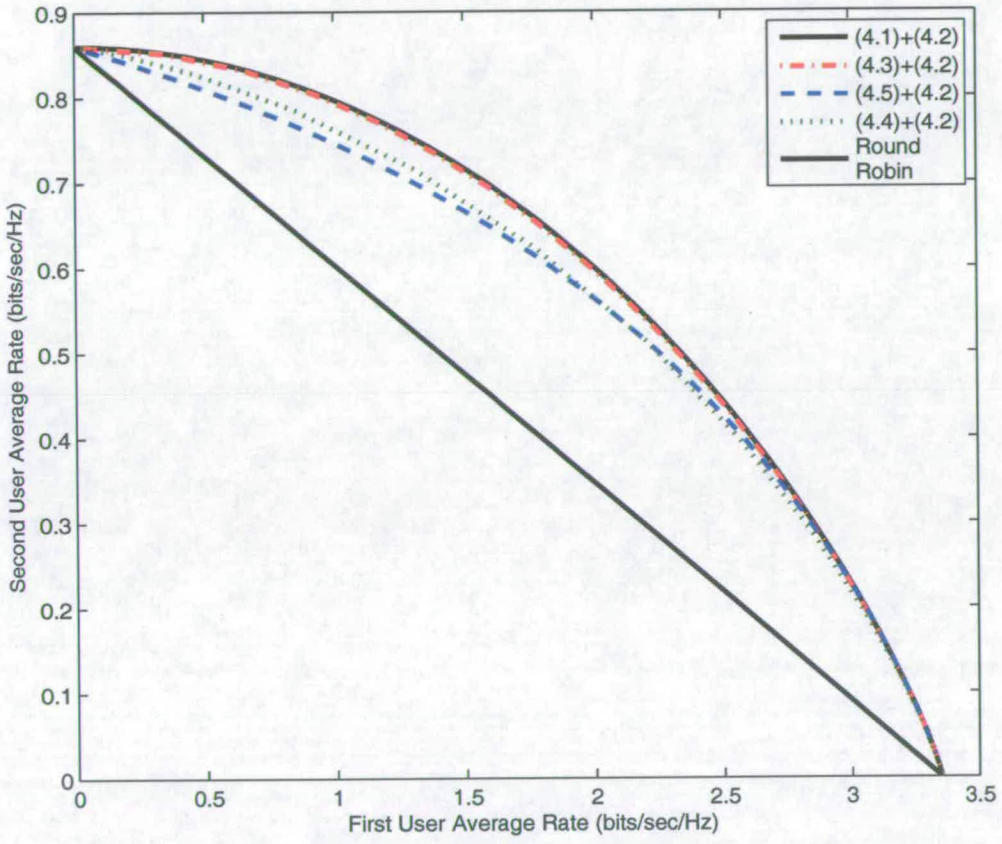


Figure 4.2: Achievable long-term average rates of the constant-power-per-block scheduling policies for the two-user case. The scheduling policies are indicated by their equation numbers in the legends; (4.2) is the equation number of the constant-power allocation policy. The channel coefficient of the first user has a Rice-distribution ($\kappa = 10$), while Rayleigh fading applies to the second user. The average channel power gain of user 1 is 10 dB higher than that of user 2. The Round-Robin policy schedules user 1 in odd-numbered channel blocks and user 2 in even-numbered blocks, regardless of their channel coefficients; power allocation is also by (4.2).

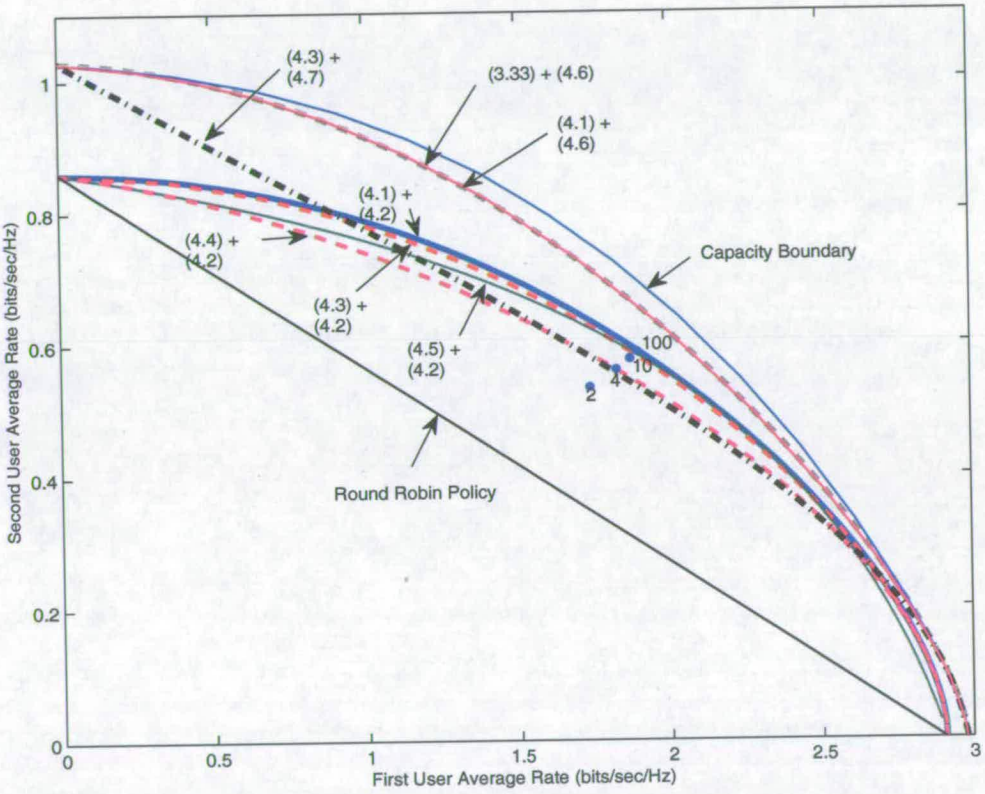


Figure 4.3: The capacity region of two-user case and the performance of various scheduling policies (section 4.1). The channels are assumed to be Rayleigh faded with 10 dB difference of the average channel power gains. The point indicated by "*" is related to the explanation of transmit power statistics in Figure 3.3. The small dots indicate the operating points of the proportional fair scheduler and the number beside each dot is the size of the window (in time slots) over which the "current" throughput (average rates) is computed in the PF scheduler.

Figure 4.3 demonstrates that power control is helpful for the users with low-spectral efficiency (low rate), while using constant power is justified when operating at high spectral efficiency. As a compromise, we may use a limited number of different power levels (as long as this does not cause interference problems) to be better able to support users with bad channels. The power allocation using water-filling in single-user selection means allocating more power when the channel is better. In multiuser communication, it also means allocating more power to the user with higher rate-reward μ. Thus, policy (4.7) is a bad choice. The selection policy (4.1), combined with power allocation policy (4.6), gives a performance very close to the theoretically optimal policy (3.31). Policy (4.3), which again performs very well under constant-power constraints, could, in principle, be used in conjunction with the power allocation policy (4.6). As the latter, however, depends on the channel noise model (it is optimal for a Gaussian channel and assumes the AWGN capacity to relate power and rate) the advantage of (4.3) (independence of a particular power-rate relation) in the constant power-case would not carry over: hence, we better use selection policy (4.1).

Furthermore, the operating points of the proportional fair (PF) scheduler [20] are also included in Figure 4.3 with different sizes of the window over which the average throughput is computed. Note that the PF scheduler does *not* allow for a trade-off between user-rates, i.e., a particular "average" point that lies *within* the capacity region is implicitly picked by the policy. Hence, PF scheduling does not allow for a flexible choice of the operating point which is a major disadvantage compared to other policies investigated.

From the performance-points of the PF scheduler in Figure 4.3 we observe that the smaller the size of the window is, the stronger are the fluctuations in the values of μ (see Section 4.1.1.1). In order to operate close to capacity boundary, the variations in the weighting factors should be minimal (constant values are preferred, see Section 4.1.3). This fact indicates that scheduling algorithms which depend on any dynamic measure (like the queue size or recently achieved rates) to adjust μ have degraded performance. When traffic load is dynamic but delay constraints are relaxed, a better approach is to define windows over which the μ-values are constant, and the values may be updated for the next window depending on the current traffic load.

4.4 Conclusions

In the literature, there exist many channel-aware scheduling schemes for the multiuser wireless channels. In this chapter, many of the known schedulers which are applicable for delay-tolerant applications over centralised wireless networks are analysed and compared based on the concepts presented in Chapter 2. The outcome of this comparison study is to understand the drawbacks of not applying the optimal scheduling policies presented in Chapter 3. A new generic mathematical framework is suggested to systematically analyse and compare the capacity regions achieved by the considered scheduling policies. The analysis is accompanied by a case-study of the two-user case to illustrate the theory. The multiuser scheduling algorithms for wireless networks are compared in terms of their efficiency in allocating the physical resources (power and bandwidth), across the whole range of possible operating points. Based on the results of the case-study, we know that some scheduling policies are good for some operating points only, as they are not generic. There are policies which can be used in generic schedulers, such as (4.3), because they have close-to-optimal performance for all operating points on the capacity boundary. Achieving any kind of fairness between users or maximising any performance metric of the network should be done by properly adjusting the control parameters of the optimal scheduling policy, and not by using a different scheduling policy.

Dynamic "on-line" variation of the weighting factors degrades the performance and will be avoided by a good scheduler if the applications are delay tolerant. However, in case of delay sensitive applications, some real-time adjustment of the control parameters is needed. This is known as the trade-off between throughput and delay constraints. If the system traffic load is dynamic, the weighting factors should be updated according to the changing conditions of the system as discussed in Chapter 5.

Chapter 5
Generic Flexible and Optimal Scheduler Structure

Good schedulers for wireless networks should have two important features: (i) *optimality in allocating the physical resources*, and (ii) *flexibility in controlling the operating point of the network*. As discussed in Chapter 2, the proper way to define optimality of a scheduler is to operate at Pareto-optimal conditions in which we can not increase the rate of one user without decreasing the rates of other users or using more physical resources. In order to operate at the Pareto-optimal points, there are unique scheduling policies that achieve these optimal conditions for the specific systems' constraints. The optimal scheduling policies have control parameters – such as weighting factors – which control the allocated resources and rates to every user. Thus, proper selection of the scheduling policies' weighting factors allows to operate at any of the points within the set of Pareto-optimal operating points. *A flexible scheduler provides the network operator with the possibility to select to operate at any of the possible Pareto-optimal points*. Thus, a flexible scheduler should allow to control the weighting factors of the optimal scheduling policies based on ,e.g., fairness criteria.

In this chapter, a flexible and optimal scheduler for centralised wireless networks is suggested. Two algorithms – offline and online versions – are suggested to control the weighting factors of a close-to-optimal scheduling policy based on the resource sharing fairness criteria. *The scheduler is proposed for systems applying the practical constraints of single user selection and constant transmitted power per channel block.* Furthermore, the suggested "on-line" algorithm to control the weighting factors of the scheduling policy provides with the possibility to enforce some restrictions on the delay of the packet transmission in a way similar to the approach applied in the well-known proportional fair scheduler [20].

A description of the suggested scheduler is provided in this chapter followed by the simulation results to show how the suggested algorithms allow to achieve the flexible fairness constraints. Furthermore, the multiuser diversity gains that can be achieved by the scheduler are analysed.

In order to make this chapter easy to follow, the detailed proofs of the derived equations are presented in Appendix C

5.1 Resource-Sharing Fairness

As well-known, maximising the sum-throughout over a wireless network – by always transmitting only over the best channel – will result in users with good channels dominating the network access and in "unfair" allocation of the air-link resources among the users. The proportional fair scheduler (PF) [20] provides a good compromise between multiuser diversity gains and fairness. However, it has the drawback that it does not provide any flexibility in controlling the resources share given to each user. As well-known, the resources demands differ depending on the services requested by the users or the pricing policy, etc. Thus, more generic scheduler structures are needed which obtain multiuser diversity gains while still maintaining the fairness and the Quality-of-Service (QoS) requirements.

Among many others one can identify two approaches in the literature that provide additional flexibility in allocating the resources to the users: proportional user-rate ratios [61] and proportional channel-access ratios [42]. The first approach provides **throughput fairness** by controlling the rates of the users relative to each other, while the latter provides **resource-sharing fairness** by controlling the amount of air-link resources allocated to the users relative to each other. However, the definition of fairness is ambiguous in the case of wireless networks. This is because some users may demand more channel access than others depending on their channels. Thus, achieving throughput fairness in wireless networks may lead to users of bad channels occupying most of the air-link resources and to severe degradation in the achievable total throughput of the system. Alternatively, if the resources are divided fairly between the users, then every user will have throughput rates according to his average channel quality without degrading the service for other users.

5.2 Background of the Suggested Scheduler

The aim of the suggested channel-aware scheduler is to achieve pre-selected resource-sharing constraints in systems applying constant transmit power and single user selection per channel block. Although one solution was provided in [42], the problem is re-visited in this work

to propose improvements from a practical perspective. In [42] it was suggested to maximise the average total system performance while satisfying pre-assigned channel-access ratios (for example, the percentage of time slots in a TDMA system) of the users. The proposal is generic and applicable to any system performance measures. The resulting scheduling policy is (4.4). The disadvantage of this policy is that it always favours the better-channel user. Although it guarantees the allocation of a share of the resources to all users, it is unfair in that the user of the better channel accesses the network with better channel relative to its average so that users with better channels obtain higher multiuser diversity gains. Furthermore, as shown in Figure 4.2, this policy is not optimal because it operates away from the boundary of the capacity region except at the maximum sum-throughput point.

As a better approach to the problem, we can apply the resource-sharing constraints as suggested in [42], but – unlike [42] – we use a different "on-line" scheduling policy (i.e. different than the one proposed in [42], refer to equation (4.4)). The proposed scheduling policy, namely the Channel-Quality-Based scheduling policy, is defined by the selection strategy (4.3). In [74], [70] the scheduling policy was suggested in two forms: maximum throughput and proportional fairness[1]. Furthermore, analysis of the performance of this policy for these two operating points was provided in [75]. The scheduling policy (4.3) is suggested in this work in a novel generic form to meet *general* resource-sharing constraints. As shown in Figure 4.2, the generic channel-quality-based scheduling policy have close-to-optimal performance over all possible operating points. Thus, it can be used in generic schedulers. Furthermore, it has an advantage over the optimal scheduling policy (4.1) that it does not need any further computations based on the channel quality measurements to obtain the feasible rates of the users. Furthermore, with probability one a single user should maximise the scheduling metric (4.4), while in (4.1) more than one user may maximise the metric based on the system supported rates.

In the new approach, the scheduler is divided structurally into two main parts. The first part decides the long-term average channel access ratios (resources-sharing) for each user ($\mathbf{ar}$), and the second part takes the channel variations into account and schedules the transmissions in order to achieve multiuser diversity gains. We are not concerned about the selection of resource-shares given to each user: this is a degree of freedom which helps to define suitable fairness criteria among users or to differentiate between users based, e.g., on the pricing policy.

[1] We prefer to call this point The Equal Resource-Share Fairness operating point. Refer to Appendix C.2 for a proof. Although the name proportional fairness was given in [70], the original proportional fairness scheduler is different because it is based on a different scheduling policy; the maximum weighted feasible rate policy.

The main contribution is the proposal of iterative algorithms that control the weighting factors μ of the scheduling policy (4.3) so that general resource-sharing constraints can be met.

Some examples of the possible approaches to decide the channel access ratio[2] include giving all users the same amount of channel access regardless of their location or in a weighted version dependent, e.g., on a pricing policy or on location by giving bad channel users a higher percentage of channel access. In addition, if some applications require maintaining long-term average rate constraints, the resource shares can be selected so that users with rate constraints receive an amount of the resources sufficient to achieve their rate constraints. Furthermore, in case of bursty applications in which the rate demands vary with time, some feedback can be applied to update the assigned channel access ratios to the users on a regular basis so that the instantaneous rate demands are taken into consideration.

As an example, if the network load is mainly data traffic, then the network operator may divide the users into two groups based on their average link quality (i.e. distance from the base station). The users within each group receive identical amount of channel resources (i.e. channel access ratios). However, a user within the bad-channel group receives twice the amount of channel access rate than a user within the good-channel group. So, if we assume that there are 4 users with good channel and 3 users with bad channel, then each user of the first group will be allowed to access the channel 0.1 of the time, while each user of the second group will access the channel 0.2 of the total time. As another example, if we assume that there are two group of served applications based on the rate requirements (one application requires twice the date rate than the other), then the users can be grouped based on their application where the users in the same group receive identical amount of resources, but different than the other group. In this example, the users in one of the two groups will receive twice the channel access ratio than the users in the other group.

5.3 Generic Channel-Quality-Based Scheduling Policy

This scheduling policy is based on the selection strategy (4.3).

[2]In general, the physical resources in wireless networks are mainly the transmission power, the time and the bandwidth. Since, we are considering constant power systems, the differentiation between users in terms of resources shares is corresponding to the shares of air-link resources (bandwidth and time) that they are receiving in the long term. The transmission power remains constant and is not controlled according to the resource sharing fairness constraints.

5.3.1 Achievable rates and access ratios with constant parameters of the scheduling policy

The long-term average achievable rates using the suggested scheduling policy (4.3) in its generic form for a given transmit power $\bar{P}$ are[3]:

$$R_i(\bar{P}) = \int_0^\infty f_{H_i}(x) \prod_{j \neq i} F_{H_j}\left(\frac{\mu_i}{\mu_j} x\right) \log\left(1 + x\frac{\bar{P}}{\sigma^2}\right) dx \tag{5.1}$$

where f_{H_i} and F_{H_i} are the stationary probability density function and the cumulative distribution functions of the fading process of user i respectively.

The percentage of channel access which user i gets (channel access ratio ar_i) is:

$$ar_i = \int_0^\infty f_{H_i}(x) \prod_{j \neq i} F_{H_j}\left(\frac{\mu_i}{\mu_j} x\right) dx \tag{5.2}$$

The proof of (5.1) is given in Chapter 4. The derivation of (5.2) is given in Appendix C.1.

5.3.2 Capacity region analysis

From Figure 5.1, which is for the Rayleigh faded channels, we see that the performance of the proposed scheduler which uses a constant power scheme is close to the ergodic capacity region boundary, obtained from information theory (requires the application of SC with SIC at the receivers, and power control), especially when the system is operating at high spectral efficiency (bits/sec/Hz). For network scenarios which involve users with bad average channel qualities causing low spectral efficiency, the performance of the proposed scheduler is not close to the capacity because it applies constant power instead of the optimal water-filling power control needed to achieve capacity. However, it should be kept in mind that the transmitted power using such A scheme may vary a lot which may cause interference problems.

[3]The weighting factors in (4.4) are assumed to be constant in the derivation of (5.1). However, constant weighting factors are applicable when the applications are delay-tolerant. If the weighting factor are updated continuously as in the real-time algorithm, then the performance will be slightly degraded. The degradation depends on how big the fluctuations are in the weighting factors.

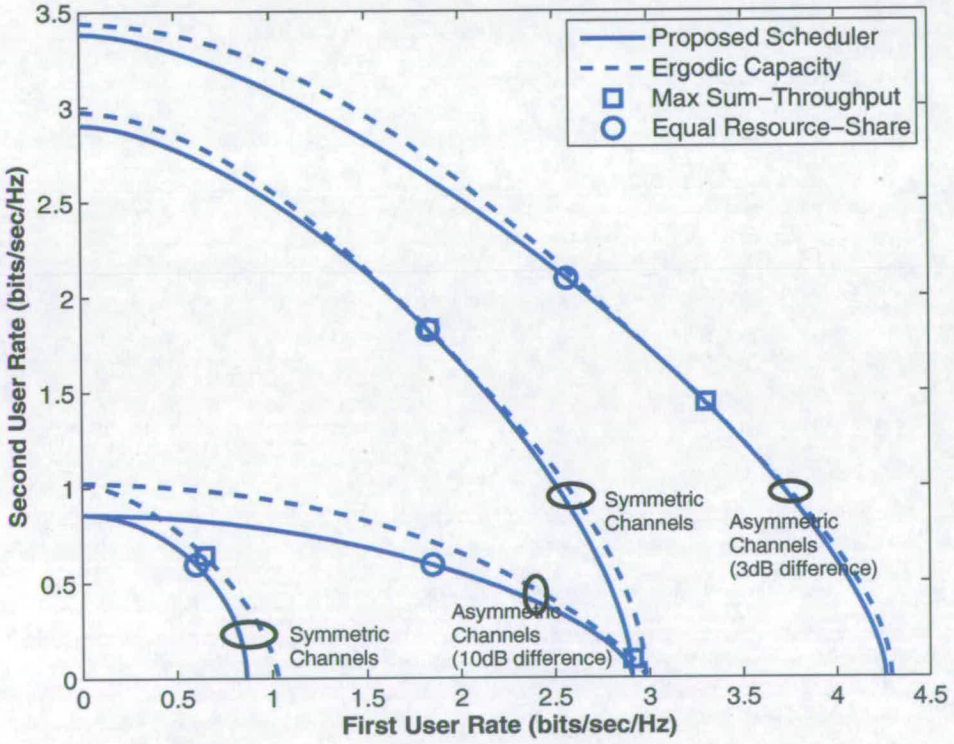


Figure 5.1: The ergodic capacity region of two-users case and the performance of the proposed scheduler. The channels of the two users are assumed to be Rayleigh faded. Four different cases in terms of the long-term average channel qualities $\bar{\mathrm{h}}$ are shown

5.4 Algorithms to Achieve Resource-Sharing Constraints

Two designs are proposed to find the appropriate weighting factors in order to achieve the pre-selected long-term average channel access ratios of the active users.

5.4.1 First design: off-line calculation of the weighting factors

The first design involves off-line calculation of the weighting factors μ_i for the different users. This design is applicable when the channels statistics are known in advance and are not changing with time. In the case of mobile users, the average channel qualities change as the users change their locations. In such cases, this design can not be applied. However, if the changes in average channels' qualities are slow, then the design can be applied by repeating the calculations from time to time once a change of the channels' average values or distributions is detected. In real-time operation, the weighting factors are constant, and thus, this design is applicable to delay-tolerant applications only.

The weighting factors μ are calculated in an iterative algorithm. The initial values of the weighting factors are chosen to be close to the actual operating point, and the updating increments in each iteration are selected such that stability and fast convergence towards the operating points are achieved.

The initial values are selected to be:

$$\mu_i^{(0)} \equiv \frac{ar_i}{\bar{h}_i} \tag{5.3}$$

ar_i is the desired access ratio. The suggested initial values are based on our observations that they give good starting points for the algorithm and help to achieve faster convergence of it, i.e., the initial choice is heuristic and based on experience.

In each iteration l we compute[4]:

$$ar_i^{(l)} = \int_0^\infty f_{H_i}(x) \prod_{j \neq i} F_{H_j}\left(\frac{\mu_i^{(l)}}{\mu_j^{(l)}} x\right) dx \tag{5.4}$$

[4]Obtained from (5.2).

Then we update μ for the next iteration:

$$\mu_i^{(l+1)} = \mu_i^{(l)} + \frac{\Delta}{\bar{h}_i}\left(ar_i - ar_i^{(l)}\right) \tag{5.5}$$

The main motivation behind this procedure is that if, for a given selection of weighting factors, a user accesses the channel more often than required, we produce a small decrement in his corresponding weighting factor (which will let the user access the channel less), and vice versa. Furthermore, with the suggested way to update μ, we guarantee that always ($\sum_i \mu_i \bar{h}_i = 1$) and this provides stability to the algorithm because it avoids the case that the weighting factors are updated independently of each other. The ratios of the weighting factors affect the operating point of the system, not their magnitudes.

The iterations terminate when the error in $\mathbf{ar}$ is within a pre-selected error tolerance. In (5.5), the selection of step size Δ affects the speed of convergence. Selecting too small Δ will make the convergence slow. However, if Δ is too high, the error will diverge. Hence a compromise for Δ must be found, which can be easily done heuristically by a variety of different choices; according to our experience this is by no means a critical issue.

5.4.2 Second design: real-time adaptation of the weighting factors

There are two main reasons for selecting a real-time adaptive approach. First, applying the first "off-line" approach requires that the channel fluctuations statistics (probability density functions) be perfectly known. This might not be the case. Second, an adaptive design is applicable in the case of mobile users, since it is based on the real-time measurements of channel conditions.

Thus, similar to the first design, the approach is to initially choose the weighting factors to be equal to

$$\widehat{\mu}_i[0] \equiv \frac{ar_i}{\bar{h}_i} \tag{5.6}$$

Alternatively, better initial values can be obtained from the off-line approach, i.e., both designs are used: the first one to have very close initial values, and the second adaptive one to maintain the required $\mathbf{ar}$ even when the users are moving.

At each time index k, one user is scheduled in each frequency band b (refer to Figure 1.1)

according to the scheduling policy:

$$m = \arg\max_i \widehat{\mu}_i[k] h_i[k,b] \tag{5.7}$$

Then, based on scheduled users, the real-time measured channel-access ratios are updated:

$$\widehat{ar}_i[k] = \left(1 - \frac{1}{t_c}\right) \widehat{ar}_i[k-1] + \frac{1}{t_c} \frac{B_i[k]}{B} \tag{5.8}$$

B is the number of frequency (flat faded) slots within the total bandwidth, and $B_i[k]$ is the number of slots in which user i is scheduled to transmit (or receive). The scalar t_c is the number of time slots which define the size of the window over which the channel access ratios are measured. The larger t_c the better. This is because high t_c prevents big fluctuations in μ. However, t_c should not be too big as this will lead to an unwanted delay in adapting the weighting factors to changes in the channels statistics. The parameter t_c allows the network operator to choose a suitable trade-off between the throughput gains and the delay restrictions.

In a similar way, the real-time measured average channel qualities are updated:

$$\widehat{\bar{h}}_i[k] = \left(1 - \frac{1}{t_c}\right) \widehat{\bar{h}}_i[k-1] + \frac{1}{t_c B} \sum_{b=1}^{B} h_i[k,b] \tag{5.9}$$

The values of μ are updated for the next time slot according to the following formula:

$$\widehat{\mu}_i[k+1] = \widehat{\mu}_i[k] + \frac{\Delta}{\widehat{\bar{h}}_i[k]} \left(ar_i - \widehat{ar}_i[k]\right) \tag{5.10}$$

Our experiments show that in the real-time system, Δ should be much smaller than for the off-line calculations. For example, in our simulation results shown in Sec 5.5, the value of Δ for the off-line calculations was in the range (0.3 to 1.0), while for the on-line approach, the value of Δ was two or three order of magnitudes smaller than the off-line case with values in the range (0.0001 to 0.005).

5.4.3 Empty queues

When a user is scheduled according to policy (4.3) but has an empty queue or a required rate less than the feasible rate supported by the instantaneous channel of the user, then this user is

not scheduled for this time slot and another user is selected. In wideband systems, if a user is scheduled to transmit on a number of frequency bands that exceeds the number of bands he actually needs, then those extra frequency bands are made available for other users.

5.5 Simulation Results

5.5.1 Simulation of the off-line calculations approach

A simulation of the off-line calculations approach was performed. It was assumed that there were six users with relative average channel qualities:

$$\bar{\mathbf{h}} = [1 \;\; 1 \;\; 0.6 \;\; 0.4 \;\; 0.4 \;\; 0.1]$$

The required access ratios $\mathbf{ar}$ were assumed to be:

$$\mathbf{ar} = [0.15 \;\; 0.1 \;\; 0.3 \;\; 0.12 \;\; 0.13 \;\; 0.2]$$

The error $||\mathbf{ar} - \mathbf{ar}^{(l)}||_2$ over the number of iterations l is shown in Figure 5.2 for different values of step size Δ. We see that after very few iterations, the error in the access ratios is negligibly small.

The reason of the almost linear error convergence with relation with the number of iterations is due to the fact that we are optimising over a hyperplane ($\sum_i ar_i = 1$). Thus, regardless of the distance or direction of the operating point from the desired point, the concavity of the surface is similar, and is actually flat (no concavity).

5.5.2 Simulation of the real-time adaptation approach

Another simulation was conducted for the real-time approach. Here, it was assumed that there were four users with relative average channel qualities:

$$\bar{\mathbf{h}} = [1 \;\; 1 \;\; 0.6 \;\; 0.2]$$

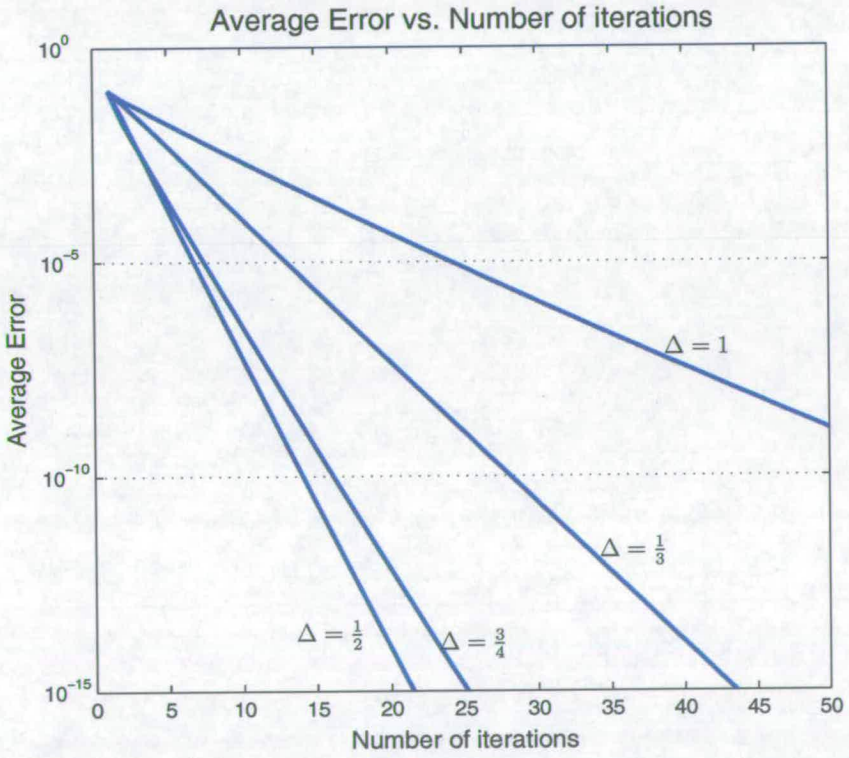


Figure 5.2: The error convergence over the iterations of the off-line algorithm to calaculate the weighting factors of the scheduling policy

The required access ratios $\mathbf{ar}$ were assumed to be:

$$\mathbf{ar} = [0.2 \;\; 0.15 \;\; 0.25 \;\; 0.4]$$

It was assumed that after some time the forth user has a better channel ($\bar{h}_4$ changes from 0.2 into 0.6). Figure 5.3 shows how the system updates itself to maintain the required channel access ratios. Again it is clear from the figure that the proposed algorithm provides quick convergence to overcome the effect of changing average channel qualities of a user. For example, when the average quality of the channel of user 4 improved, it was just for a short period that user 4 got higher channel access than required while the other users got lower channel access than their assigned ratios. However, with the application of the "real-time" algorithm, the system was able to update the weighting factors and the scheduler was able to maintain the required access ratios of the users again.

In Figure 5.3, the time scale is dependent on the parameters t_c, Δ. There is a trade-off between how fast the system response to a change in channel conditions and the fluctuations in the system's response (i.e. fluctuations in the weighting factors and in the measured access ratios in a previous window). As discussed in Chapter 4, the fluctuations in the weighting factors decreases the long term average rates. However, in order to prevent big fluctuations, t_c should have a large value. This will result in longer time scale until a system can respond to a change in channels' conditions as the system is averaging over a window defined by t_c and hence it does not detect the change directly. Thus, a compromise between rate gains and system response time is needed. Furthermore, the value of Δ does not affect the fast fluctuations in the response but rather the response pattern of the system (i.e. the ripples (peaks) in the system response until it reaches a steady state point).

5.6 Multiuser Diversity Gain Analysis

One of the approaches that is used in literature to analyse the performance of schedulers is by evaluating the multiuser diversity gains achieved by the schedulers (i.e. the analysis of the system's throughput gain over a Round-Robin system in which no channel-state information is used in the selection strategy). In [20], the multiuser diversity gain was defined as the ratio between the rates achieved by the channel-aware scheduler to the rates achieved by a Round-Robin policy. In this work, we would use a different approach to define multiuser diversity gain:

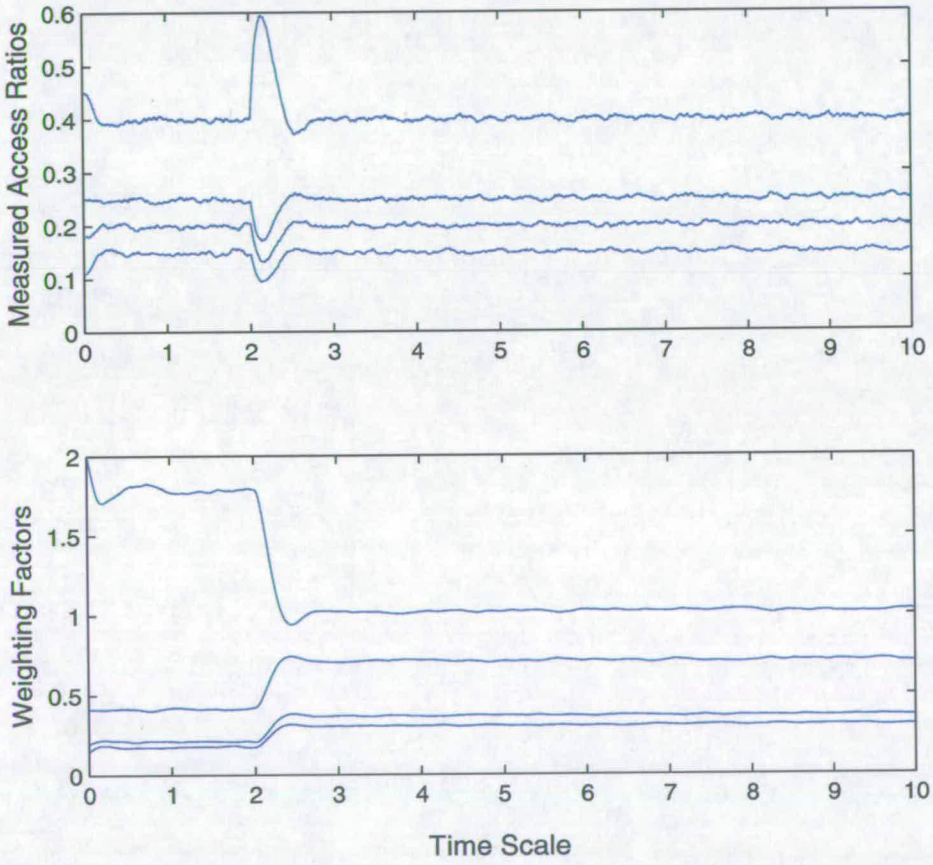


Figure 5.3: Performance of the real-time algorithm to adapt the weighting factors of the scheduling policy

the ratio between the long-term average channel quality of the channel instances over which the user is scheduled to transmit (receive) $\widetilde{h}$ to the actual long-term average channel quality of the user $\bar{h}$.

$$gain \equiv \frac{E\left[\widetilde{h}\right]}{E\left[h\right]} \tag{5.11}$$

Using the suggested scheduler, $\widetilde{h}$ has the distribution:

$$\widetilde{f_{H_i}}(x) = \frac{\prod_{j \neq i} F_{H_j}\left(\frac{\mu_i}{\mu_j} x\right)}{ar_i} f_{H_i}(x) \tag{5.12}$$

Refer to Appendix C.1 for a proof.

The reason behind selecting this definition for the multiuser diversity gain and not applying the definition in [20] is that the rate ratios depends on the system SNR (Signal to Noise Ratio). Thus, applying the approach in [20] will give different measures for the multiuser diversity gain dependent on the average SNR. However, using the selected approach (i.e. based on channel qualities), the value of the multiuser diversity gain will depend on the channel fading distribution (i.e. Rayleigh, Rice, etc..) and not on the SNR value.

There are three factors affecting the multiuser diversity gain:

- The number of active users in the system: as the number of active users in the network increases, higher multiuser diversity gains are obtained. This is because the probability of transmitting over very good channel conditions increases since the users' channels are assumed to be independent.
- The channel statistics: as the variation in the channel quality increases, higher multiuser diversity gains can be obtained. For example, The gains with Rayleigh fading is higher than with Rice fading. When the variation of the channel quality is high, it becomes more probable that when the user is scheduled, he has very good channel condition relative to the average channel quality of his link, and hence better multiuser diversity gain can be obtained.
- The channel access ratio of a user: for a given number of active users, if a user accesses the network less, he gets higher multiuser diversity gains. This is because it becomes more possible to schedule the user at the very best channel conditions of his link.

As an example on the relation of the multiuser diversity gain and the number of active users, Figure 5.4 shows the probability distribution of the normalised channel qualities over which a user is scheduled ($\widetilde{h}$) for the case of single, two, three and four users. This is for the case of equal resource-share fairness and Rayleigh faded channels. The multiuser diversity gains can be obtained directly from Figure 5.4 because the actual channels' qualities are assumed to be normalised (i.e. $\bar{h} = 1$), and hence the gain equals $E\left[\widetilde{h}\right]$ which is shown in the figure.

In the case of all users having Rayleigh fading, the multiuser diversity gain can be expressed as a function of the number of users (k):

$$gain(k) = \sum_{i=1}^{k} -\frac{(-1)^i}{i}\binom{k}{i} = \sum_{i=1}^{k}\frac{1}{i} \tag{5.13}$$

Refer to Appendix C.3 for a proof of this formula.

Figure 5.5 shows the dependence of the multiuser diversity gain on the channel statistics. It is demonstrated that higher multiuser diversity gains can be obtained when the channel variations are larger (such as in Rayleigh fading conditions).

To give an example on the relation between multiuser diversity gain and the channel access ratio of a user, Figure 5.6 shows the probability distribution of the normalised channel qualities over which a user is scheduled ($\widetilde{h}$) for the case of two users with different channel access ratios. The gain can be expressed in this case as (this is under Rayleigh fading assumption):

$$gain_i = 2 - ar_i \tag{5.14}$$

Refer to Appendix C.4 for a proof of this formula.

It is demonstrated in Figure 5.6 that higher multiuser diversity gains are obtained when the user accesses the channel less.

5.7 Conclusions

In this chapter, the schemes to adapt the control parameters of the optimal scheduling policies in order to achieve fairness constraints are considered. Since achieving throughput fairness in wireless networks may lead to users of bad channels occupying most of the air-link resources

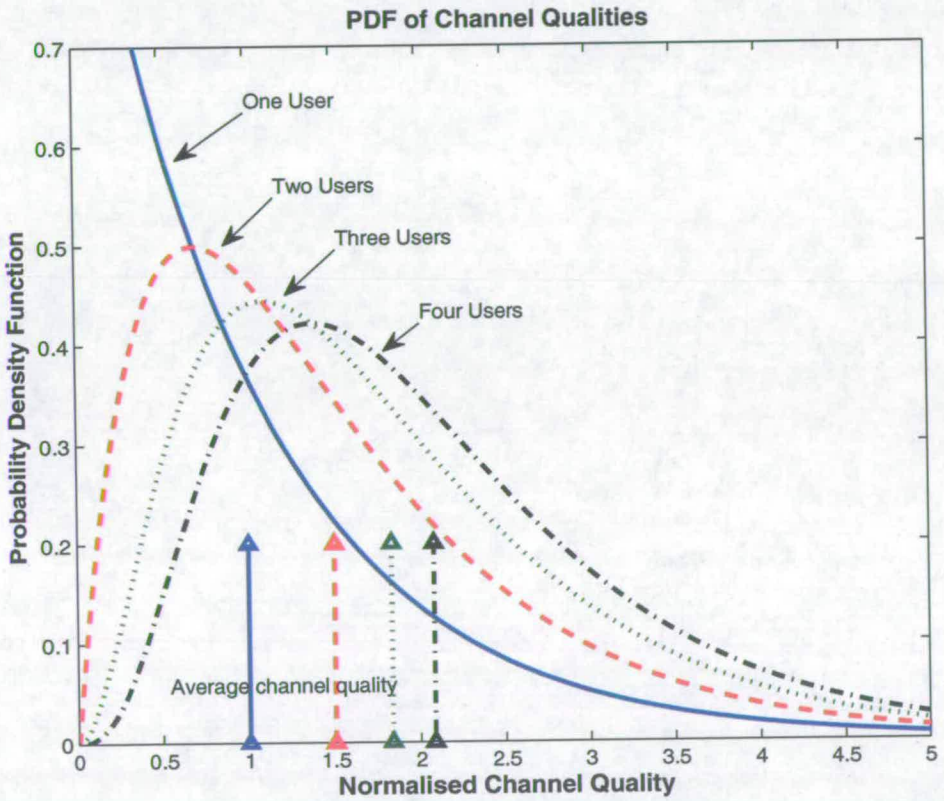


Figure 5.4: Probability density function of the normalised channel qualities over which a user is scheduled ($\widetilde{h}$) using equal resource-share fair scheduler (Rayleigh fading channels)

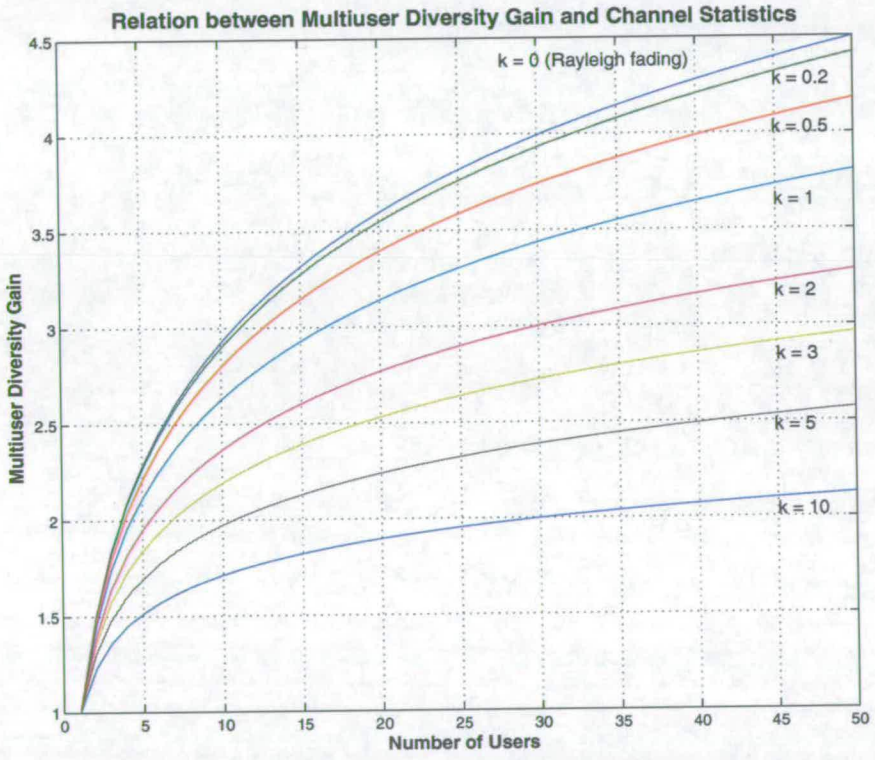


Figure 5.5: Multiuser diversity gain as a function of the number of users for the equal resource-share fairness scheduler for different type of channel statistics: Rayleigh and Rice with different values of κ.

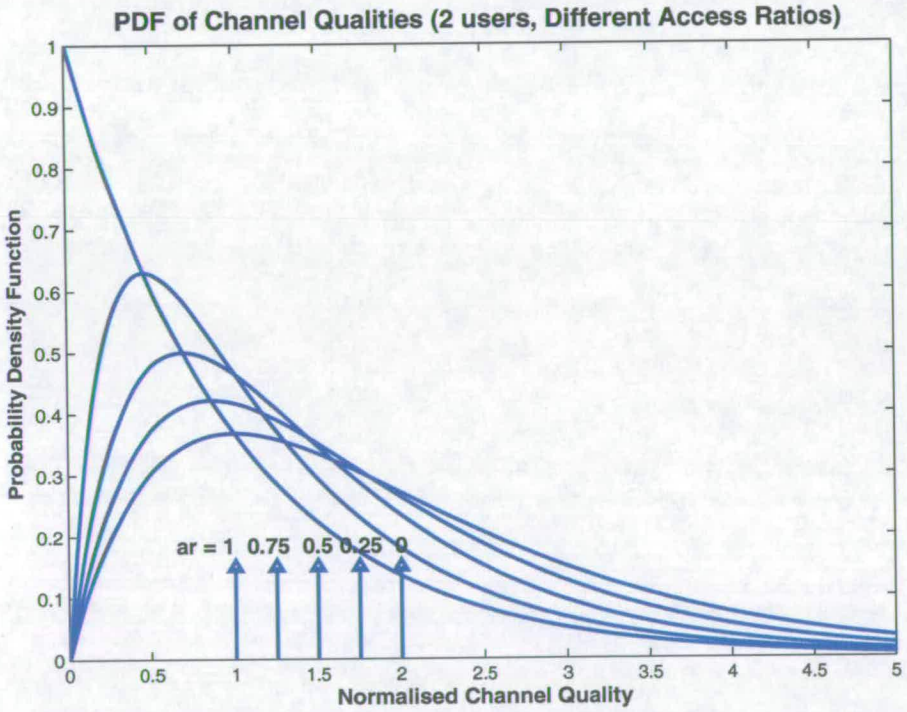


Figure 5.6: Probability density function of normalised channel qualities for two users case with different channel access ratios (Rayleigh fading channels)

and to severe degradation in the achievable total throughput of the system, resource-sharing fairness is considered instead. By applying resource-sharing fairness constraints, every user will have throughput according to his average channel quality without affecting the service of other users. Two new schemes are suggested to adapt the weighting factors in a generic channel-aware scheduler which allows one to meet pre-specified channel access ratios for the users in the system. The method has proved to be efficient, and the adaptive scheme is able to track channel variations. All schemes can be efficiently implemented and lend themselves to use in practical schedulers, e.g., in a base station or a wireless access point. The multiuser diversity gains of channel-aware scheduling schemes (relative to Round-Robin schemes) are analysed as well to study the relation between the multiuser diversity gains and the number of active users in the system, the fading statistics of the channels and the channel access ratios of the users.

Chapter 6
Conclusions and Recommendations

This chapter summarises the main results of this work by highlighting the important guidelines and key messages which are needed for the design of efficient and flexible multiuser scheduling schemes for centralised wireless networks. Furthermore, some research areas are suggested to follow on this important topic.

6.1 Conclusions

The main results and contributions of this work have covered fundamental topics related to the important research area of designing efficient schedulers for centralised wireless networks that can be flexibly controlled by the network operator. Although the example analysis of the two-user case, which is used in many parts of the thesis, does not provide an exhaustive numerical evaluation of the scheduling policies considered, it highlights important results which, qualitatively, carry over to the general case. A summary of the main results is provided below.

6.1.1 Optimality of schedulers

- Channel-aware scheduling is needed to achieve considerable performance gains in wireless networks and to achieve the capacity limits.
- Many trade-offs exist in the multiuser communication over wireless fading channels between the system capacity (throughput), and the quality-of-service of the served applications, the fairness between the users, the cost of the capacity in terms of physical resources and the system complexity.
- The best definition for the optimality of a scheduling scheme is not to maximise the total throughput of the system, but rather to operate at Pareto-optimal conditions in which no user can have higher rate without decreasing the rates of other users or increasing the amount of physical resources.

- There is no contradiction between efficient resource allocation and achieving fairness and QoS requirements, taking into consideration that the efficiency of a scheduler is by operating at Pareto-optimal conditions.
- A good scheduler uses a policy that enables operation at a Pareto-optimal point. This also means that the scheduler should operate at the boundary of the capacity region which is the set of long-term average achievable rates of the users for a given amount of physical resources.
- Maximising any utility function or maintaining fairness criteria should be done such that a suitable operating point is chosen on the capacity region's boundary.
- The optimisation problem (to find the optimal resource allocation schemes in order to operate at the capacity region's boundary) is to maximise a weighted sum of the long-term average rates of the users under the main constraint on the long-term average transmitted power as well as possible additional constraints based on the systems' capabilities.

6.1.2 Examples of optimal scheduling policies

- The optimal scheduling policy depends on the system constraints such as power control, but does not depend on the fading statistics of the users' channels.
- In order to obtain maximum possible multiuser diversity gains, the multiuser diversity should be exploited in the frequency domain in addition to the time domain.
- The optimal resource allocation scheme over flat-faded channel blocks involves applying CDMA (in uplink) or SC (in downlink) with SIC at the receivers. The number of users scheduled in a channel block varies depending on the channel conditions.
- The optimal power allocation scheme is given by the water-filling approach: more power is allocated when the channel is better, and based on the operating point of the system some users get higher average power.
- The optimal solutions (SC with SIC, and water-filling power control) are difficult if not impractical to implement.
- With a single constraint on the total power transmitted, the optimal resource allocation schemes for the uplink and the downlink are exactly identical when orthogonal signalling is used.

- Under orthogonal signalling constraints, the optimal scheduling policy is (3.33) with water-filling power control.
- Under orthogonal signalling and constant power constraints, either one user is scheduled or a maximum of two users share the channel. However, most of the time one user only is scheduled. Thus, if a single user selection per channel block is applied instead of orthogonal signalling, the performance of the system remains almost the same.
- For constant power and single user selection per block, the best scheduling policy is (3.42).
- Power control is useful at low spectral efficiency. Applying constant power constraint is justified if the system is operating at high spectral efficiencies.
- Superposition coding with successive interference cancellation at the receivers can be useful in adaptive power systems, but is not needed in constant power systems.

6.1.3 Comparison of known scheduling policies

- Some scheduling policies are good for some operating points only, as they are not generic. There are policies which can be used in generic schedulers, such as (4.3), because they have close-to-optimal performance for all operating points on the capacity boundary.
- Dynamic "on-line" variation of the weighting factors degrades the performance and will be avoided by a good scheduler if the applications are delay tolerant. However, in case of delay sensitive applications, some real-time adjustment of the control parameters is needed. This is known as the trade-off between throughput and delay constraints.
- If the system traffic load is dynamic, the weighting factors should be updated according to the changing conditions of the system.
- Achieving any kind of fairness between users or maximising any performance metric of the network should be done by properly adjusting the control parameters of the optimal scheduling policy, and not by using a different scheduling policy.

6.1.4 Generic flexible schedulers

- Achieving throughput fairness in wireless networks may lead to users of bad channels occupying most of the air-link resources and to severe degradation in the achievable total throughput of the system. Alternatively, by achieving resource-sharing fairness, every user will have throughput according to his average channel quality without affecting the service of other users.

- Two new schemes are suggested to adapt the weighting factors in a generic channel-aware scheduler which allows one to meet pre-specified channel access ratios for the users in the system. The method has proved to be efficient, and the adaptive scheme is able to track channel variations. All schemes can be efficiently implemented and lend themselves to use in practical schedulers, e.g., in a base station or a wireless access point.

- The multiuser diversity gains of channel-aware scheduling schemes (relative to Round-Robin schemes) depend on the number of active users in the system, the fading statistics of the channels and the channel access ratios of the users.

6.2 Suggested Research Areas

Examples of possible research areas to follow on the work done in this thesis include:

- To investigate the optimal resource allocation schemes for more practical system's considerations such as in MIMO systems or in systems using relays. To the best of my knowledge, the research on these topics has concentrated on point-to-point communication or on maximising the sum-throughput in multiuser operation. I suggest to investigate the optimal resource allocation schemes to achieve any of the Pareto-optimal points in these systems. In [76] a pioneering work in characterising the capacity region for MIMO BC channels was presented.

- To characterise and quantise the different trade-offs existing in wireless networks, and to study the rational and good solutions for the contradicting requirements of the system. I believe that more research is needed to design good admission controllers for wireless networks.

- To study the multiuser scheduling task taking into account the possible inaccuracy in the

channel measurement and estimation as well as the measurement delay.

- To study the Pareto-optimal scheduling schemes with strict delay constraints of the served applications, and for a mix of delay-tolerant and strict-delay applications.

Appendix A
Proofs for Chapter 3

A.1 Proofs for the Optimal Resource Allocation and Boundary Characterisations

A.1.1 Resource allocation

The optimal resource allocation can be obtained by solving the optimisation problems defined in (3.14) such that the value of λ is selected so that the condition (3.15) is maintained. Let's assume now that the value of λ is correctly selected. Later on, when the proofs for the equations to characterise the boundary of the capacity region are provided, the way to select λ by applying (3.23), (3.24) or (3.25) will become clear.

The optimal power allocation $P[k]$ for the problem in (3.14) can be obtained by the conventional way well-known in calculus by differentiating with respect to the power $P[k]$[1]. Since the relation between the rate and resources (power) is a concave function (logarithmic), there is a unique solution (i.e. value of $P[k]$) to the problem:

$$\frac{d}{dP[k]}\left[\max\left(\sum_{i=1}^{M}\mu_i R_i[k]\right) - \lambda P[k]\right] = 0 \tag{A.1}$$

This is equivalent to:

$$\frac{d}{dP[k]}\left[\max\left(\sum_{i=1}^{M}\mu_i R_i[k]\right)\right] - \lambda = 0 \tag{A.2}$$

$$\max_i\left(\frac{d}{dP[k]}\mu_i R_i[k]\right) - \lambda = 0 \tag{A.3}$$

[1] As will be discussed soon, we have to differentiate with respect to transmitted power $P[k]$ for BC channels, and with respect to received power $P_R[k]$ for MAC channels.

From the well-known relation between the rate and the power:

$$R[k] = \log\left(1+\frac{h[k]P[k]}{\sigma^2}\right) \tag{A.4}$$

$$= \int_0^{P[k]} \frac{1}{\frac{\sigma^2}{h[k]}+z}dz \tag{A.5}$$

$$= \int_0^{P_R[k]} \frac{1}{\sigma^2+z}dz \tag{A.6}$$

Although the relation (A.4) is valid in a single user channel and is not valid when we have multiple users sharing the channel and the total transmitted power, the derivative of the rate with respect to the power is the same in multiuser channels and can be obtained based on (A.5), (A.6):

$$\frac{dR_i[k]}{dP[k]} = \frac{1}{\frac{\sigma^2}{h_i[k]}+P[k]} \tag{A.7}$$

$$\frac{dR_i[k]}{dP_R[k]} = \frac{1}{\sigma^2+P_R[k]} \tag{A.8}$$

Equation (A.7) is valid for BC channels since the source of power is unique (the base station) although, from a user perspective, some of the received power is information and the remaining is interference. Similarly, (A.8) is valid for MAC channels since the receiver is unique although, from a user perspective, some of the received power is information and the remaining is interference.

By substituting (A.7) into (A.3), we obtain the result in (3.16). Since the power can not be negative, the notation ($x^+ = \max(x,0)$) is added to the solution in (3.16). Note that the substitution ($\dot{\lambda} = \sigma^2\lambda$) is applied. The result is identical in the MAC channels case, and can be done by differentiating (3.14) with respect to received power $P_R[k]$ and substituting (A.8) for the derivative.

The greedy algorithm procedure[2] to obtain the optimum rate allocation per user in a channel block is applying the same concept described above, but with gradually increasing the power and deciding the user to be allocated each small increment of the power until the total power is allocated. Note that with each small increment of power, the achievable rates of previous

[2]The noise variance is not appearing in (3.17) and (3.18) because we were scaling with it: ($\dot{\lambda} = \sigma^2\lambda$) and ($\dot{z} = \frac{z}{\sigma^2}$).

allocated power are not affected since every new small increment of allocated power is to be decoded first and then successively cancelled so that it does not affect the previously allocated power.

A.1.2 Boundary characterisation

Due to the duality of the BC and MAC channels, the equations that characterise the boundary of the capacity region of either of them is applicable to the other one. That is why there are equivalent forms to characterise the capacity region boundary.

A description is provided below for the structure of the proofs of the given equations to characterise the capacity region with an example to compute the achievable rates in MAC channels (equation(3.21)). The proofs of all other equations follow a similar approach. The proof of (3.23) is similar to the approach presented in Chapter 4.

The achievable rates in MAC channels can be computed as:

$$R_i = \int_0^\infty \frac{dR_i}{dz} P(i,z) dz \tag{A.9}$$

with

$$P(i,z) \doteq \Pr\Big(u_i(z) > u_j(z) \quad \forall j \neq i \quad \text{and} \quad u_i(z) > 0\Big) \tag{A.10}$$

where the marginal utilities are defined by (3.18), and the derivative equals:

$$\frac{dR_i}{dz} = \frac{1}{1+z} \tag{A.11}$$

To solve (A.9) we need to evaluate the probability (A.10). Using (3.18) we can state the equivalence

$$u_i(z) > 0 \iff h_i > \frac{\lambda(1+z)}{\mu_i} > 0\,. \tag{A.12}$$

Note that $\lambda > 0$, as λ is a Lagrange multiplier that introduces the "power price" (that can never be negative).

Using (A.12), the probability (A.10) can now be written as

$$P(i,z) = \Pr\left(u_i(z) > 0 \mid u_i(z) > u_j(z)\,\forall j\right) \cdot \Pr\left(u_i(z) > u_j(z)\,\forall j\right) \quad \text{(A.13)}$$

$$= \Pr\left(h_i > \frac{\lambda(1+z)}{\mu_i} \mid u_i(z) > u_j(z)\,\forall j\right) \cdot \Pr\left(u_i(z) > u_j(z)\,\forall j\right) \quad \text{(A.14)}$$

$$= \int_{\frac{\lambda(1+z)}{\mu_i}}^{\infty} f_{H_i}\left(x \mid u_i(z) > u_j(z)\,\forall j\right)dx \cdot \Pr\left(u_i(z) > u_j(z)\,\forall j\right) \quad \text{(A.15)}$$

$$= \int_{\frac{\lambda(1+z)}{\mu_i}}^{\infty} f_{H_i}\left(x, u_i(z) > u_j(z)\,\forall j\right)dx \quad \text{(A.16)}$$

$$= \int_{\frac{\lambda(1+z)}{\mu_i}}^{\infty} f_{H_i}(x) \cdot \Pr\left(u_i(z) > u_j(z)\,\forall j \mid h_i = x\right)dx \quad \text{(A.17)}$$

Since the fading processes of the users are assumed to be independent, we can write:

$$P(i,z) = \int_{\frac{\lambda(1+z)}{\mu_i}}^{\infty} f_{H_i}(x) \cdot \prod_{j \neq i} \Pr\left(u_i(z) > u_j(z) \mid h_i = x\right) dx\,. \quad \text{(A.18)}$$

Now, we need to evaluate the probability

$$\Pr\left(u_i(z) > u_j(z) \mid h_i = x\right) \quad \text{(A.19)}$$

By using (3.18), the event $u_i(z) > u_j(z)$ can be written as:

$$u_i(z) > u_j(z) \iff \frac{\mu_i}{a} - \frac{\lambda}{h_i} > \frac{\mu_j}{a} - \frac{\lambda}{h_j} \quad \text{(A.20)}$$

or, equivalently,

$$\frac{h_i(\mu_j - \mu_i) + \lambda a}{a\lambda h_i} < \frac{1}{h_j} \quad \text{(A.21)}$$

with the abbreviation $a \doteq (1+z) > 0$ and $\lambda > 0$ and $0 < \mu_i \leq 1\ \forall i$. As $\mu_j - \mu_i$ can be negative, the left-hand side of (A.21) can be negative so we have to differentiate between two

cases:

$$\text{Case A}: \quad h_i(\mu_j - \mu_i) + \lambda a > 0 \iff (\mu_j \geq \mu_i) \text{ or } \left(\mu_j < \mu_i \text{ and } h_i < \tfrac{\lambda a}{\mu_i - \mu_j}\right) \quad \text{(A.22)}$$

$$\text{Case B}: \quad h_i(\mu_j - \mu_i) + \lambda a < 0 \iff \mu_j < \mu_i \quad \text{and} \quad h_i > \tfrac{\lambda a}{\mu_i - \mu_j} \quad \text{(A.23)}$$

Case A With $a = (1 + z)$ we obtain from (A.19), (A.21) and (A.22)

$$\begin{aligned} \Pr\left(u_i(z) > u_j(z) \mid h_i = x\right) &= \Pr\left(h_j < \frac{\lambda h_i(1+z)}{\lambda(1+z) + (\mu_j - \mu_i)h_i} \middle| h_i = x\right) && \text{(A.24)} \\ &= \Pr\left(h_j < \frac{\lambda x(1+z)}{\lambda(1+z) + (\mu_j - \mu_i)x}\right) && \text{(A.25)} \\ &= F_{H_j}\left(\frac{\lambda}{\frac{\lambda}{x} + \frac{\mu_j - \mu_i}{\lambda}}\right) && \text{(A.26)} \end{aligned}$$

with $F_{H_j}(x)$ the cumulative density function of the channel j.

Case B For a negative left-hand side in (A.21) we obtain

$$\Pr\left(u_i(z) > u_j(z) \mid h_i = x\right) = \Pr\left(h_j > B\right) = 1 \quad \text{(A.27)}$$

with

$$B \doteq \frac{\lambda x(1+z)}{\lambda(1+z) + (\mu_j - \mu_i)x} < 0\,. \quad \text{(A.28)}$$

As h_j is a channel quality (power gain) and non-negative by definition, the probability (A.27) is simply "one".

Formulation of the boundary of the capacity region We write the probability

$$\Pr\left(u_i(z) > u_j(z) \mid h_i = x\right) = F_{H_j}\left(\left[\frac{\lambda}{\frac{\lambda}{x} + \frac{\mu_j - \mu_i}{\lambda}}\right]^*\right) \quad \text{(A.29)}$$

with the function $[x]^*$ defined in (3.29). When we use (A.29) in (A.18) and (A.9) we obtain the solution (3.21).

A.2 Proofs for Section 3.9.2

In Section 3.9.2, the channels are assumed to be symmetric with all users' channels having identical Rayleigh fading conditions. Furthermore, the system is assumed to be operating at the maximum sum-throughput point. Due to the symmetry of the channels in this particular example, this point is achieved when the weighting factors μ_i of all users is identical. We can select that $\mu_i = 1$ for all users.

Furthermore, due to the symmetry of the channels and having identical weighting factors, all users will have identical rates, i.e. $R_{sum} = MR_i$, where M is the total number of users in the system.

Thus, we get the following expression for R_{sum} for the constant power system, by applying (3.43) and taking into consideration that the weighting factors in (3.44) are all identical:

$$R_{sum} = M \int_0^\infty \frac{1}{\bar{h}} \exp\left(\frac{-x}{\bar{h}}\right) \left[1 - \exp\left(\frac{-x}{\bar{h}}\right)\right]^{M-1} \log\left(1 + x\frac{\bar{P}}{\sigma^2}\right) dx \qquad \text{(A.30)}$$

Since the maximum sum throughput point is obtained when all weighting factors are identical, the optimal resource allocation will not include superposition coding and hence not more than a single user will be scheduled in a channel block in both cases of constant power and adaptive power systems. That is why (3.43) is used for constant power system and (3.34) for adaptive power system.

Applying the well-known Binomial series:

$$(1-x)^N = \sum_{n=0}^{N} (-1)^n x^n \binom{N}{n} \qquad \text{(A.31)}$$

we obtain:

$$\left[1 - \exp\left(\frac{-x}{\bar{h}}\right)\right]^{M-1} = \sum_{n=0}^{M-1} (-1)^n \binom{M-1}{n} \exp\left(-n\frac{x}{\bar{h}}\right) \qquad \text{(A.32)}$$

$$\exp\left(\frac{-x}{\bar{h}}\right)\left[1 - \exp\left(\frac{-x}{\bar{h}}\right)\right]^{M-1} = \sum_{i=1}^{M} (-1)^{(i-1)} \binom{M-1}{i-1} \exp\left(-i\frac{x}{\bar{h}}\right) \qquad \text{(A.33)}$$

where the substitution $i = n + 1$ is used.

Substituting (A.33) into (A.30), we get:

$$R_{sum} = \frac{M}{\ln 2}\sum_{i=1}^{M}(-1)^{(i-1)}\binom{M-1}{i-1}\int_0^{\infty}\frac{1}{\bar{h}}\exp\left(-i\frac{x}{\bar{h}}\right)\ln\left(1+x\frac{\bar{P}}{\sigma^2}\right)dx \quad \text{(A.34)}$$

$$= \frac{M}{\ln 2}\sum_{i=1}^{M}(-1)^{(i-1)}\binom{M-1}{i-1}\frac{1}{i}\exp\left(\frac{i\sigma^2}{\bar{h}\bar{P}}\right)E_1\left(\frac{i\sigma^2}{\bar{h}\bar{P}}\right) \quad \text{(A.35)}$$

where E_1 is the exponential integral function:

$$E_1(x) \equiv \int_x^{\infty}\frac{\exp(-u)}{u}du \quad \text{(A.36)}$$

In (A.35), the integration rule (A.37) is applied:

$$\int_0^{\infty}\frac{1}{\bar{h}}\exp\left(-i\frac{x}{\bar{h}}\right)\ln\left(1+x\frac{\bar{P}}{\sigma^2}\right)dx = \frac{1}{i}\exp\left(\frac{i\sigma^2}{\bar{h}\bar{P}}\right)E_1\left(\frac{i\sigma^2}{\bar{h}\bar{P}}\right) \quad \text{(A.37)}$$

Equation (A.35) can be further simplified by applying:

$$M\binom{M-1}{i-1}\frac{1}{i} = \binom{M}{i} \quad \text{(A.38)}$$

By substituting (A.38) into (A.35) we get (3.45).

Applying the same approach, we can obtain the results for the optimal power control system. By applying (3.34) and substituting $\mu_i = 1$ for all users in (3.36), we get:

$$R_{sum} = M\int_{\lambda}^{\infty}\frac{1}{\bar{h}}\exp\left(\frac{-x}{\bar{h}}\right)\left[1-\exp\left(\frac{-x}{\bar{h}}\right)\right]^{M-1}\log\left(\frac{x}{\lambda}\right)dx \quad \text{(A.39)}$$

where λ in (A.39) can be computed by applying (3.35). We obtain:

$$\frac{\bar{P}}{\sigma^2} = M\int_{\lambda}^{\infty}\frac{1}{\bar{h}}\exp\left(\frac{-x}{\bar{h}}\right)\left[1-\exp\left(\frac{-x}{\bar{h}}\right)\right]^{M-1}\left[\frac{1}{\lambda}-\frac{1}{x}\right]dx \quad \text{(A.40)}$$

The following two integrations will be required to solve (A.39) and (A.40) respectively:

$$\int_{\lambda}^{\infty}\frac{1}{\bar{h}}\exp\left(-i\frac{x}{\bar{h}}\right)\ln\left(\frac{x}{\lambda}\right)dx = \frac{1}{i}E_1\left(\frac{i\lambda}{\bar{h}}\right) \quad \text{(A.41)}$$

$$\int_{\lambda}^{\infty} \frac{1}{\bar{h}} \exp\left(-i\frac{x}{\bar{h}}\right)\left[\frac{1}{\lambda}-\frac{1}{x}\right]dx = \frac{1}{i\lambda}\exp\left(\frac{-i\lambda}{\bar{h}}\right) - \frac{1}{\bar{h}}E_1\left(\frac{i\lambda}{\bar{h}}\right) \tag{A.42}$$

After solving (A.39) and (A.40), the results in (3.46) and (3.47) will be obtained. Note that the substitution $\dot{\lambda} = \frac{\lambda}{\bar{h}}$ is used.

Appendix B
Proofs for Chapter 4

B.1 Derivation of the Function $g_{ij}(h_i[k])$ for Scheduling Policy (3.33)

For (3.31) we obtain $y_i(h_i) = \mu_i R_i - \lambda P_i(h_i)/\sigma^2 = \mu_i \log\left(1 + h_i P_i(h_i)/\sigma^2\right) - \lambda P_i(h_i)/\sigma^2$, with $P_i(h_i)$ the power that would be allocated to user i in channel state h_i. From the power allocation (4.6) we know that the user i is (if at all) only scheduled, when the necessary condition $\frac{\mu_i}{\lambda} > \frac{1}{h_i}$ is fulfilled; otherwise, the allocated power will be "zero" and, hence, we set $g_{ij}(h_i) = 0$ for $h_i \leq \frac{\lambda}{\mu_i}$. For $h_i > \frac{\lambda}{\mu_i}$ we obtain

$$\mu_j \log\left(1 + \frac{h_j P_j(h_j)}{\sigma^2}\right) - \lambda \frac{P_j(h_j)}{\sigma^2} < \mu_i \log\left(1 + \frac{h_i P_i(h_i)}{\sigma^2}\right) - \lambda \frac{P_i(h_i)}{\sigma^2} \tag{B.1}$$

With the non-zero solution of the power allocation policy (4.6) inserted on both sides we find

$$\mu_j \log\left(h_j \frac{\mu_j}{\lambda}\right) - \mu_j + \frac{\lambda}{h_j} < \mu_i \log\left(h_i \frac{\mu_i}{\lambda}\right) - \mu_i + \frac{\lambda}{h_i} \quad \Leftrightarrow \tag{B.2}$$

$$\log\left(\frac{e^{x_j}}{x_j}\right) < \frac{\mu_i}{\mu_j}\left(-\log(x_i) + x_i\right) + 1 - \frac{\mu_i}{\mu_j} \tag{B.3}$$

with the abbreviation $x_m \doteq \frac{\lambda}{\mu_m h_m}$. If user j's channel gain does not satisfy $h_j > \lambda/\mu_j$ then (4.6) will allocate a power of zero, i.e., this user is not considered for scheduling and we can remove them from the list of candidates with which we compare user i's channel gain. By inverting the log-function we obtain:

$$\frac{e^{x_j}}{x_j} < \exp\left(\frac{\mu_i}{\mu_j} x_i\right) (x_i)^{-\mu_i/\mu_j} \exp\left(1 - \frac{\mu_i}{\mu_j}\right) \quad \Leftrightarrow \tag{B.4}$$

$$\frac{e^{x_j}}{x_j} < \exp\left(\frac{\lambda}{\mu_j h_i} + 1 - \frac{\mu_i}{\mu_j}\right) \left(\frac{\lambda}{\mu_i h_i}\right)^{-\mu_i/\mu_j} \doteq A \tag{B.5}$$

As $e^{x_j}/x_j = -\frac{1}{-x_j e^{-x_j}} = -1/q(-x_j)$ with $q(z) \doteq ze^z$, we can rewrite (B.5) according to

$$\frac{e^{x_j}}{x_j} = -\frac{1}{q(-x_j)} < A \quad \Leftrightarrow \quad q(-x_j) < -\frac{1}{A} \tag{B.6}$$

because $A > 0$ (as $\lambda, h_j, \mu_i > 0$) and $q(-x_j) = -x_j e^{-x_j} < 0$ as $x_j = \frac{\lambda}{\mu_j h_j} > 0$.

Let us briefly consider the function $y = q(w) \doteq we^w$: its inverse $w = W(y) = q^{-1}(y)$ is known as the Lambert W-function [60], and the inversion is unique for $w \in (-\frac{1}{e}, +\infty)$. For $y < 0$ we obtain $W(y) < 0$. Considering that $h_i > \lambda/\mu_i$, an analysis of the extreme values of $-\frac{1}{A}$ reveals that $-\frac{1}{A} \in (-\frac{1}{e}, 0)$, so the inversion is indeed unique and we have $-1 < W(-1/A) < 0$. Therefore, we obtain from (B.6) and (B.5): $-x_j = -\frac{\lambda}{\mu_j h_j} < W(-\frac{1}{A})$. As $\lambda, \mu_j, h_j > \frac{\lambda}{\mu_j} > 0$, we finally obtain the result in Table 4.1 (last row):

$$h_j < \frac{-\lambda}{\mu_j W\Big(-\exp\left(\frac{\mu_i}{\mu_j} - \frac{\lambda}{\mu_j h_i} - 1\right) \left(\frac{\lambda}{\mu_i h_i}\right)^{\mu_i/\mu_j} \Big)} \doteq g_{ij}(h_i) \tag{B.7}$$

Appendix C
Proofs for Chapter 5

C.1 Mathematical Derivation of Equations (5.2) and (5.12)

User i is scheduled to transmit in block n if $i = \arg\max_m \mu_m h_m[n]$. Thus, user i does not transmit in all blocks but rather in a ratio of ar_i of the total number of blocks:

$$ar_i \equiv \Pr\{i = \arg\max_m \mu_m h_m\} \tag{C.1}$$

With the assumption that the users channels are fading independently, the joint probability density function of the channel vector $\mathbf{h}$ is:

$$f(\mathbf{h}) = \prod_{m=1}^{M} f_{H_m}(h_m)$$

To illustrate how to evaluate the access ratios of the users, we start with a two-user example (refer to Figure C.1): ar_1 in this case, is the shaded region in the figure. This region satisfies:

$$\mu_1 h_1 > \mu_2 h_2$$

ar_1 can be evaluated by integration over this region:

$$\begin{aligned} ar_1 &= \int_0^{\infty} \left(\int_0^{h_2=\frac{\mu_1 h_1}{\mu_2}} f(\mathbf{h}) dh_2 \right) dh_1 \\ &= \int_0^{\infty} f_{H_1}(h_1) \left(\int_0^{\frac{\mu_1 h_1}{\mu_2}} f_{H_2}(h_2) dh_2 \right) dh_1 \\ &= \int_0^{\infty} f_{H_1}(h_1) F_{H_2}(\frac{\mu_1 h_1}{\mu_2}) dh_1 \end{aligned}$$

This can be generalised for any user and for any number M of active users:

$$ar_i = \int_0^\infty f_{H_i}(x) \prod_{j \neq i} F_{H_j}\left(\frac{\mu_i}{\mu_j} x\right) dx \tag{C.2}$$

The cumulative distribution function $\widetilde{F_i}(x)$ is defined as:

$$\begin{aligned} \widetilde{F_{H_i}}(x) &= \Pr\{h_i \leq x | i = \arg\max_m \mu_m h_m\} \\ &= \frac{\Pr\{h_i \leq x,\ i = \arg\max_m \mu_m h_m\}}{\Pr\{i = \arg\max_m \mu_m h_m\}} \\ &= \frac{\Pr\{h_i \leq x,\ i = \arg\max_m \mu_m h_m\}}{ar_i} \end{aligned}$$

Again refer to Figure C.1 for the two-user case. The shaded region with pattern is the region where:

$$h_1 \leq x \bigcap \mu_1 h_1 > \mu_2 h_2$$

In this case, we obtain:

$$\begin{aligned} \widetilde{F_{H_1}}(x) &= \frac{\int_0^x \left(\int_0^{\frac{\mu_1 h_1}{\mu_2}} f(\mathbf{h}) dh_2 \right) dh_1}{ar_1} \\ &= \frac{\int_0^x f_{H_1}(h_1) F_{H_2}\left(\frac{\mu_1 h_1}{\mu_2}\right) dh_1}{ar_1} \end{aligned}$$

The probability density function $\widetilde{f_1}(x)$ can be obtained from $\widetilde{F_{H_1}}(x)$:

$$\widetilde{f_{H_1}}(x) = \frac{d\widetilde{F_{H_1}}(x)}{dx} = \frac{f_{H_1}(x) F_{H_2}\left(\frac{\mu_1}{\mu_2} x\right)}{ar_1}$$

This can be generalised for M users:

$$\widetilde{f_{H_i}}(x) = \frac{\prod_{j \neq i} F_{H_j}\left(\frac{\mu_i}{\mu_j} x\right)}{ar_i} f_{H_i}(x) \tag{C.3}$$

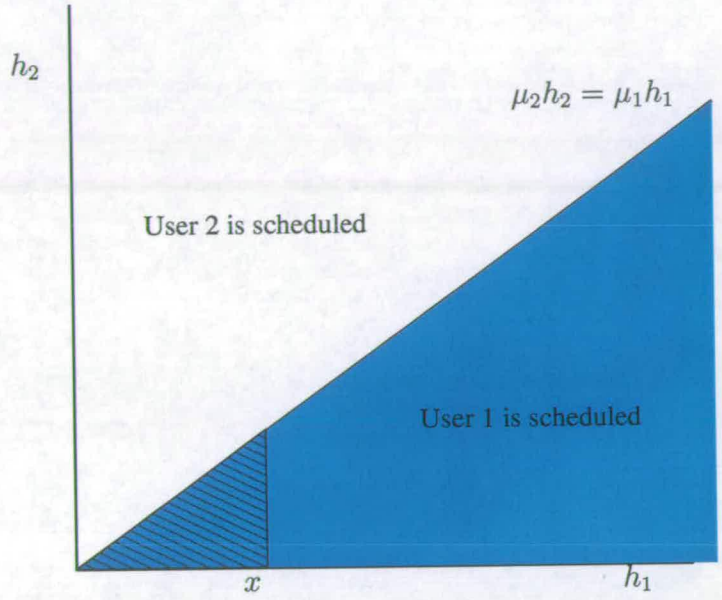


Figure C.1: Regions of channel qualities over which a user is scheduled when the scheduling policy (4.3) is applied (two-user case)

C.2 Weighting Factors to Achieve Equal Resource-Share Fairness

When all users have Rayleigh fading channels (the average values could be different), and if the weighting factors are chosen to be equal to

$$\mu_i = \frac{\alpha}{\bar{h}_i} \tag{C.4}$$

where α is any scalar, the users will access the channels equally, i.e.:

$$ar_i = \frac{1}{M} \tag{C.5}$$

with M the number of active users.

To prove this statement, refer to the formula to compute ar_i (equation (C.2)) and substitute the values of μ_i defined in equation (C.4); f_{H_i} and F_{H_i} for Rayleigh fading channels are given by (3.3) and (3.4). We obtain:

$$ar_i = \frac{1}{\bar{h}_i} \int_0^\infty \exp\left(\frac{-x}{\bar{h}_i}\right) \left[1 - \exp\left(\frac{-x}{\bar{h}_i}\right)\right]^{M-1} dx \tag{C.6}$$

By solving this integration, we obtain:

$$
\begin{aligned}
ar_i &= \frac{1}{M}\left[1-\exp\left(\frac{-\infty}{\bar{h}_i}\right)\right]^M - \frac{1}{M}\left[1-\exp\left(\frac{0}{\bar{h}_i}\right)\right]^M \\
&= \frac{1}{M}[1-0] = \frac{1}{M}
\end{aligned}
$$

Although the proof was done with the assumption of Rayleigh fading, the statement is valid for other fading distributions as long as the *normalised channel quality distributions* of the users are identical. In the following is a proof of this statement.

Let's suppose that the normalised distribution of all users channels' qualities is identical (f_n). By definition, f_n is normalised:

$$\int_0^\infty x\, f_n(x)\, dx = 1$$

Note that since f_n is the normalised distribution of channel qualities, this random variable is defined over the period $[0, \infty)$.

The probability distribution function of the channel quality of user i (h_i) which has the average value $\bar{h}_i$ is:

$$f_{H_i}(x) = \frac{1}{\bar{h}_i} f_n\left(\frac{x}{\bar{h}_i}\right) \tag{C.7}$$

Note that:

$$\int_0^\infty x\, f_{H_i}(x) dx = \int_0^\infty \frac{x}{\bar{h}_i} f_n\left(\frac{x}{\bar{h}_i}\right) dx$$

By replacement of dummy variable x:

$$x' = \frac{x}{\bar{h}_i}, \quad dx' = \frac{dx}{\bar{h}_i}$$

We obtain:

$$\int_0^\infty \frac{x}{\bar{h}_i} f_n\left(\frac{x}{\bar{h}_i}\right) dx = \bar{h}_i \int_0^\infty x' f_n\left(x'\right) dx' = \bar{h}_i$$

Furthermore, the cumulative distribution function $F_{H_i}(x)$ can be presented in terms of the normalised F_n as:

$$F_{H_i}(x') = \int_0^{x'} \frac{1}{\bar{h}_i} f_n\left(\frac{x}{\bar{h}_i}\right) dx = \int_0^{\frac{x'}{\bar{h}_i}} f_n(x) dx$$

$$F_{H_i}(x') = F_n\left(\frac{x'}{\bar{h}_i}\right) \tag{C.8}$$

Using the presentation of f_i and F_i in terms of the normalised f_n and F_n (equations (C.7) and (C.8)), and substituting the values of weighting factors defined in (C.4) into equation (C.2), we get:

$$\begin{aligned} ar_i &= \int_0^\infty \frac{1}{\bar{h}_i} f_n\left(\frac{x}{\bar{h}_i}\right)\left[F_n\left(\frac{x}{\bar{h}_i}\right)\right]^{M-1} dx \\ &= \int_0^\infty f_n\left(x'\right)\left[F_n\left(x'\right)\right]^{M-1} dx' \\ &= \frac{1}{M}\left[F_n(\infty)\right]^M - \frac{1}{M}\left[F_n(0)\right]^M \\ &= \frac{1}{M}(1-0) = \frac{1}{M} \end{aligned}$$

This concludes the proof.

If it is required to make the users access the channel equally but they have different channel distributions, then it is still possible to achieve this point by applying the adaptive design suggested in Section 5.4.

Note that the statement in this section is true when applying the scheduling policy (4.3). However, if the scheduling policy which was suggested in [20] (maximum weighted feasible rate) is applied with weighting factors achieving proportional fairness, then the users will not have identical access ratios if their average channel qualities are different. This is because, although the normalised channel qualities distributions of the users are identical, the normalised rate distributions are different according to average channel qualities.

C.3 Analysis of the Multiuser Diversity Gains for the Case of Equal Resource-Share Fairness and Rayleigh Fading Channels

The multiuser diversity gain is defined to be the ratio between the average channel quality of the channel instances over which the user is scheduled to transmit $\widetilde{h}$ to the actual average channel quality of the user.

$$gain_i \equiv \frac{E\left[\widetilde{h}_i\right]}{E\left[h_i\right]} = \frac{\int_0^\infty x\widetilde{f}_i(x)dx}{\bar{h}_i}$$

As shown in (C.3):

$$\widetilde{f}_i(x) = \frac{\prod_{j \neq i} F_j\left(\frac{\mu_i}{\mu_j} x\right)}{a r_i} f_i(x)$$

Thus, we get the following expression for $E\left[\widetilde{h}_i\right]$:

$$E\left[\widetilde{h}_i\right] = \frac{1}{a r_i \bar{h}_i} \int_0^\infty x \exp\left(\frac{-x}{\bar{h}_i}\right) \left[1 - \exp\left(\frac{-x}{\bar{h}_i}\right)\right]^{M-1} dx \tag{C.9}$$

Applying the well-known Binomial series:

$$(1-x)^N = \sum_{n=0}^{N} (-1)^n x^n \binom{N}{n}$$

and substituting equation (C.5) into equation (C.9), we get:

$$E\left[\widetilde{h}_i\right] = \frac{M}{\bar{h}_i} \sum_{n=0}^{M-1} (-1)^n \binom{M-1}{n} \int_0^\infty x \exp\left(\frac{-x}{\bar{h}_i}\right)^{(n+1)} dx \tag{C.10}$$

As well known:

$$\int_0^\infty x \exp\left(\frac{-x}{\alpha}\right)^n dx = \frac{\alpha^2}{n^2}$$

Thus, equation (C.10) can be simplified to:

$$\begin{aligned}
E\left[\widetilde{h}_i\right] &= M\bar{h}_i \sum_{n=0}^{M-1} (-1)^n \binom{M-1}{n} \frac{1}{(n+1)^2} \\
&= M\bar{h}_i \sum_{n=1}^{M} (-1)^{(n-1)} \binom{M-1}{n-1} \frac{1}{n^2} \\
&= \bar{h}_i \sum_{n=1}^{M} \frac{M(M-1)!}{n(n-1)!(M-n)!} \frac{(-1)^{(n-1)}}{n} \\
&= \bar{h}_i \sum_{n=1}^{M} \frac{M!}{n!(M-n)!} \frac{(-1)^{(n-1)}}{n} \\
&= \bar{h}_i \sum_{n=1}^{M} \frac{(-1)^{(n-1)}}{n} \binom{M}{n}
\end{aligned}$$

Thus, we prove that the multiuser diversity gain in the case of Equal Resource-Share Fairness

between M users and Rayleigh fading channels equals:

$$gain_i = \sum_{n=1}^{M} \frac{(-1)^{(n-1)}}{n} \binom{M}{n} \tag{C.11}$$

Equation (C.11) can be further simplified as:

$$gain_i = \sum_{n=1}^{M} \frac{1}{n} \tag{C.12}$$

C.4 Derivation of Equation (5.14)

When we have two users with Rayleigh fading channels, then in order to achieve the access ratios $[ar_1,\ ar_2]$, the weighting factors should be selected as:

$$\mu_1 = \alpha \frac{ar_1}{\bar{h}_1}, \quad \mu_2 = \alpha \frac{ar_2}{\bar{h}_2} \tag{C.13}$$

where α is any scalar.

This can be proved by substituting the weighting factors in (C.13) into equation (C.2):

$$\begin{aligned} ar_1 &= \int_0^\infty \frac{1}{\bar{h}_1} \exp\left(\frac{-x}{\bar{h}_1}\right) \left[1 - \exp\left(\frac{-\mu_1 x}{\mu_2 \bar{h}_2}\right)\right] dx \\ &= \int_0^\infty \frac{1}{\bar{h}_1} \exp\left(\frac{-x}{\bar{h}_1}\right) dx - \\ &\quad \int_0^\infty \frac{1}{\bar{h}_1} \exp\left(\frac{-x}{\bar{h}_1} + \frac{-ar_1 x}{ar_2 \bar{h}_1}\right) dx \\ &= 1 - \int_0^\infty \frac{1}{\bar{h}_1} \exp\left(\frac{-x}{ar_2 \bar{h}_1}\right) dx \\ &= 1 - ar_2 = ar_1 \end{aligned}$$

To compute the multiuser diversity gain of first user, we first compute $E\left[\widetilde{h_1}\right]$. $\widetilde{f_1}$ can be obtained using equation (C.3).

$$\widetilde{f_1}(x) = \frac{1}{ar_1 \bar{h}_1} \exp\left(\frac{-x}{\bar{h}_1}\right) \left[1 - \exp\left(\frac{-\mu_1 x}{\mu_2 \bar{h}_2}\right)\right] \tag{C.14}$$

By substituting the values of weighting factors in (C.13) into (C.14), we get:

$$\begin{aligned} E\left[\widetilde{h_1}\right] &= \int_0^{\infty} \frac{x}{ar_1\bar{h}_1} \exp\left(\frac{-x}{\bar{h}_1}\right)\left[1-\exp\left(\frac{-ar_1x}{ar_2\bar{h}_1}\right)\right]dx \\ &= \frac{1}{ar_1\bar{h}_1}\left[\left(\bar{h}_1\right)^2-\left(\bar{h}_1ar_2\right)^2\right] = \bar{h}_1\left(1+ar_2\right) \end{aligned}$$

Thus, we prove that:

$$gain_1 = 1 + ar_2 = 2 - ar_1 \tag{C.15}$$

Bibliography

[1] Y. Cao and V. Li, "Scheduling algorithms in broad-band wireless networks," *Proceedings of the IEEE*, vol. 89, no. 1, pp. 76–87, Jan. 2001.

[2] P. Bhagwat, A. Krishna, and S. Tripathi, "Enhancing throughput over wireless lan's using channel-state dependent packet scheduling," in *Proceedings IEEE Conference on Computer Communications (INFOCOM)*, Mar. 1996, vol. 3, pp. 1133–1140.

[3] C. Fragouli, V. Sivaraman, and M. Srivastava, "Controlled multimedia wireless link sharing via enhanced class-based queuing with channel-state dependent packet scheduling," in *Proceedings IEEE Conference on Computer Communications (INFOCOM)*, Mar. 1998, vol. 2, pp. 572–580.

[4] S. Lu and V. Bharghavan, "Fair scheduling in wireless packet networks," *IEEE/ACM Trans. Networking*, vol. 7, no. 4, pp. 473–489, Aug. 1999.

[5] T. Eugene, I. Stoica, and H. Zhang, "Packet fair queuing algorithms for wireless networks with location-dependent errors," in *Proceedings IEEE Conference on Computer Communications (INFOCOM)*, Mar. 1998, vol. 3, pp. 1103–1111.

[6] J. Gomez, A. Campbell, and H. Morikawa, "The havana framework for supporting application and channel dependent qos in wireless networks," in *ICNP*, Nov. 1999, pp. 235–244.

[7] X. Wang, G. Giannakis, and A. Marques, "A unified approach to QoS-guaranteed scheduling for channel-adaptive wireless networks," *Proceedings of the IEEE*, vol. 95, no. 12, pp. 2410–2431, Dec. 2007.

[8] H. Zhang, "Service disciplines for guaranteed performance service in packet-switching networks," *Proceedings of the IEEE*, vol. 83, no. 10, pp. 1374–1396, Oct. 2007.

[9] A. Demers, S. Keshav, and S. Shenker, "Analysis and simulation of a fair queueing algorithm," in *Proc. ACM SIGCOMM*, Sept. 1989, pp. 3–12.

[10] L. Zhang, "Virtual clock: A new traffic control algorithm for packet switching networks," in *Proc. ACM SIGCOMM*, Sept. 1990, pp. 19–29.

[11] S. Golestani, "A self-clocked fair queueing scheme for broadband applications," in *Proceedings IEEE Conference on Computer Communications (INFOCOM)*, June 1994, pp. 636–646.

[12] D. Kandlur, K. Shin, and D. Ferrari, "Real-time communication in multi-hop networks," in *Proc. 11th Int. Conf. Distributed Computer System*, May 1991, pp. 300–307.

[13] J. G. Proakis, *Digital Communications*, McGraw-Hill International Editions, third edition, 1995.

[14] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*, Cambridge University Press, May 2005.

[15] A. Goldsmith, *Wireless Communications*, Cambridge University Press, 2005.

[16] G. Kramer, I. Maric, and R. Yates, *Cooperative Communications, Foundations and Trends in Networking*, NOW Publishers Inc., 2006.

[17] F. Fitzek and M. Marcos, *Cooperation in Wireless Networks: Principles and Applications*, Springer, 2006.

[18] R. Knopp and P. A. Humblet, "Information capacity and power control in single-cell multiuser communications," in *Proceedings IEEE International Conference on Communications (ICC)*, Seattle, WA, USA, June 1995, pp. 331–335.

[19] D. Tse, "Optimal power allocation over parallel gaussian broadcast channels," in *Proceedings IEEE Information Theory Workshop*, June 1997.

[20] P. Viswanath, D. Tse, and R. Laroia, "Opportunistic beamforming using dumb antennas," *IEEE Transactions on Information Theory*, vol. 48, no. 6, pp. 1277–1294, June 2002.

[21] A. Duel-Hallen, "Fading channel prediction for mobile radio adaptive transmission systems," *Proceedings of the IEEE*, vol. 95, no. 12, pp. 2299–2313, Dec. 2007.

[22] T. Eriksson and T. Ottosson, "Compression of feedback for adaptive transmission and scheduling," *Proceedings of the IEEE*, vol. 95, no. 12, pp. 2314–2321, Dec. 2007.

[23] P. Bender, P. Black, M. Grob, R. Padovani, N. Sindhushayana, and A. Viterbi, "CDMA/HDR: A bandwidth-efficient high-speed wireless data service for nomadic users," *IEEE Communications Magazine*, vol. 38, no. 7, pp. 70–77, July 2000.

[24] P. Svedman, S. Wilson, L. Cimini, and B. Ottersten, "Opportunistic beamforming and scheduling for OFDMA systems," *IEEE Transactions on Communications*, vol. 55, no. 5, pp. 941–952, May 2007.

[25] S. Sampei and H. Harada, "System design issues and performance evaluations for adaptive modulation in new wireless access systems," *Proceedings of the IEEE*, vol. 95, no. 12, pp. 2456–2471, Dec. 2007.

[26] A. Goldsmith and S. Chua, "Variable-rate variable-power MQAM for fading channels," *IEEE Transactions on Communications*, vol. 45, no. 10, pp. 1218–1230, Oct. 1997.

[27] S. Lin and D. J. Costello, *Error Control Coding*, Pearson Prentice Hall, 2004.

[28] S. Shakkottai, T. S. Rappaport, and P. C. Karlsson, "Cross-layer design for wireless networks," *IEEE Communications Magazine*, pp. 74–80, Oct. 2003.

[29] C. Anton-Haro, P. Svedman, M. Bengtsson, A. Alexiou, and A. Gameiro, "Cross-layer scheduling for multi-user MIMO systems," *IEEE Communications Magazine*, vol. 44, no. 9, pp. 39–45, Sept. 2006.

[30] R. Berry and E. Yeh, "Cross-layer wireless resource allocation," *IEEE Signal Processing Magazine*, vol. 21, no. 5, pp. 59–68, Sept. 2004.

[31] V. Kawadia and P. R. Kumar, "A cautinary perspective on cross layer design," *available from http://black.csl.uiuc.edu/~prkumar/*, July 2003.

[32] V. Srivastava and M. Motani, "Cross-layer design: A survey and the road ahead," *IEEE Communications Magazine*, vol. 43, no. 12, pp. 112–119, Dec. 2005.

[33] D. Avidor, S. Mukherjee, J. Ling, and C. Papadias, "On asymptotically fair transmission scheduling over fading channels with measurement delay," *IEEE Transactions on Wireless Communications*, vol. 5, no. 7, pp. 1626–1633, July 2006.

[34] D. Tse and S. Hanly, "Multiaccess fading channels – Part 1: Polymatroid structure, optimal resource allocation and throughput capacities," *IEEE Transactions on Information Theory*, vol. 44, no. 7, pp. 2796–2815, Nov. 1998.

[35] M. Shaqfeh and N. Goertz, "Comments on the boundary of the capacity region of multiaccess fading channels," *to appear in IEEE Transactions on Information Theory.*

[36] M. Shaqfeh and N. Goertz, "Channel-aware scheduling with resource-sharing constraints in wireless networks," in *Proceedings IEEE International Conference on Communications (ICC)*, May 2008, pp. 4149–4153.

[37] M. Shaqfeh and N. Goertz, "Performance analysis of scheduling policies for delay-tolerant applications in centralized wireless networks," in *2008 International Symposium on Performance Evaluation of Computer and Telecommunication Systems (SPECTS 2008)*, June 2008, pp. 1–8.

[38] M. Shaqfeh, N. Goertz, and S. McLaughlin, "Organizing multiuser operation in centralized wireless networks," in *Wireless World Research Forum Meeting 20*, Apr. 2008, pp. 1–6.

[39] M. Shaqfeh and N. Goertz, "A new generic framework for comparison of flexible schedulers for delay-tolerant wireless applications," *submitted to IEEE Transactions on Communications*, June 2008.

[40] M. Shaqfeh and N. Goertz, "Ergodic capacity of block-fading gaussian broadcast and multi-access channels for single-user-selection and constant-power," *under preparation for submission into Transactions on Information Theory.*

[41] M. Shaqfeh and N. Goertz, "Systematic modification of parity-check matrices for efficient encoding of LDPC codes," in *Proceedings IEEE International Conference on Communications (ICC)*, June 2007, pp. 945–950.

[42] 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, no. 10, pp. 2053–2064, Oct. 2001.

[43] D. Park, H. Seo, H. Kwon, and B. Lee, "Wireless packet scheduling based on the cumulative distribution function of user transmission rates," *IEEE Transactions on Communications*, vol. 53, no. 11, pp. 1919–1929, Nov. 2005.

[44] A. Goldsmith and P. Varaiya, "Capacity of fading channels with channel side information," *IEEE Transactions on Information Theory*, vol. 43, no. 6, pp. 1986–1992, Nov. 1997.

[45] L. Zheng and D. Tse, "Diversity and multiplexing: A fundamental tradeoff in multiple-antenna channels," *IEEE Transactions on Information Theory*, vol. 49, no. 5, pp. 1073–1096, May 2003.

[46] D. Tse, P. Viswanath, and L. Zheng, "Diversity–multiplexing tradeoff in multiple-access channels," *IEEE Transactions on Information Theory*, vol. 50, no. 9, pp. 1859–1874, Sept. 2004.

[47] D. Tse, "Optimal power allocation over parallel Gaussian broadcast channels," *unpublished, available at www.eecs.berkeley.edu/~dtse/broadcast2.pdf.*

[48] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, March 2004.

[49] L. Li and A. Goldsmith, "Capacity and optimal resource allocation for fading broadcast channels – Part 1: Ergodic capacity," *IEEE Transactions on Information Theory*, vol. 47, no. 3, pp. 1083–1102, Mar. 2001.

[50] E. Biglieri, J. Proakis, and S. Shamai, "Fading channels: Information-theoretic and communications aspects," *IEEE Transactions on Information Theory*, vol. 44, no. 6, pp. 2619–2692, Oct. 1998.

[51] S. Hanly and D. Tse, "Multiaccess fading channels – Part 2: Delay-limited capacities," *IEEE Transactions on Information Theory*, vol. 44, no. 7, pp. 2816–2831, Nov. 1998.

[52] L. Li and A. Goldsmith, "Capacity and optimal resource allocation for fading broadcast channels – Part 2: Outage capacity," *IEEE Transactions on Information Theory*, vol. 47, no. 3, pp. 1103–1127, Mar. 2001.

[53] I. Toufik and R. Knopp, "Multiuser channel allocation algorithms achieving hard fairness," in *Proceedings IEEE Global Communications Conference (GLOBECOM)*, Nov. 2004, vol. 1, pp. 146–150.

[54] G. Caire, R. Müller, and R. Knopp, "Hard fairness versus proportional fairness in wireless communications: The single-cell case," *IEEE Transactions on Information Theory*, vol. 53, no. 4, pp. 1366–1385, Apr. 2007.

[55] T. Cover and J. Thomas, *Elements of Information Theory*, John Wiley & Sons, Inc., 1991.

[56] G. Gupta and S. Toumpis, "Power allocation over parallel Gaussian multiple access and broadcast channels," *IEEE Transactions on Information Theory*, vol. 52, no. 7, pp. 3274–3282, July 2006.

[57] N. Jindal, S. Vishwanath, and A. Goldsmith, "On the duality of Gaussian multiple-access and broadcast channels," *IEEE Transactions on Information Theory*, vol. 50, no. 5, pp. 768–783, May 2004.

[58] X. Wang, G. Giannakis, and Y. Yu, "Channel-adaptive optimal OFDMA scheduling," in *41st Annual Conference on information Sciences and Systems, 2007. CISS '07*, Mar. 2007, pp. 536–541.

[59] I. Wong and B. Evans, "Optimal OFDMA resource allocation with linear complexity to maximize ergodic weighted sum capacity," in *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Apr. 2007, vol. 3, pp. 601–604.

[60] R. Corless, G. Gonnet, D. Hare, D. Jeffrey, and D. Knuth, "On the Lambert W function," *Advances in Computational Mathematics*, vol. 5, pp. 329–359, 1996.

[61] Z. Shen, J. Andrews, and B. Evans, "Adaptive resource allocation in multiuser OFDM systems with proportional rate constraints," *IEEE Transactions on Wireless Communications*, vol. 4, no. 6, pp. 2726–2737, Nov. 2005.

[62] G. Song and Y. Li, "Utility-based resource allocation and scheduling in OFDM-based wireless broadband networks," *IEEE Communications Magazine*, vol. 43, no. 12, pp. 127–134, Dec. 2005.

[63] G. Song and Y. Li, "Cross-layer optimization for OFDM wireless networks – Part 1: Theoretical framework," *IEEE Transactions on Wireless Communications*, vol. 4, no. 2, pp. 614–624, Mar. 2005.

[64] G. Song and Y. Li, "Cross-layer optimization for OFDM wireless networks – Part 2: Algorithm development," *IEEE Transactions on Wireless Communications*, vol. 4, no. 2, pp. 625–634, Mar. 2005.

[65] 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. 39, no. 2, pp. 150–154, Feb. 2001.

[66] K. Seong, R. Narasimhan, and J. Cioffi, "Queue proportional scheduling via geometric programming in fading broadcast channels," *IEEE Journal on Selected Areas in Communications*, vol. 24, no. 8, pp. 1593–1602, Aug. 2006.

[67] Z. Han, X. Liu, Z. Wang, and K. Liu, "Delay sensitive scheduling schemes for heterogeneous QoS over wireless networks," *IEEE Transactions on Wireless Communications*, vol. 6, no. 2, pp. 423–428, Feb. 2007.

[68] S. Ryu, B. Ryu, H. Seo, and M. Shin, "Urgency and efficiency based packet scheduling algorithm for OFDMA wireless system," in *Proceedings IEEE International Conference on Communications (ICC)*, May 2005, vol. 4, pp. 2779–2785.

[69] B. Borst and P. Whiting, "Dynamic rate control algorithms for HDR throughput optimization," in *Proceedings IEEE Conference on Computer Communications (INFOCOM)*, Apr. 2001, vol. 2, pp. 976–985.

[70] L. Yang and M. Alouini, "Performance analysis of multiuser selection diversity," *IEEE Transactions on Vehicular Technology*, vol. 55, no. 6, pp. 1848–1861, Nov. 2006.

[71] D. Park and B. Lee, "QoS support by using CDF-based wireless packet scheduling in fading channels," *IEEE Transactions on Communications*, vol. 54, no. 11, pp. 2051–2061, Nov. 2006.

[72] T. Nguyen and Y. Han, "A proportional fairness algorithm with QoS provision in downlink OFDMA systems," *IEEE Communications Letters*, vol. 10, no. 11, pp. 760–762, Nov. 2006.

[73] Y. Ma, "Proportional fair scheduling for downlink OFDMA," in *Proceedings IEEE International Conference on Communications (ICC)*, June 2007, pp. 4843–4848.

[74] Lin. Yang and M. Alouini, "Performance analysis of multiuser selection diversity," in *Proceedings IEEE International Conference on Communications (ICC)*, June 2004, vol. 5, pp. 3066–3070.

[75] G. Song and Y. Li, "Asymptotic throughput analysis for channel-aware scheduling," *IEEE Transactions on Communications*, vol. 54, no. 10, pp. 1827–1834, Oct. 2006.

[76] H. Weingarten, Y. Steinberg, and S. Shamai, "The capacity region of the gaussian multiple-input multiple-output broadcast channel," *IEEE Transactions on Information Theory*, vol. 52, no. 9, pp. 3936–3964, Sept. 2006.