Vegard Hassel

Design Issues and Performance Analysis for Opportunistic Scheduling Algorithms in Wireless Networks

Thesis for the degree of philosophiae doctor

Trondheim, January 2007

Norwegian University of Science and Technology
Faculty of Information Technology, Mathematics
and Electrical Engineering
Department of Electronics and Telecommunications

NTNU
Innovation and Creativity

NTNU
Norwegian University of Science and Technology

Thesis for the degree of philosophiae doctor

Faculty of Information Technology, Mathematics
and Electrical Engineering
Department of Electronics and Telecommunications

ISBN 978-82-471-0573-3 (printed version)
ISBN 978-82-471-0587-0 (electronic version)
ISSN 1503-8181

Doctoral theses at NTNU, 2007:27

Printed by NTNU-trykk

Abstract

This doctoral thesis is a collection of six papers preceeded by an introduction. All the papers are related to design issues and performance analysis for opportunistic scheduling algorithms in cellular networks.

Opportunistic scheduling algorithms can provide higher throughput and increased quality-of-service (QoS) in wireless networks by giving priority to the users with favorable channel conditions. Such algorithms are already implemented in equipment based on wireless LAN standards, the HDR standard, the HSDPA standard, and the Mobile WiMAX standard, but are often not a part of the standard itself. The implemented algorithms are often based on intuition rather than theoretical investigations, and consequently, it is a need for a better understanding of the theoretical limits for how well such algorithms can perform and how such algorithms can be implemented in the most efficient way.

The design issues handled in this thesis are related to feedback algorithms and scheduling algorithms for increased throughput guarantees. Channel quality estimation and feedback is the basis for opportunistic scheduling, and two novel feedback algorithms are proposed to reduce the overhead from channel quality feedback. The results show that the feedback can be reduced to only obtaining feedback from the user that the system wants to schedule. An adaptive scheduling algorithm for obtaining increased throughput guarantees is also developed. Results from simulations show that this algorithm can double the throughput guarantees in modern cellular networks compared to other well-known scheduling algorithms.

The performance of opportunistic scheduling algorithms is analyzed through analytical expressions and simulations of feedback delay, fairness and throughput guarantees. It is shown how delayed feedback can lead to reduced throughput or increased bit error rate in a system with opportunistic scheduling. Closed-form expressions are also found for two types of fairness, and throughput guarantees of different well-known scheduling algorithms.

Preface

This dissertation is submitted in partial fulfillment of the requirements for the degree of *philosophiae doctor* (PhD) at the Department of Electronics and Telecommunications, Norwegian University of Science and Technology (NTNU). My main supervisor has been Professor Geir E. Øien at the Department of Electronics and Telecommunications at NTNU, while my co-supervisor has been Professor Peder Johannes Emstad at the Centre for Quantifiable Quality of Service in Communication Systems (Q2S) at NTNU.

The studies have been carried out in the period from January 2004 to December 2006, including one semester of course work, approximately one month of teaching assistant work, and approximately one month of supervising two master students and three bachelor students. The first one and a half years I mainly worked at NTNU, while the last year I have mainly worked at the University Graduate Center at Kjeller (UniK) outside Oslo. In May 2005, I visited Professor Mohamed-Slim Alouini in Tunis for a period of two weeks, and during the period from August to December 2005, I spent four months at the MAESTRO project at l'Institut National de Recherche en Informatique et en Automatique (INRIA) in Sophia Antipolis at the French Riviera. During this stay I also had the opportunity to work with Professor David Gesbert at Insitut Eurécom in Sophia Antipolis.

The work has been funded through the Faculty of Information Technology, Mathematics and Electrical Engineering at NTNU, via the NFR project "Co-Optimized Ubiquitous Broadband Access Networks" (CUBAN), while the assistantship was financed via the Department of Electronics and Telecommunications, NTNU.

Acknowledgements

First and foremost I would like to thank my supervisor Geir Øien for his valuable feedback, his positivism, and his fat wallet. I am also grateful to

David Gesbert and Mohamed-Slim Alouini for their ideas, valuable comments, and guidance. Without their help, I would never have finished this thesis on time. Hend Koubaa helped me to understand more about protocols, and I am happy that we had the opportunity to work together.

My thanks also go to my colleagues at the signal processing group at NTNU. Especially, I would like to thank "toast master" Anders Gjendemsjø for his support and humor. Pål Anders Floor, Robert Bains, and Sébastien de la Kethulle de Ryhove have also been very supportive, both socially and scientifically.

I am grateful to Philippe Nain at INRIA for letting me have an office at his group. At INRIA, my roommate Nicolas Bonneau helped me with many practical issues, and he was also the only Frenchman who had the patience to speak French with me. Thank you.

Moreover, my thanks go to UniK for lending me an office there, and I would also like to thank Frode Bøhagen, Hans Jørgen Bang, Torbjörn Ekman, and Pål Orten for their fruitful discussions at UniK.

My thanks also go to Øyvind Janbu, Frode Bøhagen, and Anders Gjendemsjø for reviewing the introduction of this thesis.

Finally, I would like to express my gratitude to my family for their support and housing during the last three years. Especially, I would like to thank my wife Tove Irene for her unlimited love, her optimism, her curly hair, her smile, and for being so energetic and supportive.

Oslo, December 2006
Vegard Hassel

Contents

Abbreviations

1xEVDO	1x evolution-data optimized
ACK	Acknowledgment packet
ACR	Adaptive continuous rate
ADR	Adaptive discrete rate
AFL	Absolute feedback load
ARQ	Automatic repeat request
AWGN	Additive white Gaussian noise
BER	Bit error rate
BF	Beamforming
CDF	Cumulative distribution function
CDM	Code-division multiplexing
CDMA	Code-division multiple access
CDP	Call dropping probability
CLT	Central limit theorem
CNR	Carrier-to-noise ratio
CSI	Channel state information
CSMA/CA	Carrier-sense multiple access with collision avoidance
CTS	Clear to send
dB	Decibel
DPC	Dirty paper coding
DSSS	Direct sequence spread spectrum
FB	Feedback packet
FC	Frame control
FCS	Fast cell selection
FCS	Frame check sequence
FDD	Frequency-division duplexing
FDM	Frequency-division multiplexing

FDMA	Frequency-division multiple access
GPS	Generalized processor sharing
GSM	Global system for mobile communication
HDR	High data rate
HOL	Head-of-the-line
HSDPA	High speed downlink packet access
HSPA	High speed packet access
IEEE	Institute of electrical and electronics engineers
ILS	Inverse-log scheduling
IP	Internet protocol
JFI	Jain's fairness index
LAN	Local area network
LOS	Line-of-sight
MAC	Medium access control
MASSE	Maximum average system spectral efficiency
MCS	Maximum carrier-to-noise ratio scheduling
MIMO	Multiple-input, multiple-output
M-LWDF	Modified largest weighted delay first
M-QAM	M-ary quadrature amplitude modulation
MUD	Multiuser diversity
NCS	Normalized carrier-to-noise ratio scheduling
NFL	Normalized feedback load
N-ORR	Normalized opportunistic round robin
OFDM	Orthogonal frequency-division multiplexing
OFDMA	Orthogonal frequency-division multiple access
ORR	Opportunistic round robin
PDF	Probability density function
PFS	Proportional fair scheduling
PLCP	Physical layer convergence protocol
PLP	Packet loss probability
PMF	Probability mass function
QAM	Quadrature amplitude modulation
QoS	Quality-of-service
QRY	Query packet
RA	Receiver address
RR	Round robin

RRM	Radio resource management
RTS	Request to send
SCSIC	Superposition coding with successive interference cancellation
SDM	Space-division multiplexing
SDMA	Space-division multiple access
SIFS	Short interframe space
SISO	Single-input, single-output
SMUD	Selective multiuser diversity
STBC	Space-time block coding
TA	Transmitter address
TDD	Time-division duplexing
TDM	Time-division multiplexing
TDMA	Time-division multiple access
TGVP	Throughput guarantee violation probability
VoIP	Voice over IP
WFQ	Weighted fair queuing
WiMAX	Worldwide interoperability for microwave access
WINNER	Wireless world initiative new radio
WLAN	Wireless LAN

Part I

Introduction

Introduction

A range of different wireless standards has been developed during the last decades [1–3]. The characteristics of these standards depend on the objectives of the wireless network design. In for example sensor networks, the main design objectives can be low power consumption and small mobile devices, while the main design objectives for cellular networks can be high throughput and fulfillment of the quality-of-service (QoS) requirements of the mobile users.

To be able to increase the throughput and to provide QoS to the users in a wireless network, *radio resource management* (RRM) should be implemented [4–6]. RRM can be defined as the real-time process of exploiting, allocating, and controlling the radio resources according to the varying number of users and applications in the network, and *opportunistic scheduling* is a vital part of an RRM system. Opportunistic scheduling denotes the process of selecting which of the mobile users that are going to transmit or receive information on the wireless channel to increase the throughput and the QoS [7, 8], and this thesis handles design and performance evaluation issues related to opportunistic scheduling algorithms in cellular wireless networks.

Fig. 1.1 shows an example of a cellular network using opportunistic scheduling. The exchange of user data is indicated with solid arrows in the figure, while the exchange of signaling information is marked with dotted arrows. In this figure, the term *transceiver* denotes both the transmitter and the receiver. The *scheduler*, located in the base station, decides which users who are going to transmit or receive. However, to take these scheduling decisions, the scheduler needs to know the channel quality of the users. This channel quality is often both estimated at the base station and locally at the mobile users. The channel quality estimates found by the mobile users need to be fed back to the base station so that the scheduler can decide which users to schedule. After the scheduling decision is taken, each of the mobile users are often notified before transmission can start.

Opportunistic scheduling algorithms are already implemented in

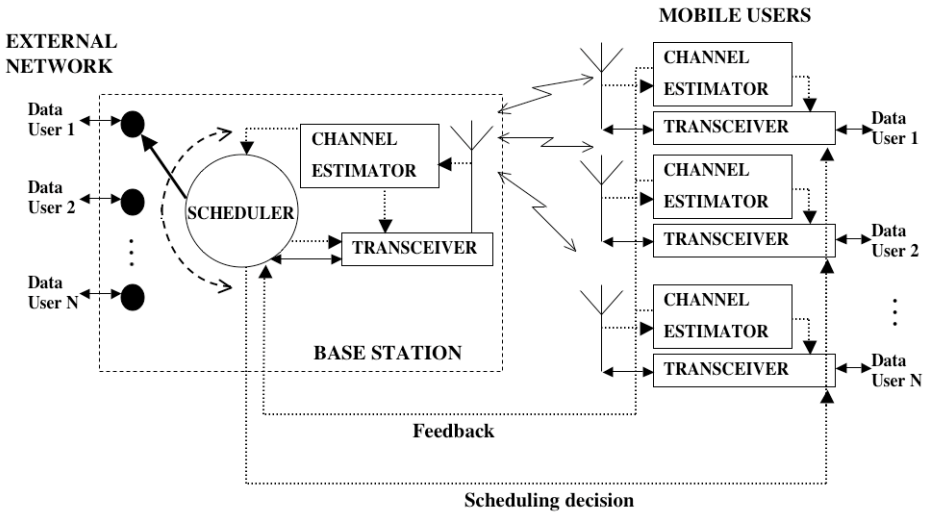

FIGURE 1.1: System model for opportunistic scheduling

equipment based on many modern wireless standards, but are often not a part of the standards themselves [3]. The implemented algorithms are often based on intuition and this thesis will therefore focus on theoretical investigations of design issues and performance issues related to opportunistic scheduling algorithms. Such theoretical investigations are vital to be able to develop more efficient scheduling algorithms.

This introduction will give a background to the six papers included in this thesis and a general overview of different scheduling issues. The introduction commences with a discussion of radio resources and RRM techniques, continues with a description of performance issues and design issues related to opportunistic scheduling algorithms, and ends with a description of the included papers, the main contributions of the thesis, and suggestions for future research. Section 1 discusses the fixed radio resources that are available in wireless networks, while Section 2 gives an overview of RRM and describes the role of opportunistic scheduling in an RRM system. Next, different performance measures for opportunistic scheduling algorithms are handled in Section 3, and four well-known opportunistic scheduling algorithms are listed in Section 4. Section 5 discusses design issues related to the physical and the Medium Access Control (MAC) layers, while Section 6 gives an overview of cross-layer design issues. Then, a description of the included papers is given in Section 7, and finally, Section 8 and Section 9 list respectively the main contributions of

the thesis and some suggestions for future research.

It should be noted that the focus of the papers in the thesis is opportunistic scheduling issues for *cellular* and/or *centralized* networks. However, to have a more complete introduction, different scheduling issues for wireless ad hoc networks are also mentioned.

1 Fixed Radio Resources in Wireless Networks

One of the main challenges in wireless networks is to use the available radio resources in the most efficient way according to the needs of the mobile users. The amount of radio resources available in a network does often change with time, however, some fixed characteristics will always constrain the amount of radio resources. Three important fixed characteristics are the *system architecture*, the *frequency spectrum* available, and the *access techniques* implemented in the system.

1.1 System Architecture

Wireless networks have either a *cellular architecture* or an *ad hoc architecture* [2]. The cellular architecture is based on stationary base stations serving the mobile customers, and we often say that this architecture is *centralized* since most of the network activity is managed by the network infrastructure, i.e., base stations, base station controllers, and switches. The ad hoc architecture is based on mobile users communicating directly with each other or by relaying via other mobile users in the network. This type of architecture is often referred to as *distributed* since each mobile user governs its own behavior. The location of the base stations, the number and type of antennas per base station or mobile terminal, the complexity of the equipment, and the available transmission power can be regarded as important components of an architecture. System architecture design is therefore a complex task with multiple challenges, e.g. for cellular networks it is critical to build the network based on extensive radio planning in order to adjust the network design according to the terrain.

1.2 Frequency Spectrum

In general, the total throughput of a wireless network will increase with the amount of frequency spectrum that is available. In most countries the radio spectrum is publicly regulated to limit the interference between different systems. This means that each network operator governs his own

share of the frequency spectrum and can deploy wireless networks that operate on these frequencies. There are also unlicensed parts of the frequency spectrum that can be used freely.

1.3 Access Techniques

Access techniques are implemented so that multiple users can share the same physical channel, and they are designed to exploit the architecture and the available frequency spectrum in the most efficient way, i.e., to increase the spectral efficiency. Spectral efficiency is defined to be the throughput per frequency spectrum available, usually measured in bits per second per Hertz. The access techniques are often based on time-division multiplexing (TDM), frequency-division multiplexing (FDM), code-division multiplexing (CDM), and space-division multiplexing (SDM) for providing wireless multiple access along the dimensions time, frequency, code, or space [1, 2].

Commonly used access techniques are ALOHA-based techniques (e.g. CSMA/CA), time-division multiple access (TDMA), frequency-division multiple access (FDMA), code-division multiple access (CDMA), or space-division multiple access (SDMA) [1, 2]. In this thesis, only TDM-based transmission will be considered, i.e., the thesis only considers systems where the data to or from different users are transmitted in time-slots, where each time-slot contains many modulation symbols. Amongst the TDM-based access techniques we also find time-division duplexing (TDD), where the time-slots are used to separate both transmission to different users and uplink and downlink transmission. It should be noted that TDM-based communication does not exclude FDM or CDM. For example, in the cellular standard HSPA, CDM is used in combination with frequency-division duplexing (FDD), while Mobile WiMAX uses orthogonal frequency-division multiplexing (OFDM) in combination with TDD [3]. Multiple-antenna techniques like diversity combining [9] and multiple-input, multiple-output (MIMO) can further enhance the efficiency of the access techniques.

2 Radio Resource Management

Architecture, frequency spectrum, and access techniques are all more or less fixed characteristics of a wireless network. However, due to the temporally and spatially varying channel quality of each of the mobile users in the network, the efficiency of the network will also vary with time. In

FIGURE 1.2: Block diagram of radio resource management system.

addition, a wireless network in full use will also have a constantly changing number of mobile users, each having changing number of applications needing to transmit or receive data over the network. RRM is related to these constantly changing characteristics of the wireless network, and as already mentioned, RRM can be defined as the real-time process of exploiting, allocating, and controlling the radio resources according to the varying number of users and applications [4–6].

In most wireless networks, the RRM system solves three main challenges, namely, (i) adapting the transmission schemes to the varying nature of the wireless links, (ii) scheduling radio resources to the mobile users according to the instantaneous channel quality and QoS requirements of the users' applications, and (iii) controlling admission of new users and applications into the network and the handoff process between cells according to the available radio resources.

Fig. 1.2 shows a block diagram that illustrates the three parts of an RRM system, and the arrows in the figure show the flow of information in the system. In addition to the internal flow of information in the RRM system, each of these three parts do also exchange information with the mobile users in the network. The process of adapting coding, modulation, and power to the channel conditions is often called *adaptive transmission* or link adaptation. This process will normally be *distributed* in the network, i.e., the unit that transmits data, adapts the coding, modulation, and power to the channel conditions. *Opportunistic scheduling algorithms* and *admission control algorithms* are often *centralized* in the network infrastructure, at the base stations or at the base station controllers. This means that these two parts of an RRM system need to collect both channel state information (CSI) from the mobile users and information about the QoS requirements of the users. In the following three paragraphs, the three parts of the RRM system will be described in further detail.

2.1 Adaptive Transmission

A wide range of adaptive transmission techniques have been developed during the last years and most modern wireless systems and standards have implemented such techniques [10–16]. Most types of wireless networks apply digital modulation where the amplitude, frequency, and/or phase of a *carrier frequency* are varied according to the *modulation symbols* being transmitted [17]. Adaptive transmission can for example be conducted by adapting the modulation constellation to the channel quality of a mobile user. In addition, adaptive transmission can be performed by varying the coding and/or the transmission power according to the channel quality or by dynamically allocating the mobile users to the carrier frequencies with the best channel quality.

In a centralized system, the transmission from the base station to the mobile users is denoted as *downlink* transmission and the transmission from the mobile users to the base station as *uplink* transmission. The process of adapting the transmission schemes to the downlink channel quality of the users, requires updated CSI estimates available at the base station. In addition, a user can use CSI estimates for the uplink to adapt his transmission. If it is assumed that the channel can vary significantly from time-slot to time-slot, CSI estimates are needed in each time-slot to conduct the adaptive transmission. CSI estimates can for example be derived from deterministically known *pilot symbols* transmitted in-between the data symbols [18, 19]. In this thesis it is assumed that such quality measurements are based on estimating the carrier-to-noise ratio (CNR), i.e., the signal strength of the received signal relative to the noise from other sources, defined as [2]:

$$\gamma = \frac{P_r}{N_0 W + P_I}, \tag{1.1}$$

where P_r [dBm] denotes the received signal power, N_0 [dBm/Hz] is the noise power spectral density, W [Hz] is the received signal bandwidth, and P_I [dBm] expresses the sum of the received power associated with the intercell and intracell interference. Throughout the thesis it is assumed that the CNR estimation can be executed perfectly.

If the network uses TDM-based transmission with different carrier frequencies on uplink and downlink, CSI estimates have to be sent from the base station to the mobile users and from the mobile users to the base station. The base station uses the received CSI estimates to adapt the transmission on the downlink, while the mobile users use the received CSI estimates to adapt the transmission on the uplink. However, if the network uses TDD as access technique, the *reciprocity* between a user's uplink and downlink

channel can be exploited [20]. This means that the channel quality of a user is the same for uplink and downlink. For such networks, the downlink can be estimated by sending pilot symbols on the uplink, and vice versa [3, 21].

2.2 Opportunistic Scheduling

Scheduling in wireless networks denotes the process of selecting which of the mobile users that are going to transmit or receive information on the wireless channel to increase the throughput and/or the QoS [7, 8]. QoS can for example be specified in terms of power consumption, buffer overflow or delay. In time-slotted systems, one user is not necessarily scheduled in each time-slot since CDM, FDM, or beamforming (BF) techniques make it possible to transmit to more users at a time. It is often said that scheduling decisions are taken by the *scheduler*. In real-life networks, this scheduler is an algorithm, often implemented centrally at the base station in cellular networks [7]. In many modern wireless standards, the scheduling is *opportunistic*; this means that the mobile users that experience favorable channel conditions, as measured by some suitable metric, are given priority to transmit or receive data. By giving priority to the users that have the best channel conditions, the system spectral efficiency and/or the QoS will increase. This in a consequence of the *multiuser diversity* (MUD) that exists between the users [22]. However, the channel conditions will often change slowly, and the time between a user is scheduled can therefore be significant, i.e., some users can be starved. To avoid such starvation it is necessary to take not only the channel conditions, but also the QoS demands of the users into account when designing scheduling algorithms.

In addition to MUD, other types of diversity can also be exploited by wireless networks to obtain better system spectral efficiency. Diversity in wireless networks can arise because of differently varying channel quality for different network technologies, different base stations, different mobile users, different frequencies, different codes, or different antennas. These kinds of diversity can be simultaneously exploited by allocating users to the base stations and/or networks with the best signal quality, scheduling the frequencies and time-slots to the users where they experience favorable signal quality, and combining the signals from different antennas to increase the spectral efficiency. How scheduling algorithms can be designed to exploit different types of diversity will be handled in Section 5.

Most of the information needed to perform opportunistic scheduling is already available from the adaptation of coding, modulation, and power. As discussed in the previous section, the CSI of the mobile users' downlink and uplink is available at the base-station. If it is assumed that all the users'

CSI is available for each time-slot, the base station can use this information to perform scheduling on a per time-slot basis.

After the scheduling decision is taken, this decision needs to be available to execute both the downlink and uplink transmission. The downlink transmission is executed by the base station while each mobile user executes the uplink transmission. Consequently, the base station needs to distribute this decision to the scheduled users for each time-slot in the uplink.

2.3 Admission Control and Handoff

The main goal of a scheduling algorithm is usually to obtain high throughput and to fulfill the QoS demands of the users' applications transmitting or receiving data in the network. However, this is an impossible task if there are too many running applications compared to the available radio resources in the network. Consequently, the number of applications in the network needs to be restricted so that the QoS of the users already in the system is not degraded in an unacceptable manner. Restricting the admission of new users into the network and restricting the execution of new applications is the mission of the admission control algorithm of an RRM system [23]. Admission control needs to be closely connected to the performance of the scheduling algorithm. If the scheduling algorithm performs well over a relatively long time horizon, the network might have spare capacity and the admission control algorithm can allow new users or applications into the system. However, if the scheduling algorithm has difficulties fulfilling the QoS requirements of the applications already running in the system, no new applications should be admitted. If a few applications are degrading the QoS of many other applications, the troublesome applications should perhaps be denied access to the network. Such denial of service procedures are often called *call dropping* [23].

Handoff is the process of transferring the servicing of a mobile user from one base station to a neighboring base station [23]. The handoff process is guided by the channel qualities that can be obtained between a mobile user and the surrounding base stations. If a mobile user experiences low quality on his channel, it might be that another base station in the area can provide better channel quality. Therefore, ideally the mobile users should measure the signal strength of the carrier frequencies of all the surrounding base stations in order to evaluate if a handoff could increase the quality of the wireless transmission. The handoff decision can either be taken by the individual users or centrally in the network infrastructure. However, the advantage of a centralized handoff control is that

the handoff process can be more tightly integrated with the other pars of the RRM-system.

A handoff can be hard or soft [23]. A hard handoff algorithm only allows the mobile users to communicate with one base station at a time, while soft handoff algorithms allow the mobile users to communicate with several base stations simultaneously. The advantage of the soft handoff process is that it is easier to guarantee QoS when the users can communicate with several base stations. Such systems are often referred to as having *load sharing* between the base stations [6]. However, in load sharing systems one user might degrade the performance of several cells at the same time. Consequently, a proper handoff algorithm should consider both the QoS of the user involved in the handoff process and the QoS of the other users that are communicating with the base stations involved in the handoff.

3 Performance Evaluation of Opportunistic Scheduling Algorithms

In the following some possible performance metrics and related issues influencing the design of wireless scheduling algorithms are listed.

3.1 Maximum Average System Spectral Efficiency (MASSE)

A fundamental aspect of all scheduling algorithms is the Maximum Average System Spectral Efficiency (MASSE) that can be obtained for a system, i.e., the maximally theoretically attainable throughput per bandwidth [bits/s/Hz], averaged over all the users in the system. The expression for the MASSE for constant-power, optimal rate adaptation, when additive white Gaussian noise (AWGN) is assumed, is defined as [24]:

$$\text{MASSE} = \sum_{i=1}^{N} p_i \int_0^{\infty} \log_2(1+\gamma) p_{\gamma_i^*}(\gamma)\, d\gamma, \tag{1.2}$$

where N is the number of users in the system, p_i is the probability of user i being selected in any time-slot (access probability), and $p_{\gamma_i^*}(\gamma)$ is the probability density function (PDF) of the CNR for the scheduling policy under study when user i is selected. The average CNR of user i is denoted as $\overline{\gamma}_i$. In this thesis, the MASSE achievable for a cell will be analyzed, however, the MASSE can also be analyzed for the wireless network as a whole. The theoretically attainable *throughput* of a cell can be seen as the MASSE multiplied by the frequency spectrum available.

It is important to note that the definition of MASSE above assumes that only one user is scheduled at a time. This assumption is also valid for all the scheduling algorithms described in the papers of this thesis. However, such scheduling algorithms are not optimal from an information theoretic perspective. According to information theory, the union of achievable rates under all multiuser strategies can be characterized by the *capacity region*. This capacity region can be obtained for the downlink by using *superposition coding with successive interference cancellation* (SCSIC) [25]. This technique is based on the base station sending information to all the mobile users in every time-slot and the coding of the signals to the different receivers are coded according to the channel quality of each of the receivers. The coding is such that a user which has a better channel quality than some other users can decode the signals intended for these users and subtract them from his own signal. This means that the user with the best channel can subtract the signals intended for all the other users in the system, while the user with the worst channel cannot subtract any interference from his signal. SCSIC can also be used to obtain the capacity region for uplink transmission and can be implemented in CDMA networks to increase the throughput significantly [26].

3.2 Fairness

Fairness in wireless networks is a measure of how equally the radio resources are allocated among the mobile users. The resources in question can be e.g. time-slots, frequencies, code-sequences, or throughput. A commonly used fairness measure is *Jain's Fairness Index* (JFI) [27]:

$$\mathrm{F}(K) = \frac{(\mathrm{E}_K[X])^2}{\mathrm{E}_K[X^2]} = \frac{(\mathrm{E}_K[X])^2}{(\mathrm{E}_K[X])^2 + \mathrm{var}(X)}, \tag{1.3}$$

where X is a random variable describing the amount of resource allocated to a user, $\mathrm{E}_K[\cdot]$ is the expectation calculated over the distribution of the resource allocation within a time-window of K [time-slots], and $\mathrm{var}(X)$ is the corresponding variance. This fairness index is bounded between zero and unity and has been used in many recent papers [28–30]. Fairness should always be evaluated for a window in time, and the scheduling algorithms that obtain high fairness over a relatively short time-window are denoted as *short-term fair*, while the algorithms that obtain high fairness over an infinite time-window are denoted as *asymptotically fair*. It should also be noted that the fairness measured according to the JFI should be compared to the MASSE performance of a network since a trade-off between MASSE and fairness can often be observed for different scheduling algorithms.

3.3 Power Consumption

Wireless networks that focus on low power consumption, e.g. sensor networks, will often seek to schedule the users that use the lowest energy per bit for their transmission [31].

3.4 Delay

Real-time, interactive applications like voice over IP (VoIP) or videoconferencing have strict requirements on what delays that can be permitted for the data received [32]. If packets are outdated, these cannot be used and will therefore be discarded by the system. Giving the right priorities to packets that have the longest delays can therefore help improving the QoS of applications with stringent delay requirements.

3.5 Buffer Overflow

Real-life systems have limited capacity in the buffers that contain data that are to be transmitted over the wireless network. If the traffic is bursty, i.e., the packet flow into the buffers varies significantly with time, and the buffers are not substantially over-dimensioned, there will be a probability for *buffer overflow* [33]. This probability can be reduced by implementing scheduling algorithms that take the delay of the packets in the buffers into account [34]. However, *flow control* in the system will also reduce the problem of buffer overflow, and the interaction between flow control and scheduling has been investigated in [35].

3.6 Throughput Guarantees

For real-time applications, the most important requirements are that the average throughput is high enough and that the packet delay is below certain delay constraints. Problems such as low battery energy and buffer overflow are often caused by the behavior of the mobile users themselves, and their applications. For example, the battery can become flat because users do not charge their mobile terminal and buffer overflow can arise if the mobile users run too many applications at the same time. Consequently, the operator of a cellular network may only be interested in guaranteeing a certain number of transmitted bits within a time-window. *Throughput guarantees*, i.e., the number of transmitted bits that can be promised to a user over a given time-window, can be used to quantify such guarantees.

Throughput guarantees can be seen as a performance criterion that represents a combination of throughput and delay. There are two types of

throughput guarantees that can be offered to the mobile users, namely *hard* or *deterministic throughput guarantees*, and *soft* or *stochastic throughput guarantees* [36]. The hard throughput guarantees promise a certain throughput over a time-window with unit probability. However, for wireless networks with a varying number of users, applications, and channel conditions, such throughput guarantees can be difficult to fulfill with absolute certainty. Consequently, soft throughput guarantees are more realistic, and thus better suited for wireless networks. The soft throughput guarantees promise a certain throughput within a time-window with a probability that is high, but less than unity.

4 Four Well-Known Opportunistic Scheduling Algorithms

In this section, four basic and well-known opportunistic algorithms for scheduling users in a time-slotted network will be presented. Evaluating the theoretical performance of such simple scheduling algorithms can lead to valuable insights that can be used for developing more advanced algorithms. The four algorithms are designed for networks where only one mobile user is scheduled in each time-slot. Some of these scheduling algorithms provide some measure of fairness in the radio resource allocation between the different users. However, these scheduling algorithms assume that the users always have data to send and do not consider the QoS requirements of the different applications in the network and can therefore not provide exact QoS guarantees. Issues related to how opportunistic scheduling algorithms can be implemented in more complex wireless networks and issues related to how QoS guarantees can be promised to the different applications will be handled in Section 5 and Section 6.

A general scheduling algorithm for scheduling a single user in each time-slot can be formulated as follows:

$$i^*(t_k) = \underset{1 \leq i \leq N}{\operatorname{argmax}} \; x_i(t_k), \tag{1.4}$$

where i^* denotes the index of the selected user, i denotes any user index, and $x_i(t_k)$ denotes the value of the chosen scheduling metric at the beginning of time-slot k. The scheduling metric is the metric that is used to decide which user is going to be scheduled, and for opportunistic scheduling algorithms $x_i(t_k)$ is typically a function of $\gamma_i(t_k)$, the CNR of user i in time-slot k.

4.1 Max Carrier-to-Noise Scheduling (MCS)

The simplest opportunistic scheduling algorithm is the one that schedules the user with the highest CNR in each time-slot [37]:

$$i^*(t_k) = \underset{1 \le i \le N}{\operatorname{argmax}} \ \gamma_i(t_k). \tag{1.5}$$

This is the most opportunistic of all scheduling algorithms for time-slotted networks, since it maximizes the MASSE. This means that the MCS policy maximally exploits the MUD in a time-slotted cell where only one user is scheduled at a time.

4.2 Normalized Carrier-to-Noise Scheduling (NCS)

To increase the fairness of the system, it can be advantageous to schedule the user that has the highest *normalized* CNR in each time-slot [38]:

$$i^*(t_k) = \underset{1 \le i \le N}{\operatorname{argmax}} \ \frac{\gamma_i(t_k)}{\overline{\gamma}_i}, \tag{1.6}$$

where $\overline{\gamma}_i$ is the average CNR of user i.

4.3 Proportional Fair Scheduling (PFS)

The most widely adopted opportunistic scheduling algorithm is the Proportional Fair Scheduling (PFS) algorithm. This algorithm is patented by Qualcomm Incorporated [39] and is also described in [22] and [40]. This algorithm has the following form:

$$i^*(t_k) = \underset{1 \le i \le N}{\operatorname{argmax}} \left(\frac{r_i(t_k)}{T_i(t_k)} \right), \tag{1.7}$$

where $r_i(t_k)$ is the rate of user i in time-slot k, and

$$T_i(t_{k+1}) = \begin{cases} \left(1 - \frac{1}{t_c}\right) T_i(t_k) + \frac{1}{t_c} r_i(t_k) & \text{if } i = i^*(t_k) \\ \left(1 - \frac{1}{t_c}\right) T_i(t_k), & \text{if } i \neq i^*(t_k), \end{cases} \tag{1.8}$$

where t_c [seconds] is a time-window over which T_i is calculated.

4.4 Opportunistic Round Robin Scheduling (ORR)

The original Opportunistic Round Robin (ORR) algorithm was presented in [41]. This original algorithm was stated as a general optimization problem where the throughput over a time-window of N time-slots should be

maximized, subject to the constraint that the N users should get exactly one time-slot each within the time-window. A more practical version of this algorithm was presented in [42], where the users are scheduled in successive rounds of N competitions. For the first time-slot in a round, the user with the highest CNR is chosen. This user is then taken out of the remaining competitions of the round, and for the next time-slot the user with the highest CNR of the remaining users is scheduled. This procedure is repeated until the last round, where the remaining user is scheduled. The scheduling process then starts over again with a new round of N competitions.

For a scenario where the users' average CNRs are spread far apart, the ORR algorithm will be non-opportunistic, i.e., the user with the highest average CNR will always be scheduled in the first time-slot of a round, the user with the second-highest average CNR will be scheduled in the second round, and so on. In this case the ORR algorithm can be combined with the NCS algorithm to yield a higher system spectral efficiency, with the user with the highest ratio $\gamma_i(t_k)/\overline{\gamma}_i$ scheduled in each competition [43]. This algorithm is denoted as the *Normalized ORR* (N-ORR) algorithm.

5 Physical and MAC Layer Design Issues for Opportunistic Scheduling Algorithms

This section discusses several issues related to the design of scheduling algorithms for wireless systems related to the physical and MAC layers, while the next section discusses how the scheduling algorithms can be designed to fulfill the QoS requirements of the applications.

5.1 Multi-Network and Multi-Cell Scheduling Issues

The previous sections have described scheduling algorithms that schedule the mobile users in a cell. However, the same types of algorithms can be used in a distributed manner where the mobile users may schedule the different networks or the different base stations that are accessible.

Multi-Network Scheduling Issues

A mobile user that has the possibility to connect to different networks based on different technologies may experience e.g. better coverage, higher throughput, better QoS, and lower costs. With an increasing number of new wireless technologies being deployed, the market for handsets that is enabled for different wireless standards is therefore likely to grow [44, 45].

The main problem related to such terminals enabled for heterogeneous wireless networks is often how the resources of the different networks should be scheduled to the users. How such multi-network scheduling algorithms can be implemented in a future wireless network architecture is described in [46].

Downlink Multi-Cell Scheduling Issues

Full frequency reuse, i.e., using the same carrier frequencies for all cells in the network, has a potential to increase the total spectral efficiency of a cellular network [47, 48]. However, since inter-cell interference limits the gain from full frequency reuse, the transmit power of the base stations should be controlled. It has been shown that binary transmit power, i.e., the base stations are transmitting at full power or not transmitting at all, can increase the spectral efficiency of CDM-based networks [49]. Based on this binary power allocation, there have recently been suggested some distributed policies for jointly allocating power to cells and scheduling mobile users within the cells [47, 48]. In addition, Kiani et. al have shown that as the number of users per cell increases, the MASSE is maximized if all cells are simultaneously transmitting and MCS is used to schedule the users [48].

Uplink Multi-Cell Scheduling Issues

Fast Cell Selection (FCS) has been proposed for both HSPA and HDR, and is based on selecting the base station that will give the best uplink channel quality for a mobile user [50, 51]. The suggested FCS algorithms mostly select the base station that can provide the best channel quality to a mobile user, and thus can be implemented in a distributed fashion at the mobiles, or centrally in the network. However, Lau has shown that for centralized FCS, this base station scheduling strategy is not optimal when opportunistic scheduling algorithms are used to schedule the users within a cell [52]. In [52], Lau has also suggested a near-optimal centralized uplink multi-cell, multi-user scheduling algorithm.

5.2 Design Issues Related to Different Access Techniques

In the previous sections, only scheduling algorithms for scheduling single users in each time-slot were handled. However, in modern wireless networks it is often possible to schedule more users in each time-slot. This section will discuss some techniques that enable a scheduling algorithm to schedule groups of users in each time-slot.

Scheduling for CDM-Based Systems

For CDM networks, several users can transmit or receive data in each time-slot using codes which are different from user to user [53–56]. Most current CDM-based standards use orthogonal codes, however as already mentioned in Section 3.1, the throughput can be increased significantly by using SCSIC [26]. Consequently, these users will interfere with each other, and *power control* is therefore important to reduce this interference in a CDM network. In addition to exploiting MUD by scheduling the users with the best channel conditions, it is also common to use FCS in the uplink of CDM networks.

Scheduling for OFDM-Based Systems

As opposed to CDM systems that use orthogonal codes, OFDM systems are orthogonal with respect to frequency. For OFDM systems it is also possible for more users to transmit or receive within a time-slot. In OFDM, each carrier frequency is split into some hundreds of sub-carriers, and each sub-carrier in each time-slot can be allocated to different users [57]. Scheduling sub-carriers in this way means that the MUD can be further exploited and that it is easier to fulfill QoS guarantees within a time-window.

Multi-Antenna Scheduling Issues

Systems with multiple antennas at both the transmitter and receiver, are often referred to as MIMO systems [58], and it was shown in [59, 60] that the Shannon capacity of a single MIMO link grows linearly with $\min(N_t, N_r)$, where N_t is the number of antennas at the transmitter and N_r is the number of antennas at the receiver. If only one user is scheduled in each time-slot, gains from *spatial diversity* and/or *spatial multiplexing* can be obtained [61]. The spatial diversity is exploited if the same signal is transmitted on all antennas and can be used to increase the reliability of reception, while the spatial multiplexing gain is achieved by transmitting different signals on each antenna to increase the throughput for a fixed reliability level [61]. It has been shown that it is a trade-off between the spatial diversity gain and the spatial multiplexing gain of a MIMO-system [61]. The spatial diversity in a MIMO system can be exploited by using e.g. Space-Time Block Coding (STBC) where the signals from the different transmit antennas are jointly encoded to obtain better error protection [62]. However, as shown in [63, 64], the gain from a scenario where STBC is used and the user with the best channel conditions is scheduled in each time-slot (MCS), is similar to the MUD gain when MCS is used in a single-input, single-output (SISO)

system, i.e., a system with single antennas both at the base station and at the mobile terminals.

As mentioned in Section 3.1, it is known that the capacity for multiuser systems can only be achieved if signals are transmitted to or from all users at the same time and SCSIC is implemented [25]. Consequently, to jointly exploit the gain from MUD and multiple antennas in MIMO systems, several users have to be scheduled in each time-slot [65]. The Shannon capacity of such a system can be obtained by using *dirty paper coding* (DPC) [66, 67], and it is shown that the multiplicative capacity gain over a MIMO system where only the user with the best channel is scheduled in each time-slot, is approximately $\min(N_t, N)$, where N is the number of mobile users [68]. DPC is a complex technique which is difficult to implement in practical systems. However, it has been shown that systems using BF where a group of users with the best semi-orthogonal channels are scheduled in each time-slot, can come close to the DPC spectral efficiency [65, 69–71].

Scheduling with Embedded Modulation

In [72] Hossain et. al suggest a modulation method for transmitting different information on the downlink to two users being able to detect different-sized quadrature amplitude modulation constellations. Let us for example assume that a combination of 4-QAM and 64-QAM is used. A 4-QAM constellation consists of one symbol in each quadrant of the coordinate system showing the in-phase and quadrature components of the constellation, while a 64-QAM constellation consists of sixteen symbols in each quadrant of this coordinate system. Let us also assume that a user with a low instantaneous CNR can only detect 4-QAM constellations while a user with a good channel quality can detect all the symbols of the 64-QAM constellation. For this scenario, one fourth of the 64-QAM constellation embedded in the 4-QAM constellation can be transmitted, i.e., if a symbol is going to be transmitted in one of the quadrants of the 4-QAM constellation, one of the sixteen 64-QAM symbols in this quadrant can be transmitted instead. Transmitting the symbols in this way, the user with the bad channel quality can receive two bits in each M-QAM constellation (four different symbols), while the user with the good channel quality can simultaneously receive four bits in each constellation (sixteen different symbols).

Based on the work in [72] the same authors evaluate the higher-layer performance of MCS where the two users with the best channel quality are scheduled in each time-slot using embedded modulation [73]. This algorithm is evaluated according to the buffer distribution, the buffer occupancy, the delay distribution and the packet loss probability (PLP), and the

numerical results show that the two-user MCS algorithm performs better than the traditional MCS algorithm, where only one user is scheduled in each time-slot, at the expense of an increase in the average transmit power.

5.3 Scheduling Issues Related to the Physical Characteristics of the Wireless Channel

Scheduling with Induced Channel Fluctuations

Rapid channel fluctuations is important for the scheduler to exploit MUD on a short time-scale in order to increase the throughput and to fulfill the QoS requirements in the system. Such rapid channel fluctuations are naturally induced if the speed of the mobile users increases [74]. However, channel fluctuations can also be induced for a more static scenario. In [22], Viswanath and Tse suggest to transmit the downlink signals on more antennas and vary the phase and amplitude of the different antennas with time. This technique is called *opportunistic beamforming and nulling*.

Scheduling with Channel Prediction

In real-life wireless networks, the CNR of the users is correlated from time-slot to time-slot. Therefore, if the CNR for each user can be predicted for a time-window ahead in time the scheduler has more information available and it can take better scheduling decisions. In [75] a predictive PFS algorithm is proposed, and the results show that the users are scheduled closer to the peaks of their actual CNR, leading to both an increase in MASSE and fairness.

5.4 Energy Efficient Scheduling

Berry and Gallager have shown that there exists a *delay-energy trade-off* for wireless transmission [76]. This means that to reduce the transmission delay in a system, the transmission power has to be increased, and vice versa. The *Lazy scheduling* algorithm exploits this trade-off by reducing the energy per bit transmitted and by lowering the transmission rate so that it takes longer time to transmit each bit [77, 78]. However, this algorithm will in many cases lead to a too long delay for many applications. It can therefore be useful to redesign this algorithm such that the energy usage per bit is instead minimized subject to a delay constraint [77, 78].

In [79] the *inverse-log scheduling* (ILS) algorithm for energy efficient scheduling in TDMA-based sensor networks is introduced. The ILS algorithm allocates transmission time according to the number of bits in the

buffers and channel quality and has a much lower complexity than the optimal scheduling algorithm. Both the centralized and distributed versions of this algorithm obtain a near-optimal energy efficiency with delay constraints in large-scale networks.

5.5 Overhead in Scheduling

According to Wikipedia [80], overhead in computer science is "any combination of excess or indirect computation time, memory, bandwidth, or other resources that are required to be utilized or expanded to enable a particular goal". For wireless communications, two significant sources of overhead are *protocol overhead* and *signaling overhead*. Protocol overhead arises because headers and/or trailers are added to the frames or packets at the different layers of the protocols stack, which reduces the resources available for transmitting information [81]. Signaling overhead or *feedback overhead* arises because information about e. g. channel state or buffer state of different mobile users needs to be distributed in the network [82]. For example, a transmitter can receive CSI estimates from a receiver in order to adapt the transmission coding and modulation to the instantaneous channel state.

Adapting the scheduling decision according to the channel conditions also requires that channel quality estimates are available at the scheduler. In order to make the most precise scheduling decisions, such estimated should obviously be as exact and up to date as possible. However, if the scheduler requests feedback more often than necessary or with a higher precision than necessary, the system will suffer from increased overhead and hence decreased system spectral efficiency. How often feedback should be collected by the scheduler depends on how fast the channel changes, which can be measured in terms of the *coherence time* of the wireless channel, i.e., the time period over which the channel quality can be considered as being constant [83]. The effect of the feedback precision has been investigated in [84, 85]. It is shown that heavy quantization of the CSI estimate being fed back will not lead to a significant reduction of the MUD gain. This means that the CSI does only need to be represented by a few bits. It has also been shown that feedback compression algorithms that exploit channel correlation in time and frequency also can significantly reduce the overhead from feedback [86, 87].

Another approach that has been pursued in order to reduce the feedback overhead is to reduce the feedback load, i.e., the number of users giving feedback. Most algorithms trying to reduce the *feedback load*, defined as the number of users feeding back their CSI estimate for each time-slot,

are based on CNR thresholds. One such algorithm is the *Selective Multiuser Diversity* (SMUD) algorithm proposed by Gesbert and Alouini [88]. This algorithm uses one CNR threshold value and the users that have a instantaneous CNR above this threshold, feed back their CSI estimate to the base station. If no users are above the threshold, a random user is selected to transmit his feedback. Qin and Berry have proposed the *Splitting algorithm*, which is designed for slotted ALOHA networks and is based on binary search [89]. Another threshold-based feedback reduction algorithm is described in [90]. For this algorithm the mobile users are probed randomly until one user is found above a threshold.

For MIMO systems, the transmitter might need to know the whole channel matrix and intuitively the CSI should therefore contain more information than for SISO systems. However, also for MIMO systems it is shown that the CSI being fed back can be heavy quantized and that only a few bits feedback can provide performance close to that with full knowledge at the transmitter [91].

Since there often are hundreds of sub-carriers in OFDM systems, and since each sub-carrier can in principle be scheduled to any user, the amount of feedback information can be significant for such systems. The correlation between the channel quality of adjacent sub-carriers can be exploited to reduce the amount of feedback. By allocating blocks of sub-carriers to each of the users, the scheduler only needs to obtain CSI for each block [92]. The length of the blocks will depend on the frequency selectivity of the wireless channel.

5.6 Scheduling for Ad Hoc Networks

Opportunistic scheduling for ad hoc networks is based on many principles from the cellular scheduling algorithms [31, 93–99]. However, for an ad hoc network, a transmitter schedules different receivers or a receiver schedules different transmitters. For IEEE 802.11-based networks, the communication process can be administered by RTS (request-to-send) and CTS (clear-to-send) packets. The RTS packets are sent from transmitters to potential receivers and when the potential receivers receive RTS packets, they reply with CTS packets. Opportunistic scheduling algorithms are based on channel estimates, hence, the CTS packets need to contain CSI for a scenario where the transmitter performs the scheduling process, while the RTS packets need to contain CSI when the receivers perform the scheduling process.

Ad hoc scheduling algorithms can also be based on a multi-hop scenario where the transmitter uses one or more other mobile users as relays

for the data that are sent to the receiver [7, 98, 100].

Since the mobile terminal typically has limited battery capacity, energy efficient scheduling is important for ad hoc networks. In addition, power control is also important for ad hoc transmission since a transmitter being close to the receiver can drown the signal from transmitters further away [99]. This is known as the *near-far problem*.

User cooperation diversity or just *cooperation diversity* has recently been proposed as a new form of spatial diversity [101, 102]. This type of diversity can be exploited in an ad hoc network where several single-antenna relays cooperate to constitute a virtual antenna array. In [103], the authors propose a scheduling algorithm that can exploit both the MUD and the cooperation diversity of an ad hoc network.

6 Cross-Layer Design Issues for Opportunistic Scheduling Algorithms

Research within the field of scheduling *packets* of *wire-line* networks has matured through extensive research during the last two decades. Much of this research has focused on scheduling algorithms similar to the Weighted Fair Queuing (WFQ) algorithm [104], a packet-based version of Generalized Processor Sharing (GPS) [105]. This is because GPS can guarantee to the different applications (sessions) that the network resources are allocated fairly and independently of the behavior of the other applications [106]. Most of the publications on packet scheduling assume that the throughput of the channel is constant.

For wireless networks, the research has mainly concentrated on how to schedule *radio resources*, e.g. time-slots, frequencies, power, and/or codes, to different mobile users. Most of these scheduling algorithms do not take the users' QoS requirements into account and mainly focuses on how to exploit the time-varying nature of the wireless channels.

Traditionally, the research on packet scheduling has concentrated mostly on QoS and fairness for different QoS classes or different applications, while opportunistic scheduling algorithms have focused on exploiting the time-varying nature of the wireless channels and to provide fairness to the different mobile users. This segregation between packet scheduling and radio resource scheduling is not efficient since none of the two types of scheduling algorithms focus both on (i) providing QoS for the applications and (ii) exploiting the time-varying characteristics of the wireless channel. It is therefore necessary to merge the scheduling of packets and the scheduling of radio resources to design *cross-layer* scheduling algorithms [107].

To be able to improve the QoS experienced by the mobile users, cross-layer scheduling algorithms need to take both the time-varying characteristics of the wireless channels and the QoS demands of the applications into account. In addition, it is often necessary to consider the characteristics of the packet load of (i) the buffers at the mobile users containing packets waiting to be transmitted over the uplink and (ii) the buffers at the base station containing packets waiting to be transmitted on the downlink to each of the users [108]. In this section, cross-layer scheduling algorithms that are designed to improve the QoS in the network will be described. Both *non-queue-aware* and *queue-aware* scheduling algorithms are considered. While non-queue-aware algorithms do not consider how the queues of the buffers can affect the QoS, the queue-aware algorithms consider effects like queuing delay, buffer overflow, and probability of empty buffers.

6.1 Non-Queue-Aware, Cross-Layer Scheduling

Physical and MAC related design issues can be analyzed by assuming that all the users are *back-logged*, i.e., that all the users in the system have non-empty buffers that always contain packets to send or receive. However, when analyzing the QoS performance of scheduling algorithms this assumption is not always correct since the number of packets in the buffers can vary significantly, and there is a relatively high probability that the buffers are empty [108, 109]. However, since the scheduling algorithms in modern cellular networks operate on time-scales that are significantly shorter than the time-scale over which the population of back-logged users change, it can nevertheless be assumed that the scheduling algorithms operate on a constant user population [109].

In [110] Andrews et al. assumed a constant user population and propose scheduling algorithms that aim at offering throughput guarantees by giving different priorities to the users depending on how far they are from fulfilling their throughput guarantees. One of the problems with this algorithm is however that it takes action only when a throughput guarantee already has been violated.

As an alternative, Borst and Whiting proposed a scheduling algorithm that tries to fulfill the throughput guarantees *before* they are violated [111]. This algorithm is also based on assuming a constant user population and is based on a mathematical proof showing that the algorithm provides the highest theoretically attainable throughput guarantees to the mobile users in a cell. This optimal algorithm can be stated as follows:

$$i^*(t_k) = \underset{1 \le i \le N}{\operatorname{argmax}} \left(\frac{r_i(t_k)}{\alpha_i} \right), \tag{1.9}$$

where α_i is a constant. However, in [111] it was not shown how the optimal α_is can be found. An additional drawback of the scheduling policy in (1.9) as proposed in [111], is however that the α_is are found based on the assumption that the throughput guarantees are to be fulfilled for a long time-window containing many time-slots.

6.2 Queue-Aware, Channel-Aware, Cross-Layer Scheduling

For time-slotted networks, the packets in the buffers are aggregated into time-slots. Consequently, empty buffers and partially filled time-slots will affect the system performance. In the recent years some publications have considered how to integrate the packet scheduling and the radio resource scheduling into queue-aware, channel-aware scheduling algorithms [34, 107, 109, 112–116]. For example, one such publication handles how to implement Weighted Fair Queuing (WFQ) when the largest share of the radio resources is given to the users with the instantaneously best channel conditions in a CDM-based network [56]. However, maybe the most well-known queue-aware, channel-aware scheduling algorithm is the Modified Largest Weighted Delay First (M-LWDF) algorithm [34, 117]:

$$i^*(t_k) = \underset{1 \le i \le N}{\operatorname{argmax}} \left(\phi_i W_i(t_k) \frac{r_i(t_k)}{\overline{r_i}} \right), \tag{1.10}$$

where $W_i(t_k)$ [seconds] is the head-of-the-line (HOL) packet delay in user i's buffer, ϕ_i is a constant denoting the priority given to user i, and $\overline{r_i}$ [bits/second] is the average rate for user i. This algorithm can be used both for uplink and downlink scheduling since W_i can denote the delay of the HOL packets in either the users' output buffers on the uplink or the buffers at the base station containing packets for downlink transmission to each of the mobile users. The advantage of this algorithm is that it takes both the channel quality and the delay of the packets into account when performing scheduling. In addition, this algorithm is proven to be *throughput optimal*. This means that the algorithm manages to keep the queues stable if this is at all feasible to do with any other algorithm, where a stable queue is defined as having a finite expected queue length. The M-LWDF algorithm can also be reformulated to guarantee a certain throughput to the users if it is used in conjunction with a token bucket control [34].

Another well-known queue-aware, channel-aware scheduling algorithm is the *exponential rule* developed by Shakkottai and Stolyar [113]. This scheduling algorithm is also proved to be throughput optimal and can also be used to provide QoS guarantees in a cellular network [118, 119].

Wu and Negi introduced the concept of *effective capacity* link model which is a model that unites the dynamics of the wireless channel and the dynamics of the packet queuing [108]. Although it is hard to develop closed-form expressions for the effective capacity for CNR distributions like Rayleigh and Rice, this model can be used as a tool to develop scheduling algorithms with QoS guarantees [120].

In [107], a general queue-aware, channel-aware scheduling algorithm providing QoS guarantees is developed. It is also thoroughly described how the adaptive coding and modulation and the scheduling algorithm is going to be implemented at the MAC layer of a IEEE 802.16-based network.

7 Contributions of the Included Papers

This thesis consists of six papers. In this section a brief summary of these papers is presented.

Paper A

Vegard Hassel, Mohamed-Slim Alouini, Geir. E. Øien, and David Gesbert, "Rate-Optimal Multiuser Scheduling with Reduced Feedback Load and Analysis of Delay Effects," published in *EURASIP Journal on Wireless Communications and Networking, Special Issue on Radio Resource Management in 3G+ Systems*, 2006.
http://www.hindawi.com/GetArticle.aspx?doi=10.1155/WCN/2006/36424

This paper is partially based on the conference paper in [121]. In Paper A, we assume that the MCS algorithm is used, and develop a new feedback algorithm for cellular networks as well as investigating the effects of delayed CSI feedback. As for the SMUD algorithm [88], the base station requests CSI feedback from the mobile users that have a CNR above a threshold value. However, as opposed to the SMUD algorithm which schedules a random user if no feedback is received, our algorithm requests feedback from all the users if no users' CNRs are above the CNR threshold. The advantage of this algorithm is that we can analytically find the CNR threshold value that minimizes the feedback load for a given number of users N. Our numerical results show that these optimal threshold values decrease the feedback load significantly for a large number of users in a cell.

The feedback collection process described above introduces increased delays in the system. Two types of delay are evaluated, namely, *scheduling delay* and *outdated CSI estimates*. Scheduling delay arises when the base station bases its scheduling decision on delayed CSI estimates while the

adaptive modulation and coding are based on updated CSI estimates. Since the scheduling decision is based on old CSI estimates, it is likely that the user with the best channel conditions is not always scheduled. This will reduce the MASSE and for large delays the system will not experience any MUD gain. When the base station uses the same delayed CSI estimates both to (i) take the scheduling decision and (ii) adapt the modulation and coding to the channel quality, the MASSE will not be affected. However, it is likely that the bit error rate (BER) will be affected since the CNR of the scheduled user may have dropped since the CSI was estimated and the modulation constellation used thus cannot be transmitted at the target BER. Our numerical results related to the delay analysis show that the system is able to perform without MASSE or BER degradation when the delays are below certain critical values that will depend on the Doppler frequency shift.

Paper B

Vegard Hassel, David Gesbert, Mohamed-Slim Alouini, and Geir E. Øien, "A Threshold-Based Channel State Feedback Algorithm for Modern Cellular Systems," accepted for publication in *IEEE Transactions on Wireless Communications*.

This paper is partially based on the conference paper in [122]. In Paper B, we analyze a feedback algorithm that is a generalization of the feedback algorithm proposed in Paper A. This algorithm uses L feedback thresholds and as opposed to the feedback algorithm in Paper A, the generalized feedback algorithm can be adapted to any scheduling algorithm. For many scheduling algorithms, e.g. PFS, the L feedback thresholds that minimize the feedback load have to be found numerically. However, for the MCS and NCS algorithms we find closed-form expressions for the feedback load and for the threshold values for a given number of thresholds L and given number of mobile users N. Our numerical results show that by employing just a few threshold values, there is a high probability of obtaining feedback from just the user that the scheduling algorithm wants to schedule. We also propose a two-step procedure that makes it possible to obtain both the optimal threshold values and the optimal number of thresholds.

Paper C

Vegard Hassel, Hend Koubaa, and Geir E. Øien, "Feedback Protocols for Increased Multiuser Diversity Gain in Cellular ALOHA-Based

Networks–A Comparative Study," Technical Report, NTNU, published at *http://www.diva-portal.org/ntnu/*, January 2007.

This report is partially based on the conference papers in [123] and [124]. In Paper C, we investigate the performance of the feedback algorithm in Paper B for a time-slotted IEEE 802.11-based network using the MCS algorithm. We propose three novel protocols for implementing the feedback algorithm and we develop closed-form expressions for the *guard time*, i.e., the time used to collect feedback for each time-slot, and the MASSE for each of these protocols. Our numerical results show that we can obtain significant MASSE gains compared to a feedback protocol that collects feedback from all the mobile users, when there are many users in the cell. However, the performance of the most efficient feedback protocols have a similar performance as the *Splitting algorithm* proposed in [89] and thus do not improve significantly on previous work by other researchers.

Paper D

Vegard Hassel, Marius Røed Hanssen, and Geir E. Øien, "Spectral Efficiency and Fairness for Opportunistic Scheduling Algorithms," submitted to *IEEE Transactions on Wireless Communications*, May 2006.

This paper is partially based on the conference paper in [125]. In Paper D, we develop a closed-form expression for the MASSE of the N-ORR algorithm. We also define the time-slot fairness and throughput fairness based on JFI [27], where JFI equal to zero denotes full unfairness and JFI equal to unity corresponds to full fairness. These two types of fairness are calculated for a time-window of K time-slots. We evaluate asymptotic fairness when K goes to infinity and we show that all scheduling algorithms that have the same probability of scheduling all users in a time-slot, will have full time-slot fairness as K goes to infinity. The corresponding asymptotic throughput fairness will only converge to unity when the channels of the users are non-fading, i.e., the users have constant CNRs, and the CNRs are the same for all the users.

We also develop closed-form expressions for the time-slot fairness and the throughput fairness as a function of K for the Round Robin (RR), MCS, NCS, and N-ORR algorithms. Our numerical results show that the N-ORR algorithm is best suited for obtaining throughput fairness for short time-windows K while the NCS algorithm is best suited for obtaining throughput fairness for long values of K.

Paper E

Vegard Hassel, Geir E. Øien, and David Gesbert, "Throughput Guarantees for Opportunistic Scheduling: A Comparative Study," accepted for publication in *IEEE Transactions on Wireless Communications*.

This paper is partially based on the conference papers in [126] and [127]. In Paper E, we develop a general formula for the approximate throughput guarantee violation probability (TGVP) of any scheduling algorithm. The TGVP quantifies the probability that a certain throughput guarantee is violated in a cellular network. We also derive closed-form expressions for the approximate TGVP for RR, MCS, NCS, and N-ORR. Such analytical expressions make it possible to calculate the QoS for the mobile users in a time-efficient way for a set of instantaneous system parameters, and the TGVP expressions can therefore be used directly in a RRM system for networks carrying real-time traffic and where the users move around at high speed. Monte Carlo simulations for two real-life wireless standards showed that our TGVP approximations are tight.

Paper F

Vegard Hassel, Sébastien de la Kethulle de Ryhove, and Geir E. Øien, "Scheduling Algorithms for Increased Throughput Guarantees in Wireless Networks,", submitted to *Workshop on Resource Allocation in Wireless Networks (RAWNET'07)*, Limassol, Cyprus, April 2007.

In Paper F, we base our work on the scheduling algorithm proposed by Borst and Whiting that obtains optimal throughput guarantees for long time-windows [111]. We find analytical expressions for how to obtain the parameters of this algorithm and we improve the performance of the algorithm by adapting the priorities of the users to the amount of bits that has already been allocated in a time-window. Our numerical results show that our adaptive algorithm has the theoretical potential of at least doubling the throughput guarantees that can be given in wireless networks based on real-life cellular standards.

8 Main Contributions of the Thesis

The main contributions of the thesis can now be summarized to be:

- The effects of CSI feedback delay on MASSE and BER are analyzed and it is shown that the performance of the MCS algorithm is not

degraded significantly if the normalized delay is kept below certain critical values.
- A general feedback algorithm with multiple feedback thresholds is developed. This algorithm can be adapted to the scheduling metric of any scheduling algorithm, and it is shown that for this algorithm it is a high probability that CSI feedback is only received from the user that the system wants to schedule.
- The performance of the general feedback algorithm is shown to be similar to the performance of the Splitting algorithm in practical cellular, time-slotted IEEE 802.11 networks.
- Closed-form expressions for the time-slot fairness and throughput fairness of the RR, MCS, NCS, and ORR algorithms are developed.
- A general expression for the approximate TGVP for any scheduling algorithm is derived and the corresponding closed-form expressions for the approximate TGVP of the RR, MCS, NCS, and ORR algorithms are developed.
- Expressions for obtaining the optimal parameters for the scheduling algorithm proposed in [111] are developed. An adaptive version of the optimal algorithm is developed and this adaptive algorithm has the theoretical potential of at least doubling the throughput guarantees that can be given in modern cellular networks.

9 Suggestions for Future Research

In this section we list some topics that can be interesting to investigate in the future:

- The Splitting algorithm in [89] can be generalized to have feedback threshold values adapted to any scheduling algorithm as done in Paper B.
- The analysis in Paper F does only evaluate the performance of the scheduling algorithms when only one user can be scheduled in each time-slot. For CDM or OFDM based networks [3], the throughput guarantee performance can be further improved by scheduling more than one user in each time-slot.
- Only a few practical scheduling algorithms for networks using beamforming have been developed [128–131]. We therefore think that such algorithms can be further developed. One concrete idea is to develop a scheduling algorithm for beamforming networks using the ideas from Paper F.

- Probably, the highest theoretically attainable throughput guarantees that can be obtained by a scheduling algorithm in a MIMO system will be based on DPC [66, 67]. By extending the results of Paper F, it should be possible to obtain this optimal scheduling algorithm.
- The assumption that the user population can be regarded as constant over the time-window over which the throughput guarantees are calculated is sometimes not completely correct. By using analytical methods based on *queuing theory* like the ones in [74, 109], the performance of the algorithm in Paper F can be investigated in further detail.
- Admission control algorithms are often based on the call dropping probability (CDP) of the system [23]. For networks carrying real-time applications, the CDP is closely linked to the TGVP and to the motion pattern of the users. Developing analytical expressions for the CDP can be helpful when implementing practical admission control algorithms.
- As far as we know, the performance of systems using joint power adaptation, rate adaptation and opportunistic scheduling has only been analyzed for continuous rates [132]. We therefore suggest to investigate the effects of discrete rates in the system. Some initial work has already been conducted for joint power and rate adaptation with discrete rates [133].
- The true performance of opportunistic scheduling in real-life networks can probably be more deeply analyzed by using network simulators like *ns2* and *OpNet*. It would be especially interesting to investigate the relationship between packet scheduling and opportunistic scheduling of time-slots.
- The field of *Network calculus* [134, 135] is mostly developed for quantifying QoS guarantees in wire-line networks. Only a few publications have handled network calculus for wireless networks [136, 137], and we think developing this powerful tool further for wireless networks can lead to new insights.

References

[1] W. Stallings, *Wireless Communications and Networks*.
Prentice-Hall, Inc., 2002.

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

[3] WiMAX Forum, "Mobile WiMAX – Part II: A Comparative Analysis." `http://www.wimaxforum.org/news/downloads/Mobile_WiMAX_Part2_Comparative_Analysis.pdf`, May 2006.

[4] J. Zander, "Radio resource management in future wireless networks: Requirements and limitations," *IEEE Communications Magazine*, vol. 35, pp. 30–36, Aug. 1997.

[5] P. Magnusson, J. Lundsjö, J. Sachs, and P. Wallentin, "Radio resource management distribution in a beyond 3G multi-radio access architecture," in *Proc. IEEE Global Communications Conference (GLOBECOM'04)*, vol. 6, pp. 3472–3477, Nov.-Dec. 2004.

[6] A. Hills and B. Friday, "Radio resource management in wireless LANs," *IEEE Radio Communications*, vol. 42, pp. 9–14, Dec. 2004.

[7] H. Fattah and C. Leung, "An overview of scheduling algorithms in wireless multimedia networks," *IEEE Wireless Communications*, vol. 9, pp. 76–83, Oct. 2002.

[8] M. Andrews, "A survey of scheduling theory in wireless data networks." `http://cm.bell-labs.com/cm/ms/who/andrews/ima.ps`. University of Minnesota, Institute of Mathematics and Its Applications' Summer Program, June-July 2005.

[9] M. K. Simon and M.-S. Alouini, *Digital Communication over Fading Channels*.
John Wiley & Sons, Inc., 2000.

[10] J. F. Hayes, "Adaptive feedback communications," *IEEE Trans. on Communications*, vol. 16, pp. 29–34, Feb. 1968.

[11] J. K. Cavers, "Variable-rate transmission for Rayleigh fading channels," *IEEE Trans. on Communications*, vol. COM-20, pp. 74–83, Feb. 1972.

[12] W. T. Webb and R. Steele, "Variable-rate QAM for mobile radio," *IEEE Trans. on Communications*, vol. 43, pp. 2223–2230, July 1995.

[13] A. J. Goldsmith and S.-G. Chua, "Variable-rate variable-power M-QAM for fading channels," *IEEE Trans. on Communications*, vol. COM-45, pp. 1218–1230, Oct. 1997.

[14] M.-S. Alouini and A. J. Goldsmith, "Adaptive modulation over Nakagami fading channels," *Kluwer Journal on Wireless Communications*, vol. 13, pp. 119–143, May 2000.

[15] K. J. Hole, H. Holm, and G. E. Øien, "Adaptive multi-dimensional coded modulation over flat fading channels," *IEEE Journal on Selected Areas in Communication*, vol. 18, pp. 1153–1158, July 2000.

[16] A. Gjendemsjø, H.-C. Yang, M.-S. Alouini, and G. E. Øien, "Joint adaptive transmission and combining with optimized rate and power allocation," in *Proc. IEEE Workshop on Signal Processing Advances in Wireless Communications (SPAWC'06)*, (Cannes, France), July 2006.

[17] B. Sklar, *Digital Communications – Fundamentals and Applications*. Prentice-Hall, Inc., 2nd ed., 2001.

[18] J. K. Cavers, "An analysis of pilot symbol assisted modulaation for rayleigh fading channels," *IEEE Trans. on Vehicular Technology*, vol. 40, pp. 686–693, Nov. 1991.

[19] D. V. Duong and G. E. Øien, "Optimal pilot spacing and power in rate-adaptive MIMO diversity system with imperfect CSI." To appear in *IEEE Trans. on Wireless Comm.*

[20] G. S. Smith, "A direct derivation of a single-antenna reciprocity relation for the time domain," *IEEE Trans. Antennas and Propagation*, vol. 52, pp. 1568–1577, June 2004.

[21] M. Sternad and D. Aronsson, "Channel estimation and prediction for adaptive OFDMA/TDMA uplinks, based on overlapping pilots," in *IEEE Int. Conference on Acoustics, Speech, and Signal Proc. (ICASSP'05)*, vol. 3, (Philadelphia, PA, USA), pp. III–861 – III–864, Mar. 2005.

[22] P. Viswanath, D. N. C. Tse, and R. Laroia, "Opportunistic beamforming using dumb antennas," *IEEE Trans. on Information Theory*, vol. 48, pp. 1277–1294, June 2002.

[23] M. Ghaderi and R. Boutaba, "Call admission control in mobile cellular networks: a comprehensive survey," *Wireless Communications & Mobile Computing*, vol. 6, pp. 69–93, Feb. 2006.

[24] A. J. Goldsmith and P. P. Varaiya, "Capacity of fading channels with channel side information," *IEEE Trans. on Information Theory*, vol. IT-43, pp. 1896–1992, Nov. 1997.

[25] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: John Wiley & Sons, Inc., 1991.

[26] M. F. Madkour, S. C. Gupta, and Y.-P. E. Wang, "Successive interference cancellation algorithms for downlink W-CDMA communications," *IEEE Trans. on Wireless Communications*, vol. 1, pp. 169–177, Jan. 2002.

[27] R. Jain, D. Chiu, and W. Hawe, "A quantitative measure of fairness and discrimination for resource allocation in shared computer systems," DEC Research Report TR-301, Digital Equipment Corporation, Maynard, MA, USA, Sept. 1984.

[28] H. Sirisena, A. Haider, M. Hassan, and K. Fawlikowski, "Transient fairness of optimized end-to-end window control," in *Proc. IEEE Global Communications Conference (GLOBECOM'03)*, (San Francisco, CA, USA), pp. 3979 – 3983, Dec. 2003.

[29] G. Berger-Sabbatel, A. Duda, O. Gaudoin, M. Heusse, and F. Rousseau, "Fairness and its impact on delay in 802.11 networks," in *Proc. of the IEEE Global Communications Conference (GLOBECOM'04)*, (Dallas, TX, USA), pp. 2967–2973, Dec. 2004.

[30] H. J. Bang, "Advanced scheduling techniques for wireless data networks." Master Thesis, Department of Physics, University of Oslo, Feb. 2005.

[31] A. J. Goldsmith and S. B. Wicker, "Design challenges for energy-constrained ad hoc wireless networks," *IEEE Wireless Communications*, vol. 9, pp. 8–27, Aug. 2002.

[32] F. Fluckiger, *Understanding Networked Multimedia: Applications and Technology*.
Englewood Cliffs, New Jersey: Prentice-Hall, Inc., 1995.

[33] M. Motani and A. T. Hoang, "An instance of multiuser diversity in wireless networks," in *Proc. IEEE Int. Symposium on Information Theory (ISIT'02)*, (Lausanne, Switzerland), pp. 443–448, June-July 2002.

[34] 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*, pp. 150–154, Feb. 2001.

[35] P. Hosein and R. Vannithamby, "On flow control and scheduling in time-shared wireless packet data channels," in *Proc. IEEE Vehicular Technology Conference (VTC'05-fall)*, vol. 1, (Dallas, TX, USA), pp. 216–220, Sept. 2005.

[36] D. Ferrari, "Client requirements for real-time communication services," *IEEE Communications Magazine*, vol. 28, pp. 65–72, Nov. 1990.

[37] R. Knopp and P. A. Humblet, "Information capacity and power control in single cell multiuser communications," in *Proc. IEEE Int. Conference on Communications (ICC'95)*, (Seattle, WA, USA), pp. 331–335, June 1995.

[38] L. Yang and M.-S. Alouini, "Performance analysis of multiuser selection diversity," in *Proc. IEEE Int. Conference on Communications (ICC'04)*, (Paris, France), pp. 3066–3070, June 2004.

[39] E. F. Chaponniere, P. J. Black, J. M. Holtzman, and D. N. C. Tse, "Transmitter directed code division multiple access system using path diversity to equitably maximize throughput." U.S. Patent #6449490, `http://www.eecs.berkeley.edu/~dtse/scheduler_patent.html`, Sept. 2002.

[40] J. M. Holtzman, "Asymptotic analysis of proportional fair algorithm," in *Proc. IEEE Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC'01)*, vol. 2, (San Diego, CA, USA), pp. F–33–F–37, Sept. 2001.

[41] S. S. Kulkarni and C. Rosenberg, "Opportunistic scheduling policies for wireless systems with short term fairness constraints," in *Proc. IEEE Global Communications Conference (GLOBECOM'03)*, (San Francisco, CA, USA), pp. 533–537, Dec. 2003.

[42] M. Johansson, "Diversity-enhanced equal access - considerable throughput gains with 1-bit feedback," in *Proc. IEEE Workshop on Signal Processing Advances in Wireless Communications (SPAWC'04)*, (Lisbon, Portugal), July 2004.

[43] Z. Ji, Y. Yang, J. Zhou, M.Takai, and R. Bagrodia, "Exploiting medium access diversity in rate adaptive wireless LANs," in *Proc. ACM Int. Conference on Mobile Computing and Networking (MOBICOM'04)*, (Philadephia, PA, USA), pp. 345–359, Sept.-Oct. 2004.

[44] A. Furuskär and J. Zander, "Multiservice allocation for multiaccess wireless systems," *IEEE Trans. on Wireless Communications*, vol. 4, pp. 174–184, Jan. 2005.

[45] F. Berggren, A. Bria, L. Badia, I. Karla, R. Litjens, P. Magnusson, F. Meago, H. Tang, and R. Veronesi, "Multi-radio resource management for ambient networks," in *Proc. IEEE Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC'05)*, vol. 2, (Helsinki, Finland), pp. 942–946, Sept. 2005.

[46] J. Sachs, H. Wiemann, P. Magnusson, P. Wallentin, and J. Lundsjö, "A generic link layer in a beyond 3G multi-radio access architecture," in *Proc. Int. Conference on Communications, Circuits and Systems (ICCCAS'04)*, vol. 1, (Chengdu, P. R. China), pp. 447–451, June 2004.

[47] J. E. Kirkebø, D. Gesbert, and S. G. Kiani, "Maximizing the capacity of wireless networks using multi-cell access schemes," in *Proc. IEEE Workshop on Signal Processing Advances in Wireless Communications (SPAWC'06)*, (Cannes, France), July 2006.

[48] S. G. Kiani, D. Gesbert, J. E. Kirkebø, A. Gjendemsjø, and G. E. Øien, "A simple greedy scheme for multicell capacity maximization," in *Proc. Int. Telecommunications Symposium (ITS'06)*, (Fortaleza, Brazil), pp. 331–335, Aug. 2006.

[49] A. Gjendemsjø, D. Gesbert, G. E. Øien, and S. G. Kiani, "Optimal power allocation and scheduling for two-cell capacity maximization," in *Proc. IEEE RAWNET (WiOpt'06)*, (Boston, MA, USA), pp. 1–6, Apr. 2006.

[50] P. Bender, P. Black, M. Grob, R. Padovani, N. Sindhushyana, and S. Viterbi, "CDMA/HDR: a bandwidth efficient high speed wireless data service for nomadic users," *IEEE Communications Magazine*, vol. 38, pp. 70–77, July 2000.

[51] A. Das, K. Balachandran, F. Khan, A. Sampath, and H.-J. Su, "Network controlled cell selection for the high speed downlink packet access in UMTS," in *Proc. IEEE Wireless Communications and Networking Conference (WCNC'04)*, (Atlanta, GA, USA), pp. 1975–1979, Mar. 2004.

[52] V. K. N. Lau, "On the macroscopic optimization of multicell wireless systems with multiuser detection and multiple antennas – uplink analysis," *IEEE Trans. on Wireless Communications*, vol. 4, pp. 1388–1393, July 2005.

[53] A. Sampath, P. S. Kumar, and J. M. Holtzman, "Power control and resource management for a multimedia CDMA wireless system," in *Proc. IEEE Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC'95)*, (Toronto, Canada), pp. 21–25, Sept. 1995.

[54] M. Elaoud and P. Ramanathan, "Adaptive allocation of CDMA resources for network level QoS assurances," in *Proc. ACM Int. Conference on Mobile Computing and Networking (MOBICOM'00)*, (Boston, MA, USA), pp. 191–199, Aug. 2000.

[55] N. Bambos and S. Kandukuri, "Power-controlled multiple access schemes for next-generation wireless packet networks," *IEEE Communications Magazine*, vol. 9, pp. 58–64, June 2002.

[56] A. Stamoulis, N. D. Sidiropoulos, and G. B. Giannakis, "Time-varying fair queueing scheduling for multicode CDMA based on dynamic programming," *IEEE Trans. on Wireless Communications*, vol. 3, pp. 512–523, Mar. 2004.

[57] C. Y. Wong, R. S. Cheng, K. B. Letaief, and R. D. Murch, "Multiuser OFDM with adaptive subcarrier, bit, and power allocation," *IEEE Journal on Selected Areas in Communication*, vol. 17, pp. 1747–1758, Oct. 1999.

[58] D. Gesbert, M. Shafi, D. Shiu, P. J. Smith, and A. Naguib, "From theory to practice: An overview of MIMO space-time coded wireless systems," *IEEE Journal on Selected Areas in Communication*, vol. 21, pp. 281–302, Apr. 2003.

[59] E. Telatar, "Capacity of multiantenna Gaussian channels," *AT&T Bell Laboratories, Tech. Memo*, June 1995.

[60] G. J. Foschini and M. J. Gans, "On limits of wireless communications in a fading environment when using multiple antennas," *Wireless Personal Communications*, vol. 6, pp. 311–335, Mar. 1998.

[61] D. Tse, "Diversity–multiplexing tradeoff in multiple-access channels," *IEEE Trans. on Information Theory*, pp. 1859–1874, Sept. 2004.

[62] S. M. Alamouti, "A simple transmit diversity scheme for wireless communications," *IEEE Journal on Selected Areas in Communication*, vol. 16, pp. 1451–1458, Oct. 1998.

[63] R. Gozali, R. M. Buehrer, and B. Woerner, "The impact of multiuser diversity on space-time block coding," *IEEE Communications Letters*, vol. 7, pp. 213–215, May 2003.

[64] E. G. Larsson, "On the combination of spatial diversity and multiuser diversity," *IEEE Communications Letters*, vol. 8, pp. 517–519, Aug. 2004.

[65] K. P. Jagannathan, S. Borst, P. Whiting, and E. Modiano, "Efficient scheduling of multi-user multi-antenna systems," in *Int. Symposium on Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks (WiOpt'06)*, (Boston, MA, USA), pp. 1–8, Apr. 2006.

[66] G. Caire and S. Shamai, "On the achievable throughput of a multiantenna Gaussian broadcast channel," *IEEE Trans. on Information Theory*, vol. 49, pp. 1691–1706, July 2003.

[67] P. Viswanath and D. N. C. Tse, "Sum capacity of the vector Gaussian broadcast channel and uplink-downlink duality," *IEEE Trans. on Information Theory*, vol. 49, pp. 1912–1921, Aug. 2003.

[68] N. Jindal and A. J. Goldsmith, "Dirty-paper coding versus TDMA for MIMO broadcast channels," *IEEE Trans. on Information Theory*, vol. 51, pp. 1783–1794, May 2005.

[69] C. B. Peel, Q. Spencer, A. L. Swindlehurst, and B. Hochwald, "Downlink transmit beamforming in multi-user MIMO systems," in *Proc. IEEE Sensor Array and Multichannel Signal Processing Workshop (SAM'04)*, (Barcelona, Spain), pp. 43–51, May 2004.

[70] T. Yoo and A. J. Goldsmith, "Optimality of zero-forcing beamforming with multiuser diversity," in *Proc. IEEE Int. Conference on Communications (ICC'05)*, (Seoul, Korea), pp. 542–546, June 2005.

[71] M. Sharif and B. Hassibi, "On the capacity of MIMO broadcast channels with partial side information," *IEEE Trans. on Information Theory*, vol. 51, pp. 506–522, Feb. 2005.

[72] J. Hossain, M.-S. Alouini, and V. K. Bhargava, "Real-time multiresolution data transmission over correlated fading channels using hierarchical constellations," in *Proc. IEEE Vehicular Technology Conference (VTC'06-spring)*, vol. 5, (Melbourne, Australia), pp. 2068–2072, May 2006.

[73] J. Hossain, M.-S. Alouini, and V. K. Bhargava, "Higher layer performance study of power-controlled hierarchical constellation-based multi-user opportunistic scheduling," in *Proc. IEEE Global Communications Conference (GLOBECOM'06)*, (San Francisco), Nov.-Dec. 2006.

[74] T. Bonald, S. C. Borst, and A. Proutière, "How mobility impacts the flow-level performance of wireless data systems," in *Proc. IEEE Joint Conference of the Computer and Communications Societies (INFOCOM'04)*, vol. 3, (Hong Kong, P. R. China), pp. 1872–1881, Mar. 2004.

[75] H. J. Bang, T. Ekman, and D. Gesbert, "A channel predictive proportional fair scheduling algorithm," in *Proc. IEEE Workshop on Signal Processing Advances in Wireless Communications (SPAWC'05)*, (New York, NY, USA), pp. 620–624, June 2005.

[76] R. A. Berry and R. G. Gallager, "Communication over fading channels with delay constraints," *IEEE Trans. on Information Theory*, vol. 48, pp. 1135–1149, May 2002.

[77] B. Prabhakar, E. Uysal-Biyikoglu, and A. E. Gamal, "Energy-efficient transmission over a wireless link via lazy packet scheduling," in *Proc. IEEE Joint Conference of the Computer and Communications Societies (INFOCOM'01)*, (Anchorage, AK, USA), pp. 386–394, Apr. 2001.

[78] E. Uysal-Biyikoglu, B. Prabhakar, and A. E. Gamal, "Energy-efficient packet transmission over a wireless link," *IEEE/ACM Trans. on Networking*, vol. 10, pp. 487–499, Aug. 2002.

[79] Y. Yao and G. B. Giannakis, "Energy-efficient scheduling for wireless sensor networks," *IEEE Trans. on Communications*, vol. 53, pp. 1333–1342, Aug. 2005.

[80] Wikipedia, "Computational overhead — Wikipedia, the free encyclopedia," 2006.
`http://en.wikipedia.org/wiki/Computational_overhead`.

[81] J. D. Cavanaugh, "Protocol overhead in IP/ATM networks." `http://www.sonic.net/support/docs/ip-atm.overhead.pdf` Minnesota Supercomputer Center, Inc., Aug. 1994.

[82] M. Codreanu, D. Tujkovic, and M. Latva-aho, "Adaptive MIMO-OFDM with low signalling overhead for unbalanced antenna systems," in *Proc. IEEE Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC'04)*, vol. 4, (Barcelona, Spain), pp. 2382–2386, Sept. 2004.

[83] J. G. Proakis, *Digital Communications*.
New York: McGraw-Hill, 3rd ed., 1995.

[84] F. Florén, O. Edfors, and B.-A. Molin, "The effect of feedback quantization on the throughput of a multiuser diversity scheme," in *Proc. IEEE Global Communications Conference (GLOBECOM'03)*, vol. 1, (San Francisco, CA, USA), pp. 497–501, Dec. 2003.

[85] M. Johansson, "Benefits of multiuser diversity with limited feedback," in *Proc. IEEE Workshop on Signal Processing Advances in Wireless Communications (SPAWC'03)*, (Rome, Italy), pp. 155–159, June 2003.

[86] H. Cheon, B. Park, and D. Hong, "Adaptive multicarrier system with reduced feedback information in wideband radio channels," in *IEEE Vehicular Technology Conference (VTC'99-fall)*, vol. 5, (Amsterdam, The Netherlands), pp. 2880–2884, Sept. 1999.

[87] H. Nguyen, J. Brouet, V. Kumar, and T. Lestable, "Compression of associated signaling for adaptive multi-carrier systems," in *IEEE Vehicular Technology Conference (VTC'04-spring)*, vol. 4, (Rome, Italy), pp. 1916–1919, May 2004.

[88] D. Gesbert and M.-S. Alouini, "How much feedback is multi-user diversity really worth?," in *Proc. IEEE Int. Conference on Communications (ICC'04)*, (Paris, France), pp. 234–238, June 2004.

[89] X. Qin. and R. Berry, "Opportunistic splitting algorithms for wireless networks," in *Proc. IEEE Int. Conference on Computer Communications (INFOCOM'04)*, (Hong Kong, P. R. China), pp. 1662 – 1672, Mar. 2004.

[90] Y. Al-Harthi, A. Tewfik, and M.-S. Alouini, "Opportunistic scheduling with quantized feedback in wireless networks," in *Proc. Int. Conference on Information Technology: Coding and Computing (ITCC'05)*, (Las Vegas, NV, USA), Apr. 2005.

[91] D. J. Love, R. W. H. Jr., W. Santipach, and M. L. Honig, "What is the value of limited feedback for MIMO channels?," *IEEE Communications Magazine*, vol. 42, pp. 54–59, Oct. 2004.

[92] J. Oh and J. M. Cioffi, "Sub-bandrate and power control for wireless OFDM," in *Proc. IEEE Vehicular Technology Conference (VTC'04-fall)*, vol. 3, (Los Angeles, CA, USA), pp. 2011–2014, Sept. 2004.

[93] P. Bhagwat, P. Bhattacharya, A. Krishna, and S. K. Tripathi, "Enhancing throughput over wireless LANs using channel state dependent packet scheduling," in *Proc. IEEE Joint Conference of the Computer and Communications Societies (INFOCOM'96)*, (San Francisco, CA, USA), pp. 1133–1140, Mar. 1996.

[94] C. Fragouli, V. Sivaraman, and M. Srivastava, "Controlled multimedia wireless link sharing via enhanced class-based queuing with channel-state dependent packet scheduling," in *Proc. IEEE Joint Conference of the Computer and Communications Societies (INFOCOM'98)*, (San Francisco, CA, USA), pp. 572–580, Mar. 1998.

[95] G. Holland, N. Vaidya, and P. Bahl, "A rate-adaptive MAC protocol for multi-hop wireless networks," in *Proc. ACM Int. Conference on Mobile Computing and Networking (MOBICOM'01)*, (Rome, Italy), pp. 236–251, July 2001.

[96] B. Sadeghi, V. Kanodia, A. Sabharwal, and E. Knightly, "Opportunistic media access for multirate ad hoc networks," in *Proc. ACM Int. Conference on Mobile Computing and Networking (MOBICOM'02)*, (Atlanta, GA, USA), pp. 24–35, Sept. 2002.

[97] V. Kanodia, A. Sabharwal, and E. Knightly, "MOAR: A multi-channel opportunistic auto-rate media access protocol for ad hoc networks," in *Proc. ACM Int. Conference on Mobile Computing and Net-*

working (MOBICOM'04), (Philadephia, PA, USA), pp. 24–35, Sept. 2004.

[98] J. Wang, H. Zhai, and Y. Fang, "Opportunistic packet scheduling and media access control for wireless LANs and multi-hop ad hoc networks," in *Proc. IEEE Wireless Communications and Networking Conference (WCNC'04)*, (Atlanta, GA, USA), pp. 1234–1239, Mar. 2004.

[99] Y. Liu, Y.-K. Kwok, and J. Wang, "Optimal power control and opportunistic fair scheduling in TH-PPM UWB ad-hoc multimedia networks," in *Proc. IEEE Wireless Communications and Networking Conference (WCNC'06)*, pp. 1693–1698, Apr. 2006.

[100] M. Hu and J. Zhang, "Two novel schemes for opportunistic multi-access," in *IEEE Workshop on Multimedia Signal Processing (MMSP'02)*, (St. Thomas, US Virgin Islands), pp. 412–415, Dec. 2002.

[101] A. Sendonaris, E. Erkip, and B. Aazhang, "User cooperation diversity. Part I. System description," *IEEE Trans. on Communications*, vol. 51, pp. 1927–1938, Nov. 2003.

[102] A. Sendonaris, E. Erkip, and B. Aazhang, "User cooperation diversity. Part II. Implementation aspects and performance analysis," *IEEE Trans. on Communications*, vol. 51, pp. 1939–1948, Nov. 2003.

[103] I. Hammerström, M. Kuhn, and A. Wittneben, "Channel adaptive scheduling for cooperative relay networks," in *Proc. IEEE Vehicular Technology Conference (VTC'04-fall)*, (Los Angeles, CA, USA), pp. 2784–2788, Sept. 2004.

[104] A. J. Demers, S. Keshav, and S. Shenker, "Analysis and simulation of a fair queueing algorithm," in *Proc. ACM Symposium on Communications Architectures and Protocols (SIGCOMM'89)*, (Austin, TX, USA), pp. 1–12, Sept. 1989.

[105] A. K. Parekh and R. G. Gallager, "A generalized processor sharing approach to flow control in integrated services networks: The single-node case," *IEEE/ACM Trans. on Networking*, vol. 1, pp. 344–357, Mar. 1993.

[106] A. Stamoulis and G. B. Giannakis, "Packet fair queueing scheduling based on multirate multipath-transparent CDMA for wireless networks," in *Proc. IEEE Joint Conference of the Computer and*

Communications Societies (INFOCOM'00), vol. 3, (Tel-Aviv, Israel), pp. 1067–1076, Mar. 2000.

[107] Q. Liu, X. Wang, and G. B. Giannakis, "A cross-layer scheduling algorithm with QoS support in wireless networks," *IEEE Trans. on Vehicular Technology*, vol. 55, pp. 839–847, May 2006.

[108] D. Wu and R. Negi, "Effective capacity: A wireless link model for support of quality of service," *IEEE Trans. on Wireless Communications*, vol. 2, pp. 630–643, July 2003.

[109] S. Borst, "User-level performance of channel-aware scheduling algorithms in wireless data networks," in *Proc. IEEE Joint Conference of the Computer and Communications Societies (INFOCOM'03)*, vol. 1, (San Francisco, CA, USA), pp. 321–331, Mar.-Apr. 2003.

[110] M. Andrews, L. Qian, and A. Stolyar, "Optimal utility based multi-user throughput allocation subject to throughput constraints," in *Proc. of the 24th annual joint conference of the IEEE Conference Computer and Communications Societies (INFOCOM'05)*, vol. 4, (Miami, FL, USA), pp. 2415–2424, Mar. 2005.

[111] S. Borst and P. Whiting, "Dynamic channel-sensitive scheduling algorithms for wireless data throughput optimization," *IEEE Trans. on Vehicular Technology*, vol. 52, pp. 569–586, May 2003.

[112] Y. Cao and V. O. K. Li, "Scheduling algorithms in broadband wireless networks," *Proc. IEEE*, vol. 89, pp. 76–87, Jan. 2001.

[113] S. Shakkottai and A. L. Stolyar, "Scheduling for multiple flows sharing a time-varying channel: The exponential rule," *American Mathematical Society Translations*, vol. 207, 2002.

[114] P. Liu, R. Berry, and M. L. Honig, "Delay-sensitive packet scheduling in wireless networks," in *Proc. IEEE Wireless Communications and Networking Conference (WCNC'03)*, vol. 3, (New Orleans, LA, USA), pp. 1627–1632, Mar. 2003.

[115] D. Rajan, A. Sabharwal, and B. Aazhang, "Delay bounded packet scheduling of bursty traffic over wireless channels," *IEEE Trans. on Information Theory*, vol. 50, pp. 125–144, Jan. 2004.

[116] J. Huang, R. Berry, and M. Honig, "Wireless scheduling with hybrid ARQ," *IEEE Trans. on Wireless Communications*, vol. 4, pp. 2801–2810, Nov. 2005.

[117] M. Andrews, K. Kumaran, K. Ramanan, A. Stolyar, R. Vijayakumar, and P. Whiting, "CDMA data QoS scheduling on the forward link with variable channel conditions," *Bell Labs Technical Memorandum*, Apr. 2000.

[118] S. Shakkottai and A. L. Stolyar, "A study of scheduling algorithms for a mixture of real and non-real time data in HDR," tech. rep., Bell Laboratories, Oct. 2000.

[119] S. Shakkottai and R. Srikant, "Scheduling real-time traffic with deadlines over a wireless channel," *ACM/Baltzer Wireless Networks Journal*, vol. 8, pp. 13–26, Jan. 2002.

[120] D. Wu and R. Negi, "Utilizing multiuser diversity for efficient support of quality of service over a fading channel," *IEEE Trans. on Vehicular Technology*, vol. 54, pp. 1198–1206, May 2005.

[121] V. Hassel, M.-S. Alouini, G. E. Øien, and D. Gesbert, "Rate-optimal multiuser scheduling with reduced feedback load and analysis of delay effects," in *Proc. European Signal Processing Conference (EUSIPCO'05)*, (Antalya, Turkey), Sept. 2005.

[122] V. Hassel, M.-S. Alouini, D. Gesbert, and G. E. Øien, "Exploiting multiuser diversity using multiple feedback thresholds," in *Proc. IEEE Vehicular Technology Conference (VTC'05-spring)*, (Stockholm, Sweden), May 2005.

[123] H. Koubaa, V. Hassel, and G. E. Øien, "Multiuser diversity gain enhancement by guard time reduction," in *Proc. IEEE Workshop on Signal Processing Advances in Wireless Communications (SPAWC'05)*, (New York, NY, USA), pp. 845–849, June 2005.

[124] H. Koubaa, V. Hassel, and G. E. Øien, "Contention-less feedback for multiuser diversity scheduling," in *Proc. IEEE Vehicular Technology Conference (VTC'05-fall)*, (Dallas, TX, USA), pp. 1574–1578, Sept. 2005.

[125] V. Hassel, M. R. Hanssen, and G. E. Øien, "Spectral efficiency and fairness for opportunistic round robin scheduling," in *Proc. IEEE Int. Conference on Communications (ICC'06)*, (Istanbul, Turkey), June 2006.

[126] V. Hassel, G. E. Øien, and D. Gesbert, "Throughput guarantees for wireless networks with opportunistic scheduling." Presented at

the *IEEE Global Communications Conference (GLOBECOM'06)*, (San Fancisco, CA, USA), Nov.-Dec. 2006.

[127] V. Hassel, G. E. Øien, and D. Gesbert, "Throughput guarantees for opportunistic scheduling: A comparative study." Presented at the *Int. Telecommunications Symposium (ITS'06)*, (Fortaleza, Brazil), Sept. 2006.

[128] A. Seeger, M. Sikora, and W. Utschick, "Combined beamforming and scheduling for high speed downlink packet access," in *Proc. IEEE Global Communications Conference (GLOBECOM'03)*, (San Francisco, CA, USA), pp. 50–54, Dec. 2003.

[129] S. Serbetli and A. Yener, "Beamforming and scheduling strategies for time slotted multiuser MIMO systems," in *Proc. IEEE Asilomar Conference on Signals, Systems, and Computers*, vol. 1, (Pacific Grove, CA, USA), pp. 1227–1231, Nov. 2004.

[130] T. Ren and R. J. La, "Downlink beamforming algorithms with intercell interference in cellular networks," in *Proc. IEEE Joint Conference of the Computer and Communications Societies (INFOCOM'05)*, vol. 1, (Miami, FL, USA), pp. 47–57, Mar. 2005.

[131] A. M. Toukebri, S. Aïssa, and M. Maier, "Resource allocation and scheduling for multiuser MIMO systems: A beamforming-based strategy," in *Proc. IEEE Global Communications Conference (GLOBECOM'06)*, (San Francisco, CA, USA), Nov.-Dec. 2006.

[132] K. Kumaran and H. Viswanathan, "Joint power and bandwidth allocation in downlink transmission," *IEEE Trans. on Wireless Communications*, vol. 4, pp. 1008–1016, May 2005.

[133] A. Gjendemsjø, G. E. Øien, and H. Holm, "Optimal power control for discrete-rate link adaptation schemes with capacity-approaching coding," in *Proc. IEEE Global Communications Conference (GLOBECOM'05)*, (St. Louis, MO, USA), Nov.-Dec. 2005.

[134] R. L. Cruz, "Quality of service guarantees in virtual circuit switched networks," *IEEE Journal on Selected Areas in Communications*, vol. 13, no. 6, pp. 1048–1056, 1995.

[135] J.-Y. Le Boudec and P. Thiran, *Network calculus: a theory of deterministic queuing systems for the Internet*, vol. 2050.
Berlin, Germany: Springer-Verlag, 2001.

[136] Y. Jiang and P. J. Emstad, "Analysis of stochastic service guarantees in communication networks: A server model," in *Proc. IFIP Int. Workshop on Quality of Service (IWQoS'05)*, (Passau, Germany), pp. 233–245, June 2005.

[137] M. Fidler, "A network calculus approach to probabilistic quality of service analysis of fading channels," in *Proc. IEEE Global Communications Conference (GLOBECOM'06)*, (San Francisco, CA, USA), Nov.-Dec. 2006.

Part II

Included papers

Paper A

Rate-Optimal Multiuser Scheduling With Reduced Feedback Load and Analysis of Delay Effects

Vegard Hassel, Mohamed-Slim Alouini, Geir E. Øien, and David Gesbert

Published in
EURASIP Journal on Wireless Communications and Networking,
Special Issue on Radio Resource Management in 3G+ Systems, 2006

Abstract

In this paper, we propose a feedback algorithm for wireless networks that always collects feedback from the user with the best channel conditions and has a significant reduction in feedback load compared to full feedback. The algorithm is based on a carrier-to-noise threshold, and closed-form expressions for the feedback load as well as the threshold value that minimizes the feedback load have been found. We analyze two delay scenarios. The first scenario is where the scheduling decision is based on outdated channel estimates, and the second scenario is where both the scheduling decision and the adaptive modulation are based on outdated channel estimates.

1 Introduction

In a wireless network, the signals transmitted between the base station and the mobile users most often have different channel fluctuation characteristics. This diversity that exists between users is called *multiuser diversity* (MUD) and can be exploited to enhance the capacity of wireless networks [1]. One way of exploiting MUD is by *opportunistic scheduling* of users, giving priority to users having good channel conditions [2, 3]. Ignoring the feedback loss, the scheduling algorithm that maximizes the average system spectral efficiency among all time division multiplexing (TDM) based algorithms, is the one where the user with the highest carrier-to-noise ratio (CNR) is served in every time-slot [2]. Here, we refer to this algorithm as *Max CNR Scheduling* (MCS).

To be able to take advantage of the MUD, a base station needs feedback from the mobile users. Ideally, the base station only wants feedback from the user with the best channel conditions, but unfortunately each user does not know the CNR of the other users. Therefore, in current systems like Qualcomm's High Data Rate (HDR) system, the base station collects feedback from all the users [4].

One way to reduce the number of users giving feedback is by using a *CNR threshold*. For the *selective multiuser diversity* (SMUD) algorithm, it is shown that the feedback load is reduced significantly by using such a threshold [5]. For this algorithm only the users that have a CNR above a CNR threshold should send feedback to the scheduler. If the scheduler does not receive feedback, a random user is chosen. Because the best user is not chosen for every time-slot, the SMUD algorithm however introduces a reduction in system spectral efficiency. In addition it can be hard to set the threshold value for this algorithm. Applying a high threshold value will lead to low feedback load, but will additionally reduce the MUD gain and hence the system spectral efficiency. Using a low threshold value will have the opposite effect: the feedback load reduction is reduced, but the spectral efficiency will be higher.

The feedback algorithm proposed here is inspired by the SMUD algorithm, in the sense that this new algorithm also employs a feedback threshold. However, if none of the users succeed to exceed the CNR threshold, the scheduler requests full feedback, and selects the user with the highest CNR. Consequently, the MUD gain [1] is maximized, and still the feedback load is significantly reduced compared to the MCS algorithm. Another advantage with this novel algorithm is that for a specific set of system parameters it is possible to find a threshold value that minimizes the feedback load.

For the new feedback algorithm we choose to investigate two impor-

tant issues, namely, (i) how the algorithm can be optimized, and (ii) the consequences of delay in the system. The first issue is important because it gives theoretical limits for how well the algorithm will perform. The second issue is important because duration of the feedback collection process will often be significant and this will lead to a reduced performance of the opportunistic scheduling since the feedback information will be outdated. The consequences of delay are analyzed by looking separately at two different effects: (a) the system spectral efficiency degradation arising because the scheduler does not have access to instantaneous information about CNRs of the users, and, (b) the bit-error-rate (BER) degradation arising when both the scheduler and the mobile users do not have access to instantaneous channel measurements.

Contributions. We develop closed-form expressions for the feedback load of the new feedback algorithm. The expression for the threshold value which minimizes the feedback load is also derived. In addition we obtain new closed-form expressions for the system spectral efficiency degradation due to the *scheduling delay*. Finally, closed-form expressions for the effects of *outdated channel estimates* are obtained. Parts of the results have previously been presented in [6].

Organization. The rest of this paper is organized as follows. In Section 2, we present the system model. The feedback load is analyzed in Section 3, while Section 4 and Section 5 analyze the system spectral efficiency and BER, respectively. In Section 6 the effects of delay are discussed. Finally, Section 7 lists our conclusions.

2 System Model

We consider a single cell in a wireless network where the base station exchanges information with a constant number N of mobile users which have identically and independently distributed (i.i.d.) CNRs with an average of $\overline{\gamma}$. The system considered is TDM based, i.e., the information in transmitted in time-slots with a fixed length. We assume flat fading channels with a coherence time of one time-slot, which means that the channel quality remains roughly the same over the whole time-slot duration and that this channel quality is uncorrelated from one time-slot to the next. The system uses adaptive coding and modulation, i.e., the coding scheme, the modulation constellation and the transmission power used depends on the CNR of the selected user [7]. This has two advantages. On one hand, the spectral efficiency for each user is increased. On the other hand, because the rate of the users are varied according to their channel conditions, it makes

it possible to exploit MUD.

We will assume that the users always have data to send and that these user data are robust with respect to delay, i.e., no real-time traffic is transmitted. Consequently, the base station only has to take the channel quality of the users into account when it is performing scheduling.

The proposed feedback algorithm is applicable in at least two different types of cellular systems. The first system model is a time-division duplex (TDD) scenario where the same carrier frequency is used for both uplink and downlink. We can therefore assume a reciprocal channel for each user, i.e., the CNR is the same for the uplink and the downlink for a given point in time. The system uses the first half of the time-slot for downlink and the last half for uplink transmission. The users measure their channel for each downlink transmission and this measurement is fed back to the base station so that it can decide which user is going to be assigned the next time-slot. The second system model is a system where different carriers are used for uplink and downlink. For the base station to be able to schedule the user with the best downlink channel quality, the users must measure their channel for each downlink transmission and feed back their CNR measurement. For both system models the users are notified about the scheduling decision in a short broadcast message from the base station between each time-slot.

3 Analysis of the Feedback Load

The first step of the new feedback algorithm is to ask for feedback from the users that are above a CNR threshold value γ_{th}. The number of users n being above the threshold value γ_{th} is random and follow a *binomial distribution* given by:

$$\Pr(n) = \binom{N}{n}(1 - P_\gamma(\gamma_{th}))^n P_\gamma^{N-n}(\gamma_{th}), \quad n = 1, 2, \cdots, N, \tag{A.1}$$

where $P_\gamma(\gamma)$ is the cumulative distribution function (CDF) of the CNR for a single user. The second step of the feedback algorithm is to collect full feedback. Full feedback is only needed if all users' CNRs fail to exceed the threshold value. The probability of this event is given by inserting $\gamma = \gamma_{\text{th}}$ into

$$P_{\gamma^*}(\gamma) = P_\gamma^N(\gamma), \tag{A.2}$$

where γ^* denotes the CNR of the user with the best channel quality.

We now define the *normalized feedback load* (NFL) to be the ratio between the average number of users transmitting feedback, and the total number

of users. The NFL can be expressed as a the average of the ratio $\frac{n}{N}$, where n is the number of users giving feedback:

$$\begin{aligned}\bar{F} &= \frac{N}{N}P_\gamma^N(\gamma_{th}) + \sum_{n=1}^{N}\frac{n}{N}\binom{N}{n}(1-P_\gamma(\gamma_{th}))^n P_\gamma^{N-n}(\gamma_{th}) \\ &= P_\gamma^N(\gamma_{th}) + (1-P_\gamma(\gamma_{th}))\sum_{n=1}^{N}\binom{N-1}{n-1}(1-P_\gamma(\gamma_{th}))^{n-1}P_\gamma^{N-n}(\gamma_{th}) \\ &= P_\gamma^N(\gamma_{th}) + (1-P_\gamma(\gamma_{th}))\sum_{k=0}^{N-1}\binom{N-1}{k}(1-P_\gamma(\gamma_{th}))^{k}P_\gamma^{N-1-k}(\gamma_{th}) \\ &= 1 - P_\gamma(\gamma_{th}) + P_\gamma^N(\gamma_{th}), \quad N = 2,3,4,\cdots, \end{aligned} \tag{A.3}$$

where the last equality is obtained by using binomial expansion [8, Eq. (1.111)]. For $N = 1$ full feedback is needed, and $\bar{F} = 1$. In that case the feedback is not useful for multiuser scheduling, but for being able to adapt the base station's modulation according to the channel quality in the reciprocal TDD system model described in the previous section.

A plot of the feedback load as a function of γ_{th} is shown in Fig. A.1 for $\overline{\gamma}$= 15 dB. It can be observed that the new algorithm reduces the feedback significantly compared to a system with full feedback. It can also be observed that one threshold value will minimize the feedback load in the system for a given number of users.

The expression for the threshold value that minimizes the average feedback load can be found by differentiating (A.3) with respect to γ_{th} and setting the result equal to zero:

$$\gamma_{th}^* = P_\gamma^{-1}\left(\left(\frac{1}{N}\right)^{\frac{1}{N-1}}\right), \quad N = 2,3,4,\cdots, \tag{A.4}$$

where $P_\gamma^{-1}(\cdot)$ is the inverse CDF of the CNR. In particular, for a Rayleigh fading channel, with CDF $P_\gamma(\gamma) = 1 - e^{-\gamma/\overline{\gamma}}$, the optimum threshold can be found in a simple closed-form as:

$$\gamma_{th}^* = -\overline{\gamma}\ln\left(1-\left(\frac{1}{N}\right)^{\frac{1}{N-1}}\right), \quad N = 2,3,4,\cdots. \tag{A.5}$$

4 System Spectral Efficiencies for Different Power and Rate Adaptation Techniques

To be able to analyze the system spectral efficiency we choose to investigate the *maximum average system spectral efficiency* (MASSE) theoretically attain-

able. The MASSE [Bit/Sec/Hz] is defined as the maximum average sum of spectral efficiency for a carrier with bandwidth W [Hz].

4.1 Constant Power and Optimal Rate Adaptation

Since the best user is always selected, the MASSE of the new algorithm is the same as for the MCS algorithm. To find the MASSE for such a scenario, the probability density function (PDF) of the highest CNR among all the users has to be found. This PDF can be obtained by differentiating (A.2) with respect to γ. Inserting the CDF and PDF for Rayleigh fading channels ($p_\gamma(\gamma) = (1/\overline{\gamma})e^{-\gamma/\overline{\gamma}}$), and using binomial expansion [8, Eq. (1.111)], we obtain:

$$p_{\gamma^*}(\gamma) = \frac{N}{\overline{\gamma}} \sum_{n=0}^{N-1} \binom{N-1}{n} (-1)^n e^{-(1+n)\gamma/\overline{\gamma}}. \tag{A.6}$$

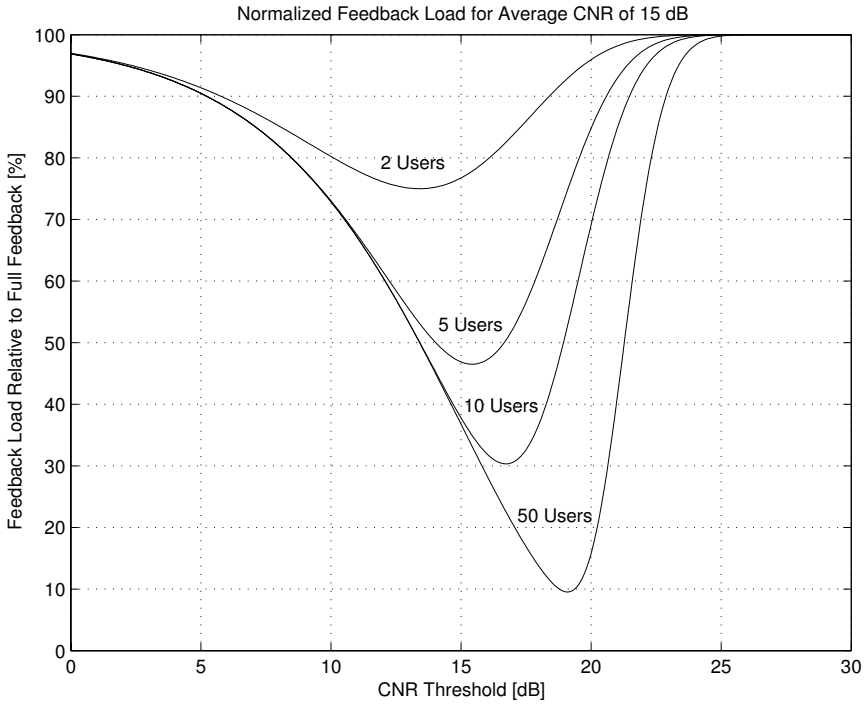

FIGURE A.1: Normalized feedback load as a function of γ_{th} with $\overline{\gamma} = 15$ dB.

Inserting (A.6) into the expression for the spectral efficiency for optimal rate adaptation found in [9], the following expression for the MASSE can be obtained [10, Eq. (44)]:

$$\begin{aligned}\frac{< C >_{ora}}{W} &= \int_0^\infty \log_2(1+\gamma) p_{\gamma^*}(\gamma) d\gamma \\ &= \frac{N}{\ln 2} \sum_{n=0}^{N-1} \binom{N-1}{n} (-1)^n \frac{e^{\frac{1+n}{\overline{\gamma}}}}{1+n} E_1\left(\frac{1+n}{\overline{\gamma}}\right), \end{aligned} \quad \text{(A.7)}$$

where *ora* denotes *optimal rate adaptation* and $E_1(\cdot)$ is the *first order exponential integral function* [8].

4.2 Optimal Power and Rate Adaptation

It has been shown that the MASSE for optimal power and rate adaptation can be obtained as [10, Eq. (27)]:

$$\begin{aligned}\frac{< C >_{opra}}{W} &= \int_0^\infty \log_2\left(\frac{\gamma}{\gamma_0}\right) p_{\gamma^*}(\gamma) d\gamma \\ &= \frac{N}{\ln 2} \sum_{n=0}^{N-1} \binom{N-1}{n} \frac{(-1)^n}{1+n} E_1\left(\frac{(1+n)\gamma_0}{\overline{\gamma}}\right), \end{aligned} \quad \text{(A.8)}$$

where *opra* denotes *optimal power and rate adaptation* and γ_0 is the optimal cut-off CNR level below which data transmission is suspended. This cut-off value must satisfy [9]:

$$\int_{\gamma_0}^\infty \left(\frac{1}{\gamma_0} - \frac{1}{\gamma}\right) p_{\gamma^*}(\gamma)\, d\gamma = 1. \quad \text{(A.9)}$$

Inserting (A.6) into (A.9), it can subsequently be shown that the following cut-off value can be obtained for Rayleigh fading channels [10, Eq. (24)]:

$$\sum_{n=0}^{N-1} \binom{N-1}{n} (-1)^n \left(\frac{e^{-(1+n)\gamma_0/\overline{\gamma}}}{(1+n)\gamma_0/\overline{\gamma}} - E_1\left(\frac{(1+n)\gamma_0}{\overline{\gamma}}\right)\right) = \frac{\overline{\gamma}}{N}. \quad \text{(A.10)}$$

5 M-QAM Bit-Error-Rates

The BER of coherent M-ary quadrature amplitude modulation (M-QAM) with two-dimensional Gray coding over an additive white Gaussian noise (AWGN) channel can be approximated by [11]:

$$\mathrm{BER}(M,\gamma) \approx 0.2 \exp\left(-\frac{3\gamma}{2(M-1)}\right). \quad \text{(A.11)}$$

The constant-power *adaptive continuous rate* (ACR) M-QAM scheme can always adapt the rate to the instantaneous CNR. From [12] we know that the constellation size for continuous-rate M-QAM can be approximated by $M \approx \left(1 + \frac{3\gamma}{2K_0}\right)$, where $K_0 = -\ln(5\,\mathrm{BER}_0)$ and BER_0 is the target BER. Consequently, it can be easily shown that the theoretical constant-power ACR M-QAM scheme always operates at the target BER.

For physical systems only integer constellation sizes are practical, so now we restrict the constellation size M_k to 2^k where k is a positive integer. This adaptation policy is called *adaptive discrete rate* (ADR) M-QAM, and the CNR range is divided into $K+1$ *fading regions* with constellation size M_k assigned to the kth fading region. Because of the discrete assignment of constellation sizes in ADR M-QAM, this scheme has to operate at a BER lower than the target. The average BER for ADR M-QAM using constant power can be calculated as [12]:

$$< \mathrm{BER} >_{adr} = \frac{\sum_{k=1}^{K} k\,\overline{\mathrm{BER}_k}}{\sum_{k=1}^{K} k\,p_k}, \tag{A.12}$$

where

$$\overline{\mathrm{BER}_k} = \int_{\gamma_k}^{\gamma_{k+1}} \mathrm{BER}(M_k, \gamma)\, p_{\gamma^*}(\gamma)\, d\gamma, \tag{A.13}$$

and

$$p_k = \left(1 - e^{-\gamma_{k+1}/\overline{\gamma}}\right)^N - \left(1 - e^{-\gamma_k/\overline{\gamma}}\right)^N \tag{A.14}$$

is the probability that the scheduled user is in the fading region k for CNRs between γ_k and γ_{k+1}.

Inserting (A.11) and (A.6) into (A.13) we obtain the following expression for the average BER within a fading region:

$$\overline{\mathrm{BER}_k} = \frac{0.2N}{\overline{\gamma}} \sum_{n=0}^{N-1} \binom{N-1}{n} (-1)^n \frac{e^{-\gamma_k a_{k,n}} - e^{-\gamma_{k+1} a_{k,n}}}{a_{k,n}}, \tag{A.15}$$

where $a_{k,n}$ is given by

$$a_{k,n} = \frac{1+n}{\overline{\gamma}} + \frac{3}{2(M_k - 1)}. \tag{A.16}$$

When power adaptation is applied, the BER approximation in (A.11) can be written as [11]:

$$\mathrm{BER}_{pa}(M, \gamma) \approx 0.2 \exp\left(-\frac{3\gamma}{2(M-1)} \frac{S_k(\gamma)}{S_{av}}\right), \tag{A.17}$$

where $S_k(\gamma)$ is the power used in fading region k and S_{av} is the average transmit power. Inserting the continuous power adaptation policy given by [11, Eq. (29)] into (A.17) shows that the ADR M-QAM scheme using optimal power adaptation always operates at the target BER. Correspondingly, it can be shown that the continuous-power, continuous-rate M-QAM scheme always operates at the target BER.

6 Consequences of Delay

In the previous sections, it has been assumed that there is no delay from the instant where the channel estimates are obtained and fed back to the scheduler, to the time when the optimal user is transmitting. For real-life systems, we have to take delay into consideration. We analyze, in what follows, two delay scenarios. In the first scenario, a *scheduling delay* arises because the scheduler receives channel estimates, takes a scheduling decision, and notifies the selected user. This user then transmits, but at a possibly different rate. The second scenario deals with *outdated channel estimates*, which leads to both a scheduling delay as well as suboptimal modulation constellations with increased BERs.

Outdated channel estimates have been treated to some extent in previous publications [12, 13]. However, the concept of scheduling delay has in most cases been analyzed for wire-line networks only [14, 15]. Although some previous work have been done on scheduling delay in wireless networks [16], scheduling delay has to the best of our knowledge not been looked into for cellular networks.

6.1 Impact of Scheduling Delay

In this subsection we will assume that the scheduling decision is based on a perfect estimate of the channel at time t, whereas the data are sent over the channel at time $t+\tau$. We will assume that the link adaptation done at time $t+\tau$ is based on yet another channel estimate taken at $t+\tau$. To investigate the influence of this type of scheduling delay, we need to develop a PDF for the CNR at time $t+\tau$, conditioned on channel knowledge at time t. Let α and α_τ be the channel gains at time t and $t+\tau$, respectively. Assuming that the average power gain remains constant over the time delay τ for a slowly-varying Rayleigh channel, (i.e. $\Omega = E[\alpha^2] = E[\alpha_\tau^2]$), and using the same approach as in [12] it can be shown that the conditional PDF $p_{\alpha_\tau|\alpha}(\alpha_\tau|\alpha)$ is

given by:

$$p_{\alpha_\tau|\alpha}(\alpha_\tau|\alpha) = \frac{2\alpha_\tau}{(1-\rho)\,\Omega}\, I_0\left(\frac{2\sqrt{\rho}\,\alpha\alpha_\tau}{(1-\rho)\,\Omega}\right) e^{-\frac{(\alpha_\tau^2+\rho\alpha^2)}{(1-\rho)\,\Omega}}. \tag{A.18}$$

where ρ is the correlation factor between α and α_τ and $I_0(\cdot)$ is the *zeroth-order modified Bessel function of the first kind* [8]. Assuming Jakes Doppler spectrum, the correlation coefficient can be expressed as $\rho = J_0^2(2\pi f_D \tau)$, where $J_0(\cdot)$ is the *zeroth-order Bessel function of the first kind* and f_D [Hz] is the maximum Doppler frequency shift [12]. Recognizing that (A.18) is similar to [17, Eq. (A-4)] gives the following PDF at time $t+\tau$ for the new feedback algorithm, expressed in terms of γ_τ and $\overline{\gamma}$ [17, Eq. (5)]:

$$p_{\gamma_\tau^*}(\gamma_\tau) = \sum_{n=0}^{N-1} \binom{N}{n+1} (-1)^n \frac{\exp\left(-\frac{\gamma_\tau}{\overline{\gamma}(1-\rho\frac{n}{n+1})}\right)}{\overline{\gamma}(1-\rho\frac{n}{n+1})}. \tag{A.19}$$

Note that for $\tau = 0$ ($\rho = 1$) this expression reduces to (A.6), as expected. When τ approaches infinity ($\rho = 0$) (A.19) reduces to the Rayleigh PDF for one user. This is logical since for large τs, the scheduler will have completely outdated and as such useless feedback information, and will end up selecting users independent of their CNRs.

Inserting (A.19) into the capacity expression for optimal rate adaptation in [9, Eq. (2)], then using binomial expansion, integration by parts, L'Hôpital's rule, and [8, Eq. (3.352.2)], it can be shown that we get the following expression for the MASSE:

$$\begin{aligned}\frac{< C >_{ora}}{W} &= \int_0^\infty \log_2(1+\gamma_\tau)\, p_{\gamma_\tau^*}(\gamma_\tau) d\gamma_\tau \\ &= \frac{1}{\ln 2}\sum_{n=0}^{N-1}\binom{N}{n+1}(-1)^n e^{\frac{1}{\overline{\gamma}(1-\rho\frac{n}{n+1})}} E_1\left(\frac{1}{\overline{\gamma}\left(1-\rho\frac{n}{n+1}\right)}\right).\end{aligned} \tag{A.20}$$

Using a similar derivation as for the expression above it can furthermore be shown that we get the following expression for the MASSE using both optimal power and rate adaptation:

$$\begin{aligned}\frac{< C >_{opra}}{W} &= \int_0^\infty \log_2\left(\frac{\gamma_\tau}{\gamma_0}\right) p_{\gamma_\tau^*}(\gamma_\tau) d\gamma_\tau \\ &= \frac{1}{\ln 2}\sum_{n=0}^{N-1}\binom{N}{n+1}(-1)^n E_1\left(\frac{\gamma_0}{\overline{\gamma}\left(1-\rho\frac{n}{n+1}\right)}\right),\end{aligned} \tag{A.21}$$

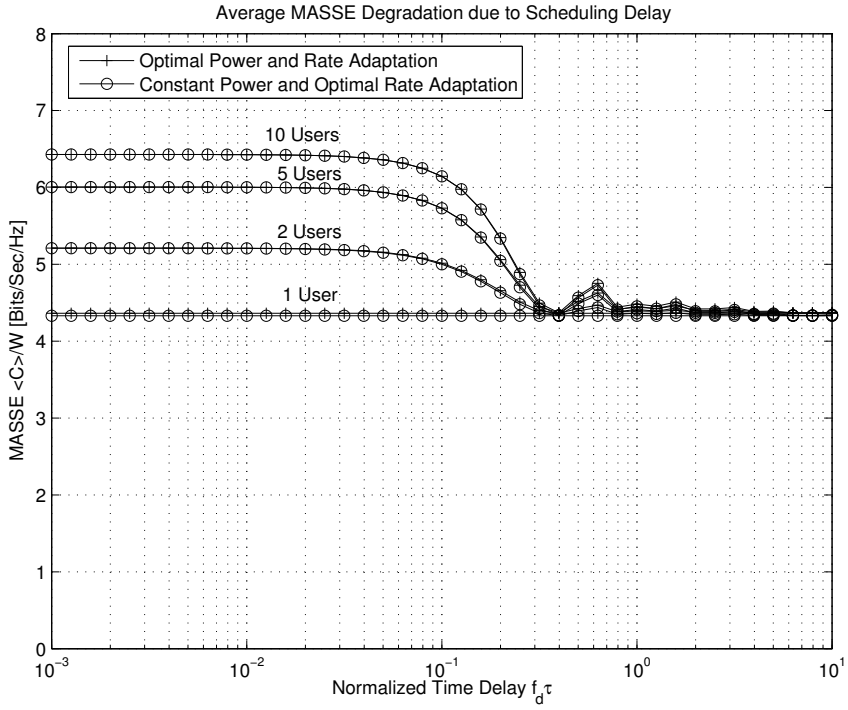

FIGURE A.2: Average degradation in MASSE due to scheduling delay for (i) optimal power and rate adaptation and (ii) optimal rate adaptation.

with the following power constraint:

$$\sum_{n=0}^{N-1} \binom{N}{n+1} (-1)^n \left(\frac{e^{-\frac{1}{\overline{\gamma}(1-\rho\frac{n}{n+1})}}}{\gamma_0} - \frac{E_1\left(\frac{1}{\overline{\gamma}(1-\rho\frac{n}{n+1})}\right)}{\overline{\gamma}(1-\rho\frac{n}{n+1})} \right) = 1. \tag{A.22}$$

Again, for zero time delay ($\rho = 1$), (A.20) reduces to (A.7), (A.21) reduces to (A.8), and (A.22) reduces to (A.10), as expected.

Fig. A.2 shows how scheduling delay affects the MASSE for 1, 2, 5, and 10 users. We see that both optimal power and rate adaptation and optimal rate adaptation are equally robust with regard to the scheduling delay. Independent of the number of users, we see that the system will be able to operate satisfactory if the normalized delay is below the critical value of $2 \cdot 10^{-2}$. For normalized time delays above this value, we see that the MASSE converges towards the MASSE for one user, as one may expect.

6.2 Impact of Outdated Channel Estimates

We will now assume that the transmitter does not have a perfect outdated channel estimate available at time $t+\tau$, but only at time t. Consequently, both the selection of a user and the decision of the constellation size have to be done at time t. This means that the channel estimates are outdated by the same amount of time as the scheduling delay. The constellation size is thus not dependent on γ_τ, and the time delay in this case does not affect the MASSE. However, now the BER will suffer from degradation because of the delay. It is shown in [12] that the average BER, conditioned on γ, is

$$\mathrm{BER}(\gamma) = \frac{0.2\gamma}{\gamma + \overline{\gamma}(1-\rho)K_0} \cdot e^{-\frac{\rho K_0 \gamma}{\gamma + \overline{\gamma}(1-\rho)K_0}}. \tag{A.23}$$

The average BER can be found by using the following equation:

$$< \mathrm{BER} >_{acr} = \int_0^\infty \mathrm{BER}(\gamma) p_{\gamma^*}(\gamma)\, d\gamma. \tag{A.24}$$

For discrete rate adaptation with constant power, the BER can be expressed by (A.12), replacing $\overline{\mathrm{BER}}_k$ with $\overline{\mathrm{BER}}'_k$, where:

$$\overline{\mathrm{BER}}'_k = \int_{\gamma_k}^{\gamma_{k+1}} \int_0^\infty \mathrm{BER}(M_k, \gamma_\tau) p_{\gamma_\tau|\gamma}(\gamma_\tau|\gamma) d\gamma_\tau p_{\gamma^*}(\gamma) d\gamma. \tag{A.25}$$

Inserting (A.6), (A.11) and (A.18) expressed in terms of γ_τ and γ into (A.25), we obtain the following expression for the average BER within a fading region:

$$\overline{\mathrm{BER}}'_k = \frac{0.2N}{\overline{\gamma}} \sum_{n=0}^{N-1} \binom{N-1}{n} (-1)^n \frac{e^{-\gamma_k c_{k,n}} - e^{-\gamma_{k+1} c_{k,n}}}{d_{k,n}}, \tag{A.26}$$

where $c_{k,n}$ is given by

$$c_{k,n} = \frac{1+n}{\overline{\gamma}} + \frac{3\rho}{3\overline{\gamma}(1-\rho) + 2(M_k - 1)}, \tag{A.27}$$

and $d_{k,n}$ by

$$d_{k,n} = \frac{1+n}{\overline{\gamma}} + \frac{3(1+n-\rho n)}{2(M_k-1)}. \tag{A.28}$$

Note that for zero delay ($\rho = 1$) $c_{k,n} = d_{k,n} = a_{k,n}$, and (A.26) reduces to (A.15), as expected.

Because we are interested in the average BER only for the CNRs for which we have transmission, the average BER for continuous-power, continuous-rate M-QAM is

$$< \text{BER} >_{acr,pa} = \frac{\int_{\gamma_K}^{\infty} \text{BER}(\gamma) p_{\gamma^*}(\gamma)\, d\gamma}{\int_{\gamma_K}^{\infty} p_{\gamma^*}(\gamma)\, d\gamma}. \tag{A.29}$$

Correspondingly, the average BER for the continuous-power, discrete-rate M-QAM case is given by:

$$< \text{BER} >_{adr,pa} = \frac{\int_{\gamma_0^* M_1}^{\infty} \text{BER}(\gamma) p_{\gamma^*}(\gamma)\, d\gamma}{\int_{\gamma_0^* M_1}^{\infty} p_{\gamma^*}(\gamma)\, d\gamma}. \tag{A.30}$$

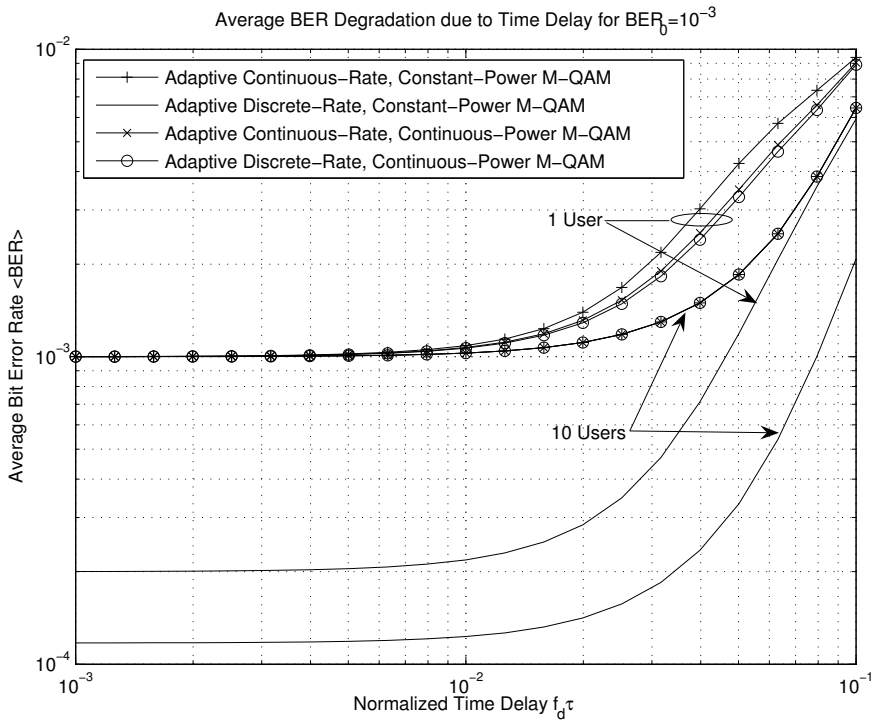

FIGURE A.3: Average BER degradation due to time delay for M-QAM rate adaptation with $\overline{\gamma} = 15$ dB, 5 fading regions and $\text{BER}_0 = 10^{-3}$.

Fig. A.3 shows how outdated channel estimates affect the average BER for 1 and 10 users. We see that the average system BER is satisfactory as

long as the normalized time delay again is below the critical value 10^{-2} for the adaptation schemes using continuous power and/or continuous rate. The constant-power, discrete-rate adaptation policy is more robust with regard to time delay.

7 Conclusion

We have analyzed a scheduling algorithm that has optimal spectral efficiency, and reduced feedback compared with full feedback load. We obtain a closed-form expression for the CNR threshold that minimizes the feedback load for this algorithm. Both the impact of scheduling delay and outdated channel estimates are analytically and numerically described. For both delay scenarios plots show that the system will be able to operate satisfactorily with regard to BER when the normalized time delays are below certain critical values.

References

[1] P. Viswanath, D. N. C. Tse, and R. Laroia, "Opportunistic beamforming using dumb antennas," *IEEE Trans. Inform. Theory*, vol. 48, pp. 1277–1294, June 2002.

[2] R. Knopp and P. A. Humblet, "Information capacity and power control in single cell multiuser communications," in *Proc. IEEE Int. Conf. on Communications (ICC'95)*, (Seattle, WA, USA), pp. 331–335, June 1995.

[3] M. Andrews, K. Kumaran, K. Ramanan, A. Stolyar, P. Whiting, and R. Vijayakumar, "Providing quality of service over a shared wireless link," *IEEE Communications Mag.*, pp. 150–154, Feb. 2001.

[4] P. Bender, P. Black, M. Grob, R. Padovani, N. Sindhushyana, and S. Viterbi, "CDMA/HDR: A bandwidth-efficient high-speed wireless data service for nomadic users," *IEEE Communications Mag.*, vol. 38, pp. 70–77, July 2000.

[5] D. Gesbert and M.-S. Alouini, "How much feedback is multi-user diversity really worth?," in *Proc. IEEE Int. Conf. on Communications (ICC'04)*, (Paris, France), pp. 234–238, June 2004.

[6] V. Hassel, M.-S. Alouini, G. E. Øien, and D. Gesbert, "Rate-optimal multiuser scheduling with reduced feedback load and analysis of delay effects." Presented at *European Signal Processing Conference (EUSIPCO'05)*, Antalya, Turkey, Sept. 2005.

[7] K. J. Hole and G. E. Øien, "Spectral efficiency of adaptive coded modulation in urban microcellular networks," *IEEE Trans. on Veh. Technol.*, vol. 50, pp. 205–222, Jan. 2001.

[8] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products*. San Diego, CA, USA: Academic Press, 6th ed., 2000.

[9] A. J. Goldsmith and P. P. Varaiya, "Capacity of fading channels with channel side information," *IEEE Trans. Inform. Theory*, vol. IT-43, pp. 1896–1992, Nov. 1997.

[10] M.-S. Alouini and A. J. Goldsmith, "Capacity of Rayleigh fading channels under different adaptive transmission and diversity-combining techniques," *IEEE Trans. on Veh. Technol.*, vol. 48, pp. 1165–1181, July 1999.

[11] A. J. Goldsmith and S.-G. Chua, "Variable-rate variable-power M-QAM for fading channels," *IEEE Trans. on Communications*, vol. COM-45, pp. 1218–1230, Oct. 1997.

[12] M.-S. Alouini and A. J. Goldsmith, "Adaptive modulation over Nakagami fading channels," *Kluwer Journal on Wireless Communications*, vol. 13, pp. 119–143, May 2000.

[13] D. L. Goeckel, "Adaptive coding for time-varying channels using outdated fading estimates algorithms," *IEEE Trans. on Communications*, pp. 844–855, June 1999.

[14] S. Bolis, E. G. Economou, and P. G. Philokyprou, "Scheduling delay protocols integrating voice and data on a bus lan," *IEE Proceedings I Communications Speech and Vision*, vol. 139, pp. 402–412, Aug. 1992.

[15] H.-H. Chen and W.-T. Tea, "Hierarchy schedule sensing protocol for CDMA wireless networks – performance study under multipath, multiuser interference, and collision-capture effect," *IEEE Trans. on Mobile Computing*, vol. 4, pp. 178–188, March/April 2005.

[16] K.-W. Hung and T.-S. Yum, "Fair and efficient transmission scheduling in multihop packet radio networks," in *Proc. IEEE Global Communications Conf. (GLOBECOM'92)*, vol. 1, (Orlando, FL), pp. 6–10, Dec. 1992.

[17] J. H. Barnard and C. K. Pauw, "Probability of error for selection diversity as a function of dwell time," *IEEE Trans. on Communications*, vol. 37, pp. 800–803, Aug. 1989.

Paper B

A Threshold-Based Channel State Feedback Algorithm for Modern Cellular Systems

Vegard Hassel, David Gesbert, Mohamed-Slim Alouini, and Geir E. Øien

Accepted for publication in
IEEE Transactions on Wireless Communications

Abstract

In this paper we propose a channel state feedback algorithm that uses multiple feedback thresholds to reduce the number of users transmitting feedback to a minimum. The users are polled with lower and lower threshold values and only the users that are above a threshold value transmit feedback to the base station. We show how this feedback algorithm can be used for any scheduling algorithm and show how closed-form expressions for the optimal threshold values can be obtained for two well-known scheduling algorithms. Finally, we propose a two-step optimization procedure for optimizing the feedback algorithm for real-life cellular standards.

1 Introduction

In modern wireless networks, *adaptive coding and modulation* are implemented so that the mobile users and base stations can adapt their transmission rate to the quality of the wireless channel [1]. This adaptation does not only increase the spectral efficiency of the wireless links between the base station and the mobile users, but can also be exploited further by the base station through *opportunistic scheduling* [2]. Opportunistic scheduling increases the system spectral efficiency by giving priority to mobile users when they have good channel quality. Opportunistic scheduling algorithms in cellular networks are often executed by the base station. This means that the base station needs to know the instantaneous channel quality of the users in the system and schedule the users based on this knowledge. In modern wireless standards like Mobile WiMAX, HSPA, and 1xEVDO, opportunistic scheduling algorithms can be implemented to schedule users for every time-slot in the down-link [3–5]. Therefore, when the channels are rapidly varying, most of the opportunistic scheduling algorithms are based on having available channel quality estimates for *all* the mobile users in every time-slot. If all the mobile users are going to feed back their channel quality estimates to the base station for each time-slot, a significant share of the battery energy will be used on transmission of overhead information instead of useful data traffic. In addition, for many wireless systems, collecting carrier-to-noise ratio (CNR) estimates from all the users will lead to a significant delay before the transmission of useful data can start.

Three main directions have previously been pursued to reduce the degradation due to feedback, namely, (i) feedback quantization, (ii) feedback compression, and (iii) feedback load reduction. Publications investigating the first approach have shown that heavy quantization of the channel state information (CSI) being fed back, will not lead to a significant reduction of the system gain [6, 7]. Correspondingly, the quantization of the beamforming vector being fed back, has been investigated for multi-antenna systems [8]. The second approach exploits the channel correlation in time and frequency to design compression algorithms that reduce the feedback overhead significantly [9, 10]. Most algorithms trying to reduce the feedback load, i.e., the number of users feeding back channel state information, are based on CNR thresholds [11, 12]. One threshold-based algorithm that uses a single CNR threshold value was proposed by Gesbert and Alouini [12]. The mobile users that have a CNR above this threshold value transmit feedback to the scheduler. The algorithm in [12] does not always obtain feedback from the user with the highest CNR since it will

always be a possibility that all users are below the threshold value and a random user has to be chosen.

The goal of this paper is to conduct a theoretical investigation of a novel feedback algorithm that is a generalization of the algorithm in [12]. This generalization is based on two main ideas, namely, (i) adapting the feedback threshold value to the *scheduling metric* of the scheduling algorithm, i.e. the metric that is used to decide which user is going to be scheduled in a time-slot, and (ii) using multiple feedback thresholds to collect feedback from the *preferred user*, i.e. the user that the scheduling algorithm prefers to schedule. For any scheduling algorithm, our proposed algorithm leads to a significant reduction of the number of users transmitting feedback. This will reduce the power consumption of the mobile users, and also reduce the time to collect feedback for many cellular systems.

Previous publications related to feedback load reduction are all based on the assumption that the base station always tries to collect feedback from the user with the highest CNR. This is because it is assumed that the scheduling algorithm used by the system is Max CNR Scheduling (MCS), where the user with the highest CNR is scheduled in every time-slot. Since the MCS algorithm can be unfair in many cases, other scheduling algorithms are often preferred, and we thus propose to adapt the feedback threshold values to account for any scheduling metric.

By employing multiple feedback thresholds, the base station can conduct the feedback collection process by polling the users sequentially from the highest threshold value down to the lowest threshold value until feedback from one or more users is received. Our numerical results show that by employing just a few number of thresholds, our algorithm will lead to a high probability for receiving feedback from only one user.

Contributions. We propose the novel channel state feedback algorithm as described above. In Section 3, we assume that the time to collect feedback is negligible compared to the time used to transmit user data. We find an analytical expression for the feedback load when the number of threshold values L is predetermined and when a general scheduling metric is assumed. Based on this general expression, we obtain specific expressions for the feedback load for the MCS and the NCS algorithms, and use these expressions to find the feedback threshold values minimizing the feedback load for each of these two algorithms. In Section 4, we argue that the time used to collect feedback is in fact non-negligible in many real-life systems. For such systems, we propose a two-step method for optimizing both the threshold values and the number of thresholds L.

2 System Model

We consider a time-division multiplexed (TDM) based wireless system with a single base station scheduling N users. The transmission rate and the scheduling decision is based on CNR estimates being fed back from one or more users for each time-slot. If we assume that we have reciprocity between up-link and down-link, which can be assumed in e.g. Mobile WiMAX, these CNR estimates can be used to schedule users in both the down-link and up-link. If this is not the case, the system model is only valid for down-link scheduling. We assume that the channels of all users are independent, flat-fading channels with average CNR $\overline{\gamma}_i$, where the index i denotes the ith user. In order to have a roughly constant CNR level, $\gamma_i(t)$, within each time-slot, it is assumed that the duration of a time-slot is shorter than the coherence time of the channels.

The feedback threshold values vary with the scheduling metric $x_i(t)$ used by the scheduling algorithm. Examples of scheduling metrics of different scheduling algorithms will be given in the beginning of Section 3. The base station searches for users in the whole range of their scheduling metric and we denote the feedback thresholds for user i by $x_{\text{th},i,L} > x_{\text{th},i,L-1} > \cdots > x_{\text{th},i,0}$. Assuming that the x_i-ranges are from zero to infinity, we set $x_{\text{th},i,L} = \infty$ and $x_{\text{th},i,0} = 0$, and let the base station start polling each of the users with $x_{\text{th},i,L-1}$ for user i. Since $x_{\text{th},i,L}$ is never used to search for the users, we say that we have L threshold values. Note that since the feedback thresholds are covering the whole x_i-range of the users, it is ensured that feedback will be received from at least one user. In practice, it is often a need for calculating the threshold values on-line. Since a threshold value can be expressed as a function of the CNR of a user, the optimization of the threshold values does not have to be performed by the users since the base station can always poll each of the users with CNR values that correspond to the threshold values $x_{\text{th},i,l}$, $l = 0, ..., L$.

The goal of this paper is to conduct a general theoretical analysis of the proposed feedback algorithm. However, to conduct an analysis of the true performance of the algorithm in different real-life networks, the system model needs further specification. For wireless systems based on WLAN standards, HSPA, Mobile WiMAX or 1xEVDO, a performance analysis of our algorithm has to be conducted by considering system specific parameters and protocols (See e.g. [13]). For HSPA, which uses code-division multiplexing (CDM) in both the up-link and the down-link, more users can transmit or receive data simultaneously. This means that CNR estimates can be fed back by several users simultaneously. Also for such systems our proposed feedback algorithm can be implemented to reduce the number of

users transmitting feedback. The main gain for the system will be a lower power consumption for the users; in addition we also get less interference on the feedback channel, and hence also a lower bit-error-rate for the CNR estimates being fed back. For CDM-based systems, opportunistic scheduling algorithms often have to pick a subset of the users to transmit or receive within a time-slot. Our proposed feedback algorithm can also be used to obtain feedback from such a subset of preferred users.

3 Optimizing the Algorithm for a Fixed Number of Feedback Thresholds

In this section we assume that the time to collect feedback is negligible compared to the time used to transmit user data. This means that we focus on how the number of users transmitting feedback can be minimized for a given value of L. By denoting the scheduling metric as $x_i(t)$ for user i in time-slot t, we have $x_i(t) = \gamma_i(t)$ for the MCS algorithm, where $\gamma_i(t)$ denotes the instantaneous CNR of user i in time-slot t. For the Normalized CNR Scheduling algorithm (NCS) the corresponding scheduling metric is $x_i(t) = \gamma_i(t)/\overline{\gamma}_i$, where $\overline{\gamma}_i$ is the average CNR of user i [14]. For this scheduling algorithm the threshold values have to be optimized for searching for the preferred users in the $\gamma_i(t)/\overline{\gamma}_i$-range of the users. Likewise, the feedback algorithm can be designed for the Proportional Fair Scheduling (PFS) algorithm by optimizing the feedback thresholds to search for the preferred user in the $r_i(t)/T_i$-range of the users, where $r_i(t)$ denotes the instantaneous rate of user i and T_i denotes a weighted sum of the rate allocated to this user [15].

3.1 Feedback Thresholds for a General Scheduling Metric

To evaluate the performance of our feedback algorithm for a general scheduling metric $x_i(t)$, we have to find an expression for the *normalized feedback load* (NFL), which expresses the average share of users that give feedback for each time-slot. It can be shown that the NFL can be obtained as:

$$\bar{\mathrm{F}}_{\mathrm{gen}} = \frac{1}{N}\sum_{l=0}^{L-1}\sum_{\Psi}|\Psi|\prod_{i\in\Psi}(P_{x_i}(x_{\mathrm{th},i,l+1}) - P_{x_i}(x_{\mathrm{th},i,l}))\prod_{j\notin\Psi}P_{x_j}(x_{\mathrm{th},j,l}), \qquad \text{(B.1)}$$

where $\Psi \neq \emptyset$ denotes any subset of users, including the set of all users, while $P_{x_i}(\cdot)$ denotes the cumulative distribution function (CDF) of the scheduling metric of user i. For many scheduling algorithms it can be hard

to find closed-form expressions for the distributions $P_{x_i}(\cdot)$ for $i = 1, ..., N$. For such scheduling algorithms these distributions have to be estimated on-line. Such estimations can be performed at the base station, based on the CNR estimates being fed back, by using for example *order statistic filter banks* [16]. To obtain the optimal feedback thresholds, the feedback thresholds that minimize (B.1) have to be obtained. For the PFS algorithm and many other algorithms, the distributions $P_{x_i}(\cdot)$ are not known and a numerical optimization procedure has to be employed. However, for the MCS and the NCS algorithms, the distributions $P_{x_i}(\cdot)$ are known and we can obtain closed-form expressions for the optimal feedback threshold values.

3.2 Feedback Thresholds for the MCS Algorithm

If we assume that the users have the same distribution of their CNRs with an average of $\overline{\gamma}$, the NFL for the MCS algorithm can be expressed as:

$$\bar{\mathrm{F}}_{\mathrm{MCS}} = \frac{1}{N} \sum_{l=0}^{L-1} \sum_{n=1}^{N} n \binom{N}{n} (P_\gamma(\gamma_{\mathrm{th},l+1}) - P_\gamma(\gamma_{\mathrm{th},l}))^n \cdot P_\gamma^{N-n}(\gamma_{\mathrm{th},l}), \qquad \text{(B.2)}$$

where $P_\gamma(\gamma)$ is the CDF of the CNR for a single user and $\gamma_{\mathrm{th},l}$ denotes the lth threshold value. This expression was found by calculating the expected number of users that give feedback for each threshold value, and summing all these feedback loads. The expression is normalized by dividing by the number of users. Using the binomial expansion formula [17], (B.2) can be written as:

$$\bar{\mathrm{F}}_{\mathrm{MCS}} = \sum_{l=0}^{L-1} \left(P_\gamma(\gamma_{\mathrm{th},l+1}) - P_\gamma(\gamma_{\mathrm{th},l})\right) \cdot P_\gamma^{N-1}(\gamma_{\mathrm{th},l+1}). \qquad \text{(B.3)}$$

A plot of the NFL for the optimal threshold values is shown in Fig. B.1 as a function of L for different number of users with Rayleigh channels with $\overline{\gamma} = 15$ dB. We see that the NFL converges to $1/N$ as L grows large. This is logical since the more thresholds there are in the system, the more likely is it that only one user will have a CNR value between two adjacent thresholds. To investigate how the feedback load scales with the number of users, we have also plotted the *absolute feedback load* (AFL) in Fig. B.2. The AFL expresses the average number of users transmitting feedback and we can observe that the decrease in AFL as a function of the number of thresholds is higher for a high number of users.

To optimize the thresholds, we take the gradient of (B.3) with respect to the threshold values and set it equal to zero, which gives the following

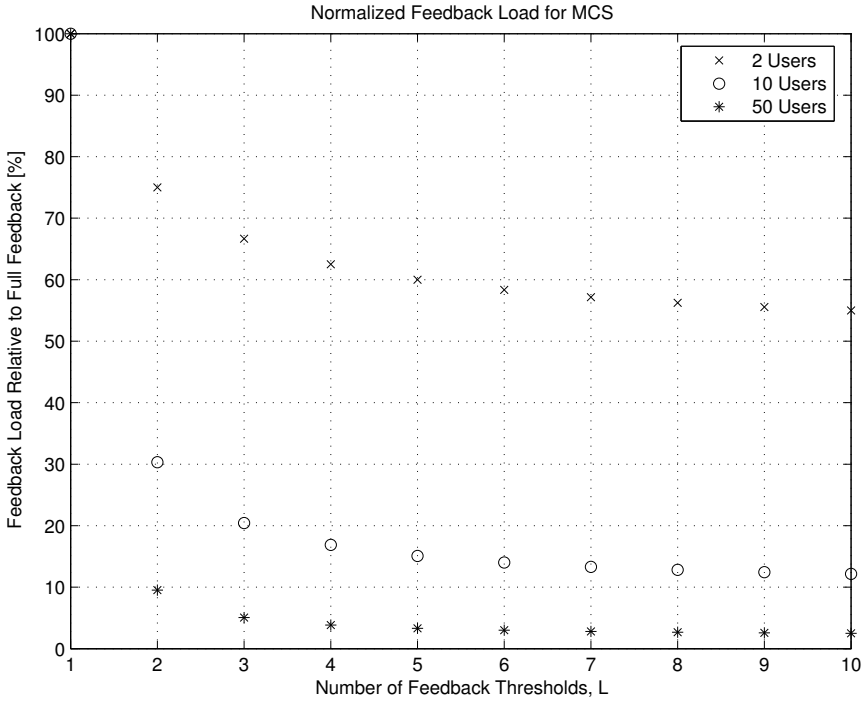

FIGURE B.1: Minimum normalized feedback load for the MCS algorithm as a function of L for different number of users with Rayleigh fading channels with $\overline{\gamma} = 15$ dB.

expression for the optimal threshold values:

$$\gamma^*_{\text{th},l} = P_\gamma^{-1}\left(S_l \cdot P_\gamma(\gamma^*_{\text{th},l+1})\right), \quad l = 1,2,3,\cdots, L-1, \tag{B.4}$$

where $P_\gamma^{-1}(\cdot)$ is the inverse CDF of the CNR for a single user, and the constants S_l are given by:

$$S_l = \begin{cases} N^{\frac{1}{1-N}}, & l = 1 \\ [N-(N-1)S_{l-1}]^{\frac{1}{1-N}}, & l = 2,3,\cdots, L-1, \end{cases} \tag{B.5}$$

with $N \geq 2$. The set of equations in (B.4) has a recursive nature. One way to calculate these threshold values is to start by calculating $\gamma_{\text{th},L-1}$. This value can easily be found since $\gamma_{\text{th},L}$ is defined to be infinity. Knowing $\gamma_{\text{th},L-1}$, (B.4) can be used to calculate all threshold values down to $\gamma_{\text{th},1}$. It is also

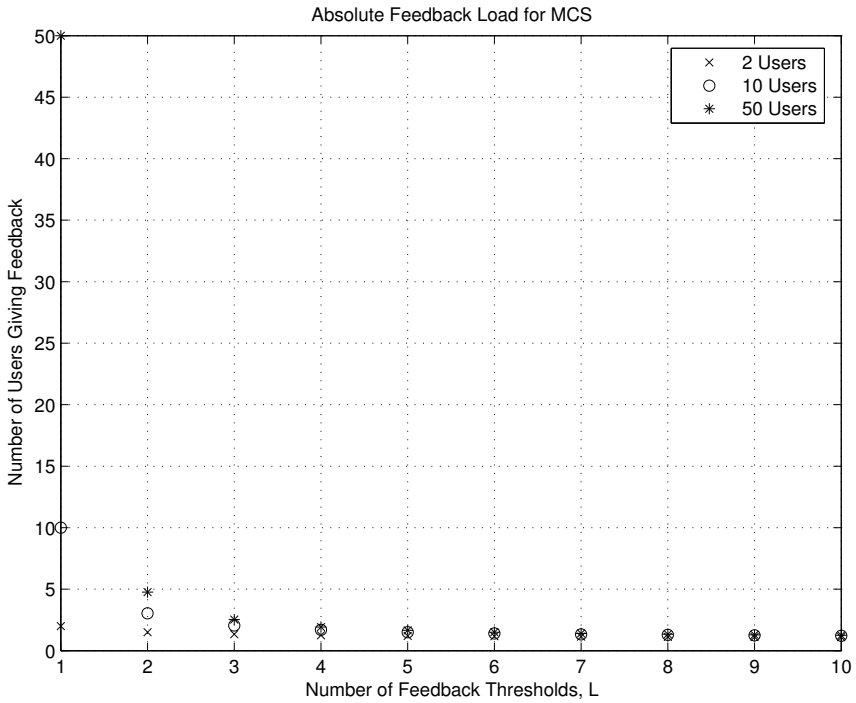

FIGURE B.2: Minimum absolute feedback load for the MCS algorithm as a function of L for different number of users with Rayleigh channels with $\overline{\gamma} = 15$ dB.

possible to express the threshold values as the sum of the average CNR and a constant (in dB). By writing (B.4) in the form:

$$P_\gamma(\gamma^*_{\text{th},l}) = S_l \cdot P_\gamma(\gamma^*_{\text{th},l+1}), \quad l = 1, 2, 3, \cdots, L-1, \tag{B.6}$$

and exploiting the fact that $P_\gamma(\gamma_{\text{th},L}) = 1$, we can write (B.4) as:

$$\gamma^*_{\text{th},l} = P_\gamma^{-1}\left(\prod_{i=l}^{L-1} S_i\right), \quad l = 1, 2, 3, \cdots, L-1. \tag{B.7}$$

For the Rayleigh, Nakagami, and Rice distributions, the inverse CDF $P_\gamma^{-1}(\cdot)$ equals $\overline{\gamma}$ multiplied by a constant which is only dependent on the number of users [18]. Consequently, the threshold values in dB will be a sum of $\overline{\gamma}$ and a constant.

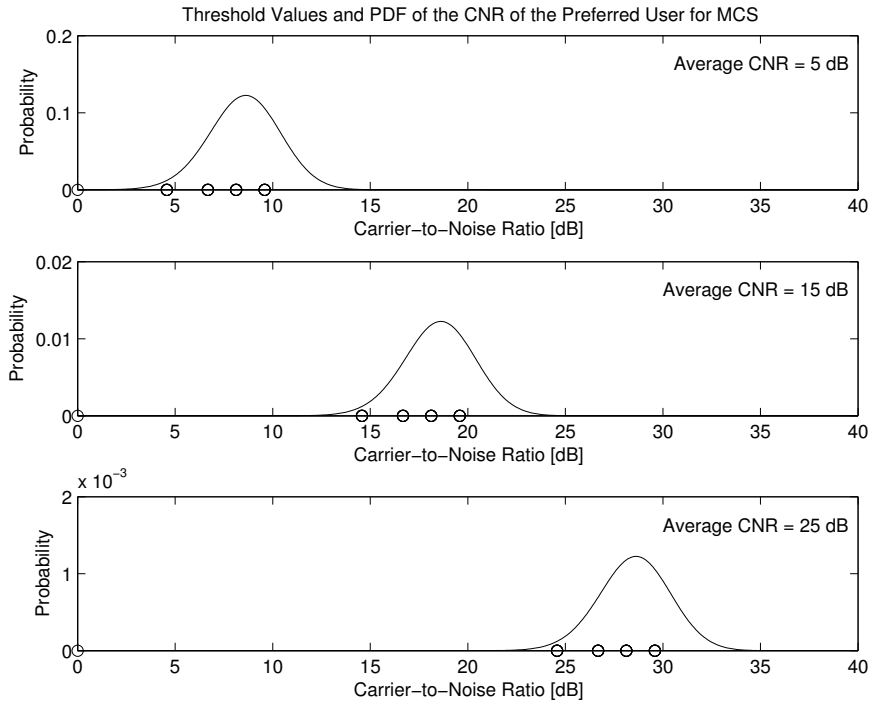

FIGURE B.3: Three sets of feedback threshold values for the MCS algorithm for $L = 5$ and 10 users with Rayleigh fading channels with $\overline{\gamma} = 5$ dB, $\overline{\gamma} = 15$ dB, and $\overline{\gamma} = 25$ dB, respectively. The PDF of the CNR of the user with the highest CNR is also shown for $\overline{\gamma} = 5$ dB, $\overline{\gamma} = 15$ dB, and $\overline{\gamma} = 25$ dB, respectively.

Fig. B.3 shows three sets of threshold values for three cases where we have ten users having Rayleigh distributed channels with $\overline{\gamma} = 5$ dB, $\overline{\gamma} = 15$ dB, and $\overline{\gamma} = 25$ dB, respectively. The threshold values are identical for all users and are shown as small rings. Each set of threshold values contains five CNR values ($L = 5$). By comparing the threshold values of the three different sets, we see that the threshold values in dB are a sum of $\overline{\gamma}$ and a constant. The probability density functions (PDF) of the best user among ten users is also shown for each of the three $\overline{\gamma}$-values. These plots show that the probability of finding the best user below $\gamma_{\text{th},1}$ is quite small. Consequently, the probability of full feedback is low.

3.3 Feedback Thresholds for the NCS Algorithm

The PDFs of the scheduling metric $x_i(t) = \chi_i(t) = \gamma_i(t)/\overline{\gamma}_i$ can be obtained by doing a transformation of the PDF of the CNR of user i, γ_i [17, (2.1.8)]. If the Rayleigh, Nakagami or Rice distributions listed in [18] are used in this transformation, it can be shown that the resulting PDFs are independent of i. Therefore, if we assume that the users' CNRs have the same distributions with different averages, it can be shown that the PDF of $\chi_i(t)$ is the same for all users. Since this PDF and the corresponding CDF are independent of i, they can be denoted $p_\chi(\chi)$ and $P_\chi(\chi)$, respectively. Using these distributions for χ, we can obtain an expression for the NFL in similar way as in the previous section:

$$\bar{\mathrm{F}}_{\mathrm{NCS}} = \sum_{l=0}^{L-1} \left(P_\chi(\chi_{\mathrm{th},l+1}) - P_\chi(\chi_{\mathrm{th},l}) \right) \cdot P_\chi^{N-1}(\chi_{\mathrm{th},l+1}), \tag{B.8}$$

where $\chi_{\mathrm{th},l}$ denotes the lth threshold value. Taking the gradient of (B.8) with respect to the threshold values, we obtain a similar expression for the threshold values as we did for the MCS feedback threshold values:

$$\chi^*_{\mathrm{th},l} = P_\chi^{-1}\left(S_l \cdot P_\chi(\chi^*_{\mathrm{th},l+1}) \right), \quad l = 1, 2, 3, \cdots, L-1, \tag{B.9}$$

where $P_\chi^{-1}(\cdot)$ is the inverse CDF of $\chi(t)$, and the constants S_l are given by (B.5).

The plots of the NFL and AFL for the optimal threshold values in (B.8) as a function of L for different number of users with Rayleigh distributed channels, is identical to Figs. B.1 and B.2, respectively. It can be shown that the identical feedback load for the MCS and NCS algorithms arises as a consequence of the similarities between the CDFs of the scheduling metrics when the users have the same average CNR for the MCS algorithm. Since the feedback thresholds are adapted to the scheduling metric of the NCS algorithm, the feedback load is independent of the average CNRs of the users.

Fig. B.4 shows five feedback threshold values of the NCS algorithm for ten users with Rayleigh fading channels with different average CNRs. The corresponding PDF of χ for the preferred user is also shown. It should be noted that the threshold values are identical for all users. However, if the threshold values are converted to the corresponding CNR values, $\gamma_{\mathrm{th},i,l} = \overline{\gamma}_i \chi_{\mathrm{th},l}$, they will differ from user to user.

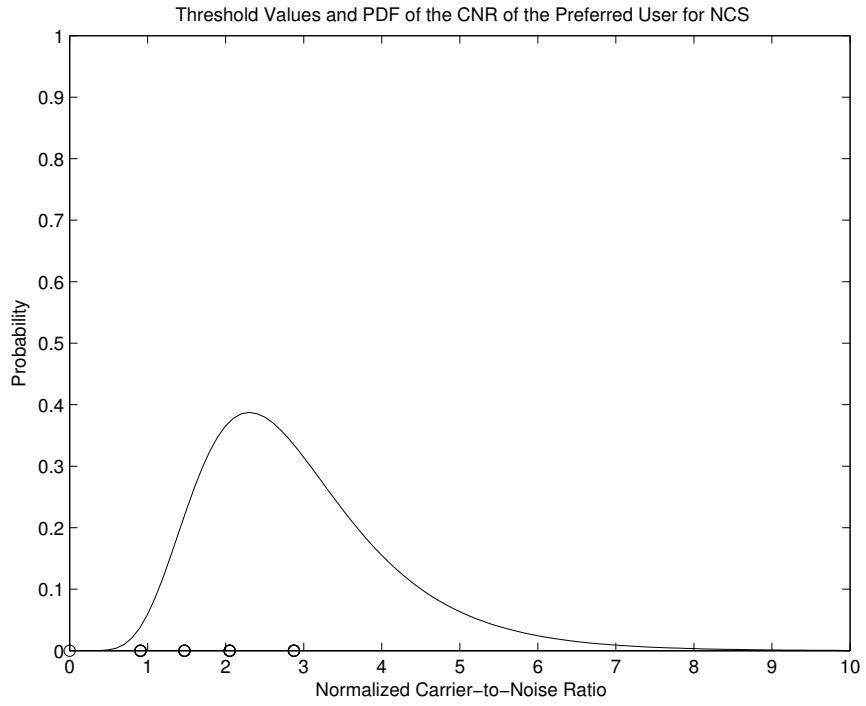

FIGURE B.4: One set of feedback threshold values for the NCS algorithm for $L = 5$ and 10 users with Rayleigh fading channels with different average CNRs. The PDF of the normalized CNR of the user with the highest normalized CNR is also shown.

4 A Two-Step Procedure for Optimizing the Threshold Values and the Number of Thresholds

In the previous section, it was assumed that the time it takes for the scheduler to conduct the polling process, take a scheduling decision, and distribute this decision is negligible. In practical systems this process will have to be conducted within a *guard time* at the beginning of the time-slots. To have the highest possible utilization of the system, we thus want that (a) feedback is received from the preferred user (or subset of users), (b) the power consumption of the users is minimized and (c) the guard time is reduced. As previously explained, both (a) and (b) are achieved by using our feedback algorithm. The guard time can be split into two components, namely, (i) the time used to poll the users with lower and lower feedback

thresholds, and (ii) the time used to receive feedback from one or more users. Both the number of users, the value of L, and the threshold values will affect these two time contributions. Setting all the thresholds to zero will minimize (i). However, this will maximize the feedback load and hence also (a). Consequently, the threshold values have to be set to non-zero values and thus (i) is strongly dependent on the number of users and the value of L.

Based on the discussion above, we see that it is often favorable both for the power consumption and the guard time length that the threshold values are set to minimize the feedback load. However, the guard time will also be affected by the value of L. We therefore propose a two-step optimization procedure to obtain the threshold values and the value of L. In the first step, the threshold values are set to those who minimize the feedback load. In the second step, the value of L that minimizes the guard time is found numerically, based on the threshold values from step one. In [13], we performed the two-step optimization procedure described above, for an IEEE 802.11 system, and we refer to [13] for numerical results on this procedure.

5 Conclusions

We have proposed a new channel state feedback algorithm for modern cellular networks. Compared to previously published works, our algorithm is based on two novel concepts, namely, (i) adapting the feedback threshold value to the scheduling algorithm implemented in the system, and (ii) employing multiple feedback thresholds to reduce the number of users transmitting feedback to a minimum. Our feedback algorithm leads to a significant decrease in the power consumption of the mobile users and also in the time used to collect feedback for many systems. The proposed feedback algorithm can be implemented for any scheduling metric, but in most cases the optimal threshold values have to be found numerically. However, for the MCS and the NCS algorithms we obtained elegant closed-form expressions for the optimal threshold values. Finally, we proposed a two-step optimization procedure for obtaining the threshold values and the number of thresholds in real-life wireless networks.

References

[1] W. T. Webb and R. Steele, "Variable-rate QAM for mobile radio," *IEEE Trans. Commun.*, vol. 43, pp. 2223–2230, July 1995.

[2] R. Knopp and P. A. Humblet, "Information capacity and power control in single cell multiuser communications," in *Proc. IEEE Int. Conf. on Communications (ICC'95)*, (Seattle, WA, USA), pp. 331–335, June 1995.

[3] WiMAX Forum, "Mobile WiMAX – Part II: A Comparative Analysis." `http://www.wimaxforum.org/news/downloads/Mobile_WiMAX_Part2_Comparative_Analysis.pdf`, May 2006.

[4] A. Farrokh and V. Krishnamurthy, "Opportunistic scheduling for streaming users in high-speed downlink packet access (HSDPA)," in *Proc. IEEE Global Communications Conf. (GLOBECOM'04)*, vol. 6, (Dallas, TX, USA), pp. 4043–4047, Nov.-Dec. 2004.

[5] Y.-J. Choi and S. Bahk, "QoS scheduling for multimedia traffic in packet data cellular networks," in *Proc. IEEE Int. Conf. on Communications (ICC'03)*, vol. 1, (Anchorage, AK, US), pp. 358–362, May 2003.

[6] F. Florén, O. Edfors, and B.-A. Molin, "The effect of feedback quantization on the throughput of a multiuser diversity scheme," in *Proc. IEEE Global Communications Conf. (GLOBECOM'03)*, vol. 1, (San Francisco, CA, USA), pp. 497–501, Dec. 2003.

[7] M. Johansson, "Benefits of multiuser diversity with limited feedback," in *Proc. IEEE Workshop on Signal Processing Advances in Wireless Communications (SPAWC'03)*, (Rome, Italy), pp. 155–159, June 2003.

[8] K. K. Mukkavilli, A. Sabharwal, E. Erkip, and B. Aazhang, "On beamforming with finite rate feedback in multiple-antenna systems," *IEEE Trans. Inform. Theory*, vol. 49, pp. 2562–2579, Oct. 2003.

[9] H. Cheon, B. Park, and D. Hong, "Adaptive multicarrier system with reduced feedback information in wideband radio channels," in *Proc. IEEE Vehicular Technology Conf. (VTC'99-fall)*, vol. 5, (Amsterdam, The Netherlands), pp. 2880–2884, Sept. 1999.

[10] H. Nguyen, J. Brouet, V. Kumar, and T. Lestable, "Compression of associated signaling for adaptive multi-carrier systems," in *Proc. IEEE Vehicular Technology Conf. (VTC'04-spring)*, vol. 4, (Rome, Italy), pp. 1916–1919, May 2004.

[11] X. Qin. and R. Berry, "Opportunistic splitting algorithms for wireless networks," in *Proc. IEEE International Conference on Computer Communications (INFOCOM'04)*, (Hong Kong, P. R. China), pp. 1662 – 1672, Mar. 2004.

[12] D. Gesbert and M.-S. Alouini, "How much feedback is multi-user diversity really worth?," in *Proc. IEEE Int. Conf. on Communications (ICC'04)*, vol. 1, (Paris, France), pp. 234–238, June 2004.

[13] H. Koubaa, V. Hassel, and G. E. Øien, "Multiuser diversity gain enhancement by guard time reduction." In *Proc. IEEE Workshop on Signal Processing Advances in Wireless Communications (SPAWC'05)*, (New York, NY, USA), June 2005.

[14] L. Yang and M.-S. Alouini, "Performance analysis of multiuser selection diversity," in *Proc. IEEE Int. Conf. on Communications (ICC'04)*, vol. 5, (Paris, France), pp. 3066–3070, June 2004.

[15] P. Viswanath, D. N. C. Tse, and R. Laroia, "Opportunistic beamforming using dumb antennas," *IEEE Trans. Inform. Theory*, vol. 48, pp. 1277–1294, June 2002.

[16] R. Suoranta, K.-P. Estola, S. Rantala, and H. Vaataja, "PDF estimation using order statistic filter bank," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Proc. (ICASSP'94)*, vol. 3, (Adelaide, Australia), pp. III–625–III–628, Apr. 1994.

[17] G. Casella and R. L. Berger, *Statistical Inference*. Belmont, CA, USA: Duxbury Press, 1990.

[18] H. Holm, G. Øien, M.-S. Alouini, D. Gesbert, and K. J. Hole, "Optimal design of adaptive coded modulation schemes for maximum average spectral efficiency," in *Proc. IEEE Workshop on Signal Processing Advances in Wireless Communications (SPAWC'03)*, (Rome, Italy), pp. 403–407, June 2003.

Paper C

Feedback Protocols for Increased Multiuser Diversity Gain in Cellular ALOHA-Based Networks – A Comparative Study

Vegard Hassel, Hend Koubaa, and Geir E. Øien

Technical Report, NTNU.
Published at *http://www.diva-portal.org/ntnu/*, January 2007

Abstract

Multiuser diversity (MUD) underlies much of the recent work on scheduling design in wireless networks. This form of diversity can for example be exploited by opportunistically scheduling the mobile user with the best channel quality [1]. In cellular networks exploiting MUD, the base station collects channel state information (CSI) from the mobile users. The process of obtaining CSI will be performed within a *guard time*, and the length of this guard time will depend on the feedback protocol implemented. In this context, it has already been shown that by applying multiple carrier-to-noise ratio thresholds, the number of mobile users giving feedback can be significantly decreased [2]. However, it has not been evaluated how the algorithm in [2] can be implemented in protocols for real-life networks. In this paper we analyze feedback protocols for reducing the guard time and resolving the feedback contention problem in a cellular, slotted ALOHA-based network. We propose three new feedback protocols based on the algorithm in [2] and we develop closed-form expressions for the guard time duration and the system spectral efficiency of these protocols. We also compare the three new protocols with the Splitting algorithm proposed by Qin and Berry [3] and a new and modified version of this algorithm. Plots show that the spectral efficiency in an IEEE 802.11 network can increase significantly for a high number of users when the Modified Splitting algorithm is used.

1 Introduction

In a wireless network, the signals transmitted from a base station to different mobile users often have different channel fluctuation characteristics. This diversity that exists between the mobile users is called *multiuser diversity* (MUD) and can be exploited to increase the throughput of wireless networks [1]. One way of exploiting MUD is by means of *opportunistic scheduling* of users, giving priority to users having favorable channel conditions [4, 5]. The Max Carrier-to-noise Scheduling (MCS) algorithm, where the user with the best channel quality is scheduled in each time-slot, maximizes the MUD in a time-slotted network. To be able to take advantage of the MUD, the base station needs feedback from the mobile users about their respective channel conditions. If the MCS algorithm is used, the base station only needs feedback from the user with the best channel conditions, but unfortunately each user does not know the carrier-to-noise ratio (CNR) of the other users. Therefore, in current cellular standards like Qualcomm's High Data Rate (HDR) system, the base station collects feedback from all the users [6]. In a time-slotted cellular network that exploits MUD, the base station can use the first part of the time-slot to collect feedback from the users and to decide which user to schedule [3]. We call this first part of the time-slot the *guard time*. Collecting feedback from all the users in a system can lead to a significant guard time and hence it is important to investigate alternative protocols for obtaining feedback.

One way of reducing the guard time is by implementing feedback algorithms that utilize *CNR thresholds* to reduce the number of users giving feedback and still be able to exploit MUD. At least two different types of such threshold-based feedback algorithms have already been proposed. The first type was initially proposed by Gesbert and Alouini and is based on a single CNR threshold value [7]. The users that have a CNR above this value give feedback to the scheduler. This algorithm does not always find the user with the highest CNR because there will always be a possibility that all users are below the threshold value, and in this case a random user is chosen. A generalized version of this algorithm has also been proposed [2]. By using several threshold values, the scheduler can request feedback in a successive fashion starting out with the highest of the threshold values. If the lowest threshold value is zero, the user with the highest CNR will always be found.

The second type of threshold-based feedback algorithm was proposed by Qin and Berry and is based on the ideas from binary search [3]. The proposed *Splitting algorithm* finds the user with the best channel quality by using an iterative procedure to update two CNR threshold values when the

users are using a common ALOHA channel.

Contributions. For the Splitting algorithm, the guard time has already been analyzed for a slotted ALOHA channel. However, the multiple threshold algorithm in [2] has not yet been analyzed for a slotted ALOHA channel and it is therefore hard to decide which of the two threshold-based algorithms that perform best. In this paper we propose three new cellular ALOHA protocols for the algorithm in [2] and compare the performance of these algorithms with the Splitting algorithm as well as with a new and modified version of the Splitting algorithm [1].

Organization of the paper. The remainder of this paper is organized as follows. We outline the system model and the problem formulation in Section 2, and present the five feedback protocols under study in Section 3. In Section 4 and Section 5 we develop analytical expressions for the guard time and the system spectral efficiency, respectively. Section 6 discusses how the protocols should be optimized and presents plots comparing the guard time and the system spectral efficiency of the resulting five feedback protocols in an IEEE 802.11 network. Finally, our conclusions are listed in Section 7.

2 System Model and Problem Formulation

2.1 General System Model

We consider the downlink of a single-carrier cellular network where the base station wants to transmit data to N mobile users which have identically and independently distributed (i.i.d.) CNRs with an average of $\overline{\gamma}$. The channel is ALOHA-based, i.e., all the users can access the network at the same time. When more users transmit packets simultaneously, this will result in a collision and the information in the packets will be destroyed. The system uses time-slotted transmission and for each time-slot with duration T_{TS}, the base station schedules a user which will receive data. We assume slowly varying fading channels with a coherence time that is longer than one time-slot. This means that the same transmission rate is used for the whole time-slot. The system uses adaptive coding and modulation, i.e., the coding scheme and the modulation constellation used depend on the CNR of the selected user [10]. This has two advantages. On one hand, the spectral efficiency for each user is increased. On the other hand, because the rate of the users are varied according to their channel conditions, it makes it possible to exploit MUD.

[1] This paper is partially based on the work in [8] and [9]

To be able to select the user which will receive data, the base station needs to receive channel state information (CSI) estimates from one or more users. Such CSI estimates can be obtained from pilot symbols that are transmitted in-between the data symbols. For the three feedback protocols that are based on [2], we use L feedback thresholds denoted by $\gamma_{\text{th},L} > \gamma_{\text{th},L-1} > \cdots > \gamma_{\text{th},0}$ to search for the users in a sequential manner. For convenience we define $\gamma_{\text{th},L} = \infty$ and $\gamma_{\text{th},0} = 0$, so that we can search for mobile users within the whole CNR range. Initially, we search for users that have a CNR above $\gamma_{\text{th},L-1}$. If no users are found, the feedback threshold is lowered to $\gamma_{\text{th},L-2}$, and we search for users that have a CNR above this threshold. The algorithm lowers the threshold value sequentially until one or more users are found. We denote the CNR interval where the first user is found as the *successful interval* and process of checking for users within one interval is referred to as a *trial*.

2.2 Further Specifications for an IEEE 802.11-Based Network

We want to investigate the gain from using multiple feedback thresholds in a cellular IEEE 802.11 network [11]. In such networks, the access mechanism is ALOHA-based, and one of the main problems that can arise in such networks is *collisions* between packets. To avoid collisions, a handshaking mechanism is often used between the transmitter and the receiver before starting any data transmission. The transmitter sends a *Request To Send* (RTS) packet to the receiver asking if he can transmit. The receiver replies with a *Clear to Send* (CTS) packet if he is ready for data reception. If we want to deploy the proposed feedback protocols in an IEEE 802.11 network, we can use packets similar to RTS and CTS to conduct the feedback collection process. Consequently, we define four different packets based on the general frame format defined in the IEEE 802.11 standard [11]:

- Query (QRY) packet
- Feedback (FB) packet
- Reservation (RES) packet
- Acknowledgment (ACK) packet

The QRY packet is used by the base station to initiate the feedback collection process. This packet contains the addresses of all the users that have data to receive and the number of thresholds L applied. As shown in [2], each of the users can calculate the feedback threshold values from the number of users N, the number of thresholds L, and the average CNR $\overline{\gamma}$ of the users. When all the users have calculated the threshold values, the feedback collection process can start. We denote the duration of this packet,

including the packet processing time and the propagation delay, as T_{QRY} [seconds].

The FB packet is transmitted by the mobile users and contains the CSI estimate of a user's channel. This packet is also used for all the five protocols handled in this paper. Including packet processing time and propagation delay, this packet has the duration T_{FB} [seconds].

The RES packet is transmitted by a mobile user to inform the base station that he is not in the successful interval (Ranked Single-User Feedback protocol) or that he has a CNR between the two current threshold values (Splitting algorithm). Although the RES packet does not contain any information (See Section 6.1), it makes the base station able to detect if one or more users are between two threshold values. The total time to transmit and process this packet is denoted T_{RES} [seconds].

The ACK packet is transmitted by the base station to inform all the mobile users in the system about the status of a recent FB or RES packet transmission. If no packets were transmitted, this packet contains *0*, while for a successful packet transmission the ACK contains *1*. However, when two packets have collided, this packet contains *e*, denoting an erroneous transmission. It should be noted that not all FB and RES packets need no be followed by an ACK packet. The aggregated transmission and packet processing time of this packet is denoted T_{ACK} [seconds].

In IEEE 802.11-based networks, all these packets are transmitted at the base rate of the system and we assume that the bit error probability of these packets are zero.

2.3 Problem Formulation

The main goal of this paper is to propose and analyze three protocols based on the feedback algorithm proposed in [2] and compare these protocols with the Splitting algorithm, both in its original and modified version, for an IEEE 802.11-based network. We want to evaluate the different feedback protocols according to their *Maximum Average System Spectral Efficiency* (MASSE) performance. The MASSE [bits/sec/Hz] is defined as the maximum average spectral efficiency that is possible within a cell, averaged over all the N mobile users. To be able to investigate the MASSE, the guard time, i.e., the duration of the feedback collection process, has to be quantified. This guard time analysis will be conducted in Section 4.

3 Proposed Feedback Protocols

In this section we will give an overview of the five different feedback protocols handled in this paper. The first three protocols are new and are based on the algorithm proposed in [2]. The fourth protocol is the Splitting algorithm introduced in [3] and the fifth protocol is a new and modified version of this algorithm.

3.1 Ranked Full Feedback

For this protocol, all the users that are above the current threshold value are allowed to transmit their CSI estimate simultaneously. For the first trial the users that have a CNR above $\gamma_{\text{th},L-1}$ are allowed to transmit feedback. If there are none, the threshold is successively lowered to $\gamma_{\text{th},L-2}$, $\gamma_{\text{th},L-3}$, $\cdots$, $\gamma_{\text{th},0}$. Consequently, the threshold is successively lowered until feedback is successfully transmitted or a collision occurs. Each trial is assigned the duration $T_{\text{FB}} + T_{\text{ACK}}$, so that an FB packet followed by an ACK packet can be transmitted. Thanks to the ACK, all the users in the system will be informed if other users transmitted feedback.

If a feedback transmission happens without a collision (ACK=*1*), the guard time is over. However, if a collision occurs (ACK=*e*), the contention problem is solved by letting all the users transmit their feedback sequentially depending on their *rank* in the system. The rank is simply an ordering pre-assigned by the base station. All the users will transmit their feedback to the base station during a time $N \cdot T_{\text{FB}}$; hence the user with the highest CNR is guaranteed to be found, which will maximize the MUD gain in the cell.

3.2 Ranked Single-User Feedback

As for the Ranked Full Feedback protocol, the Ranked Single-User Feedback protocol also lowers the threshold values in the same successive fashion, giving all the users the opportunity to transmit their feedback simultaneously for each trial. The duration $T_{\text{FB}} + T_{\text{ACK}}$ is assigned to each trial and the guard time is over if a successful FB packet transmission occurs. However, instead of letting all users transmit their feedback if a collision occurs, only the user with the highest rank within the successful interval transmits his feedback. When a collision occurs, the user with the highest rank is first given the opportunity to transmit his FB packet. If this user is within the successful interval, the FB packet is transmitted, a *1*-ACK packet is broadcasted, and the guard time is over. However, if a user is not within

the successful interval, he transmits a RES packet and the base station will broadcast an ACK=*0* to inform the other users that this user's transmission is finished. Now, the user with the second highest rank will be given the opportunity to transmit an FB packet. This process continues until one of the users have transmitted an FB packet and the base station has broadcasted a *1*-ACK. For this protocol, the base station will not receive CSI feedback from all the users in the cell and, hence the user with the highest CNR is not always scheduled. Consequently, a certain MUD degradation will be experienced. However, the guard time will decrease, which will contribute to an increase in the overall MASSE. This protocol can also lead to an unfairness problem: If the rank of the users is fixed, the users with the highest rank will on average be selected more often than the users with lower rank. To have a more fair protocol, the rank of the users can be changed from time to time.

3.3 Exponential Backoff

For this protocol, as for the two protocols above, all the users are given the opportunity to transmit their FB packets simultaneously for each trial until a successful feedback transmission or a collision occurs. Each trial has the duration $T_{\text{FB}} + T_{\text{ACK}}$. For this protocol, the contention problem is solved by using a tailored version of the Exponential Backoff scheme [12]. If only one user is above a threshold, he will successfully feed back his CSI and the guard period will be over. However, if a collision takes place, the feedback transmission probability is lowered for the users within the successful interval and these users are again given the possibility to transmit their feedback within a time T_{FB}. After this time period the base station broadcasts an ACK packet to inform the users about the status of the feedback collection process. If more collisions are experienced (ACK=*e*), the transmission probability for the users within the successful interval is lowered one more time. The transmission probability is not changed if no users are transmitting feedback (ACK=*0*). This process will continue until one user has conducted a successful feedback transmission (ACK=*1*).

It can be shown that for n users contending, $1/n$ will be the transmission probability that maximizes the probability for a successful transmission. In [2] it has also been shown that the most probable number of users participating in a collision is two. Consequently, for the Exponential Backoff protocol the transmission probability is halved for each feedback collision. This protocol gives an increase in the fairness since a random user within the successful interval transmits feedback. However, the user with

the highest CNR is not always feeding back his CSI and the MUD gain is not maximized.

3.4 Splitting Algorithm

The Splitting algorithm was proposed by Qin and Berry in [3] and uses principles from binary search to look for the user with the highest CNR. This protocol uses two threshold values and the users that have a CNR in the interval between these thresholds should transmit a RES packet simultaneously. The goal is that only the user with the best channel quality should be captured between the two thresholds. Initially, the highest threshold equals infinity and lowest threshold equals the value that maximizes the probability of having one user in the interval between the two thresholds. If more users have a CNR in the interval between the two thresholds, the base station broadcasts an *e*-ACK and the interval is split in two by increasing the lowest threshold value. However, if no users transmit a RES packet within the interval, a *0*-ACK is broadcasted and the highest threshold value is set to the lowest threshold value and the lowest threshold value is lowered. If only one user transmits his RES packet, the base station knows that this is the user with the highest CNR and a *1*-ACK is broadcasted. Finally, this user can transmit his CSI estimate by using an FB packet and the guard time is over. In [3] it is proven that maximally 2.5 iterations are needed on average to find the user with the best channel quality.

3.5 Modified Splitting Algorithm

As will be clear from Section 6.1, the RES packet is only slightly shorter than the FB packet in an IEEE 802.11-based network. We therefore propose a modification to the Splitting algorithm where an FB packet is used for the iteration process instead of a RES packet. For this protocol the iteration process will be slightly longer than for the original Splitting algorithm, however; the total guard time for an IEEE 802.11-based network will be shorter since it is not necessary to transmit an FB packet after the iteration process.

4 Guard Time Analysis

The goal of this section is to develop analytical expressions for the guard time for the Ranked Full Feedback protocol, the Ranked Single-User Feedback protocol, and the Exponential Backoff protocol. These guard time ex-

pressions will be needed in the expressions for the MASSE (See Section 5). To make the analysis simpler, we assume that the duration of the QRY packet is zero. Since the QRY broadcast time is the same for all the feedback protocols described above, this assumption will not affect the difference in guard time between the different protocols. Even if feedback is requested from all the users, a similar QRY packet needs to be broadcasted to inform the users about the order of their feedback transmission, since the users that have data to receive can change from time-slot to time-slot.

For the three proposed feedback protocols based on [2], the number of intervals checked before the successful interval is reached, is identical. The number of threshold values checked *before* the successful interval is found (number of trials), denoted M, will influence the guard time significantly. M can be modeled as a discrete random variable, and the probability that M has the value l can be expressed as follows:

$$\Pr(M = l) = P_\gamma^N(\gamma_{\text{th},L-l}) - P_\gamma^N(\gamma_{\text{th},L-l-1}),\ l = 0, 1, \cdots, L-1, \tag{C.1}$$

where $P_\gamma(\cdot)$ is the cumulative distribution function (CDF) of the CNR for one user. This equation expresses the probability of one or more users being in interval l while the rest of the users have lower CNR levels. The expected number of trials before the successful interval can now be expressed as:

$$\mathrm{E}[M] = \sum_{l=0}^{L-1} l\,[P_\gamma^N(\gamma_{\text{th},L-l}) - P_\gamma^N(\gamma_{\text{th},L-l-1})], \tag{C.2}$$

where $\mathrm{E}[\cdot]$ denotes the expectation operator.

4.1 Guard Time for Ranked Full Feedback

The time duration after the successful interval is found can be expressed as the sum of $T_{\text{G,coll},l}$ and $T_{\text{G,nocoll},l}$, where the former is the guard time contribution in the case a collision takes place in the successful interval l and the latter is the guard time contribution in the case only one user is found in the successful interval l. The expected values of these guard time contributions can be expressed as:

$$\mathrm{E}[T_{\text{G,coll},l}] = [(N+1)T_{\text{FB}} + T_{\text{ACK}}] \cdot \sum_{n=2}^{N} p(l, n), \tag{C.3}$$

and

$$\mathrm{E}[T_{\text{G,nocoll},l}] = (T_{\text{FB}} + T_{\text{ACK}}) \cdot p(l, 1), \tag{C.4}$$

where $p(l,n)$ denotes the joint probability mass function (PMF) associated with the event of having n users in the successful interval l, i.e., $< \gamma_{\text{th},l}, \gamma_{\text{th},l+1}]$ [2]:

$$p(l,n) = \binom{N}{n} (P_\gamma(\gamma_{\text{th},l+1}) - P_\gamma(\gamma_{\text{th},l}))^n (P_\gamma(\gamma_{\text{th},l}))^{N-n}. \quad \text{(C.5)}$$

Now, the total expected guard time for the Ranked Full Feedback protocol can be expressed as:

$$\mathrm{E}[T_\mathrm{G}] = (T_\mathrm{FB} + T_\mathrm{ACK}) \cdot \mathrm{E}[M] + \sum_{l=0}^{L-1} \mathrm{E}[T_{\mathrm{G,coll},l}] + \sum_{l=0}^{L-1} \mathrm{E}[T_{\mathrm{G,nocoll},l}], \quad \text{(C.6)}$$

for $L > 1$. For $L = 1$, all users will be within the successful interval. Therefore, collisions can be avoided, and the guard time equals the guard time for the Full Feedback protocol, $T_\mathrm{G} = N \cdot T_\mathrm{FB}$.

4.2 Guard Time for Ranked Single-User Feedback

As for the Ranked Full Feedback protocol, the time duration after the successful interval l is found can be expressed as the sum of the time contributions $T_{\mathrm{G,coll},l}$ and $T_{\mathrm{G,nocoll},l}$. The expected time contribution from the case where no collision takes place, $T_{\mathrm{G,nocoll},l}$, is the same as for the Ranked Full Feedback protocol given in (C.4). The expression for the time contribution in the case of a collision yields:

$$\begin{aligned} \mathrm{E}[T_{\mathrm{G,coll},l}] &= 2(T_\mathrm{FB} + T_\mathrm{ACK}) \sum_{n=2}^{N} p(l,n) \\ &+ (T_\mathrm{RES} + T_\mathrm{ACK}) \sum_{n=2}^{N} \sum_{k=0}^{N-n} k \binom{N-k-1}{n-1} \\ &\times (P_\gamma(\gamma_{\text{th},l+1}) - P_\gamma(\gamma_{\text{th},l}))^n P_\gamma(\gamma_{\text{th},l})^{N-n}, \end{aligned} \quad \text{(C.7)}$$

where the first factor appears because one FB-collision arises when the successful interval is found and one FB packet is transmitted because the user with the highest rank within the successful interval feeds back his CSI, while the second factor is derived in Appendix 1. The total expression for the expected guard time is the same as in (C.6). As for the Ranked Full Feedback protocol, the guard time expression is only valid for $L > 1$. For $L = 1$, only the user with the highest rank feeds back his CSI, which gives $T_\mathrm{G} = T_\mathrm{FB}$. This CSI estimate is used to adapt the coding and modulation.

4.3 Guard Time for Exponential Backoff

The Exponential Backoff scheme can be described by the Markov chain shown in Fig. C.1. Considering any successful interval l, we define the state $I = i$ as the number of collisions that have occurred. When the first collision occurs, the protocol goes to state $i = 1$ where the transmission probability is q^i. For each new collision, the state is incremented, and the time contribution from switching to a new state is T_{FB}+T_{ACK}. As mentioned in Section 3.3, the value of q is one half, so the transmission probability is halved for each state. The probability for successful feedback transmission in state $I = i$ is $\text{P}_{\text{succ}} = nq^i(1-q^i)^{n-1}$, where n denotes the number of contending users. Correspondingly, the probability that none of the users are transmitting feedback in state $I = i$ is equal to $\text{P}_{\text{stay}} = (1-q^i)^n$. Because the sum of all transition probabilities from one state equals unity, the probability for going to the next state is $\text{P}_{\text{next}} = 1-(1-q^i)^n - nq^i(1-q^i)^{n-1}$. The joint probability of entering state $I = i$, and having n contending users in the successful interval l, can be written as a sum of the probabilities of the mutually exclusive events in the previous state that lead to the next:

$$\begin{aligned}\pi(i,l,n) &= \pi(i-1,l,n)\cdot \text{P}_{\text{next}} \sum_{k=0}^{\infty} (\text{P}_{\text{stay}})^k \\ &= \pi(i-1,l,n)\frac{1-(1-q^{i-1})^n - nq^{i-1}(1-q^{i-1})^{n-1}}{1-(1-q^{i-1})^n}, \quad \text{(C.8)}\end{aligned}$$

for $i \geq 1$. For $n \geq 2$ and $i = 1$, $\pi(i,l,n)$ equals the probability that there are multiple users in the successful interval, consequently $\pi(1,l,n) = p(l,n)$. For $n = 1$, there are no collisions ($i = 0$) and we have $\pi(0,l,n) = p(l,n)$.

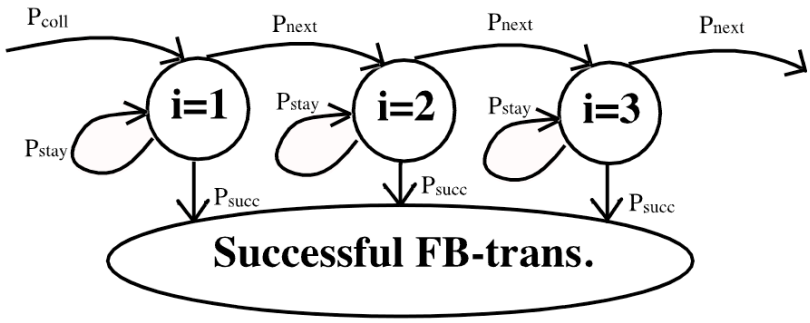

FIGURE C.1: Markov chain illustrating the exponential backoff scheme.

By nesting the recursive relationship in (C.8) down to $i = 2$ and using the relations $\pi(1,l,n) = p(l,n)$ and $\pi(0,l,n) = p(l,n)$, we obtain:

$$\pi(i,l,n) = p(l,n) \prod_{m=1}^{i-1} \frac{1-(1-q^m)^n - nq^m(1-q^m)^{n-1}}{1-(1-q^m)^n}, \qquad \text{(C.9)}$$

for $i \geq 0$ and $n \geq 1$. Note that the value $i = 0$ can only arise when $n = 1$, and that the product in this expression reduces to unity when $i = 0$ or $i = 1$. Now we can insert (C.5) into (C.9) and find all the transition probabilities $\pi(i,l,n)$ for any number of contending users n in any successful interval l.

To find the number of T_{FB}+T_{ACK} used due to no feedback transmission, we calculate the probability of staying k transmission attempts in state $I = j$:

$$\Pr(K_1 = k) = (1 - \text{P}_{\text{stay}}) \cdot (\text{P}_{\text{stay}})^k = (1-(1-q^j)^n) \cdot ((1-q^j)^n)^k. \qquad \text{(C.10)}$$

This is a *geometric distribution*, and consequently, the expected number of T_{FB}+T_{ACK} used in state $I = j$, K_1, can be shown to be [13, (1.113)]:

$$\text{E}[K_1|j,n] = \frac{(1-q^j)^n}{1-(1-q^j)^n}. \qquad \text{(C.11)}$$

Summing this expression over all the states *before and including* state $I = i$, for a successful feedback transmission in state $I = i$, and using the law of total expectation, the expected number of T_{FB}+T_{ACK} before experiencing a successful feedback transmission, K_2, can be found as:

$$\begin{aligned} \text{E}[K_2] &= \sum_{l=0}^{L-1} \sum_{n=2}^{N} \sum_{i=1}^{\infty} \sum_{j=1}^{i} \text{E}[K_1|j,n] \cdot \pi(i,l,n) \cdot \text{P}_{\text{succ}} \sum_{k=0}^{\infty} (\text{P}_{\text{stay}})^k \\ &= \sum_{l=0}^{L-1} \sum_{n=2}^{N} \sum_{i=1}^{\infty} \sum_{j=1}^{i} \frac{(1-q^j)^n}{1-(1-q^j)^n} \cdot \pi(i,l,n) \cdot \frac{nq^i(1-q^i)^{n-1}}{1-(1-q^i)^n}. \qquad \text{(C.12)} \end{aligned}$$

Denoting the number of collisions by K_3, the expected number of collisions before successful feedback transmission can be found in a similar way:

$$\text{E}[K_3] = \sum_{l=0}^{L-1} \sum_{n=2}^{N} \sum_{i=1}^{\infty} i \cdot \pi(i,l,n) \cdot \frac{nq^i(1-q^i)^{n-1}}{1-(1-q^i)^n}. \qquad \text{(C.13)}$$

The expected guard time can now be found as:

$$\text{E}[T_{\text{G}}] = (T_{\text{FB}} + T_{\text{ACK}})(1 + \text{E}[M] + \text{E}[K_2] + \text{E}[K_3]), \qquad \text{(C.14)}$$

where the single T_{FB}+T_{ACK} denotes the time it takes for the user to transmit his FB packet successfully. As for the two ranked protocols, the first collision will be avoided when $L = 1$ (all users are within the successful interval). Therefore, one T_{FB}+T_{ACK} has to be deducted from the expression of the expected guard time in (C.14) for $L = 1$.

4.4 Guard Time for the Splitting Algorithm

To calculate the expected guard time for the Splitting algorithm and the Modified Splitting algorithm for different number of users, we have used [3, Eq. (13)] in combination with [3, Eq. (6)].

5 Analysis of the Maximum Average System Spectral Efficiency

In this section we derive expressions for the MASSE for all the feedback protocols, taking the degradation due to the guard time into account in each case. The expressions are first presented in a general form which holds for any channel fading distribution, and then closed-form expressions are presented for i.i.d. Rayleigh fading channels.

5.1 Spectral Efficiency When the User With Highest CNR is Selected

The MASSE of the Full Feedback protocol can be expressed as follows:

$$\begin{aligned}\text{MASSE}_{\text{FF}} &= \frac{T_{\text{TS}} - N \cdot T_{\text{FB}}}{T_{\text{TS}}} \int_0^\infty \log_2(1+\gamma) p_{\gamma^*}(\gamma) d\gamma \\ &= \frac{T_{\text{TS}} - N \cdot T_{\text{FB}}}{T_{\text{TS}}} \frac{N}{\ln 2} \sum_{n=0}^{N-1} \binom{N-1}{n} \frac{(-1)^n}{1+n} e^{(1+n)/\overline{\gamma}} E_1\left(\frac{1+n}{\overline{\gamma}}\right),\end{aligned} \tag{C.15}$$

where $p_{\gamma^*}(\gamma) = N \cdot P_\gamma^{N-1}(\gamma) \cdot p_\gamma(\gamma)$ is the probability density function (PDF) of the CNR of the user with the highest CNR, $p_\gamma(\gamma)$ being the PDF of the CNR of a single user. T_{TS} is the total time assigned for a transmission, with the guard time included.

Both the Ranked Full Feedback protocol and the Splitting algorithm will lead to a selection of the user with the highest CNR. When the user with

the highest CNR is always chosen to receive or transmit, the following expression for the MASSE is employed [4]:

$$\text{MASSE}_{\text{best}} = \sum_{l=0}^{L-1} \frac{T_{\text{TS}} - \text{E}_l[T_{\text{G}}]}{T_{\text{TS}}} \int_{\gamma_{\text{th},l}}^{\gamma_{\text{th},l+1}} \log_2(1+\gamma) p_{\gamma^*}(\gamma) d\gamma, \qquad \text{(C.16)}$$

where $\text{E}_l[T_{\text{G}}]$ is the expected guard time given that interval l is the successful interval. The relation between $\text{E}_l[T_{\text{G}}]$ and $\text{E}[T_{\text{G}}]$ found in the previous section can be expressed as follows:

$$\text{E}[T_{\text{G}}] = \sum_{l=0}^{L-1} \text{E}_l[T_{\text{G}}] p_N(l), \qquad \text{(C.17)}$$

where $p_N(l)$ is the PMF of l being the successful interval with N users in the system:

$$p_N(l) = P_\gamma^N(\gamma_{\text{th},l+1}) - P_\gamma^N(\gamma_{\text{th},l}). \qquad \text{(C.18)}$$

The corresponding expression for $\text{E}_l[T_{\text{G}}]$ for the Ranked Full Feedback protocol is given by:

$$\text{E}_l[T_{\text{G}}] = (L-l-1)\cdot(T_{\text{FB}} + T_{\text{ACK}}) + \frac{T_{\text{G,coll},l} + T_{\text{G,nocoll},l}}{p_N(l)}, \qquad \text{(C.19)}$$

where the expressions for $T_{\text{G,coll},l}$ and $T_{\text{G,nocoll},l}$ are given by (C.3) and (C.4), respectively. For $L = 1$ all the users will be in the successful interval and consequently we will have full feedback load, $\text{E}_l[T_{\text{G}}] = N \cdot T_{\text{FB}}$.

By using the derivation shown in Appendix 2, we obtain the MASSE for a Rayleigh fading channel given in (C.20), where $E_1(x) = \int_1^\infty e^{-xt}/t\,dt$ is the first order exponential integral function.

5.2 Spectral Efficiency When One Random User Within the Successful Interval is Selected

The Ranked Single-User Feedback protocol and the Exponential Backoff protocol will both choose a random user within the successful interval. Observing that picking a random user within the successful interval is similar to having quantized feedback, we can utilize the results from previous publications to develop an expression for the system spectral efficiency. Modifying [14, Eq. (17)] it can be shown that the spectral efficiency can be written as:

$$\text{MASSE}_{\text{single}} = \sum_{l=0}^{L-1} \frac{T_{\text{TS}} - \text{E}_l[T_{\text{G}}]}{T_{\text{TS}}} \frac{p_N(l)}{p_1(l)} \int_{\gamma_{\text{th},l}}^{\gamma_{\text{th},l+1}} \log_2(1+\gamma) p_\gamma(\gamma) d\gamma, \qquad \text{(C.21)}$$

$$\begin{aligned}\text{MASSE}_{\text{best}} &= \sum_{l=0}^{L-1} \frac{T_{\text{TS}} - \text{E}_l[T_{\text{G}}]}{T_{\text{TS}}} \frac{N}{\ln 2} \sum_{n=0}^{N-1} \binom{N-1}{n} \frac{(-1)^n}{1+n} \\ &\times \left[\ln(1+\gamma_{\text{th},l}) \cdot e^{-\frac{(1+n)\gamma_{\text{th},l}}{\overline{\gamma}}} - \ln(1+\gamma_{\text{th},l+1}) \cdot e^{-\frac{(1+n)\gamma_{\text{th},l+1}}{\overline{\gamma}}} \right] \\ &+ \sum_{l=0}^{L-1} \frac{T_{\text{TS}} - \text{E}_l[T_{\text{G}}]}{T_{\text{TS}}} \frac{N}{\ln 2} \sum_{n=0}^{N-1} \binom{N-1}{n} \frac{(-1)^n}{1+n} \\ &\times e^{\frac{(1+n)}{\overline{\gamma}}} \left(E_1\left(\frac{(1+n)(\gamma_{\text{th},l}+1)}{\overline{\gamma}} \right) - E_1\left(\frac{(1+n)(\gamma_{\text{th},l+1}+1)}{\overline{\gamma}} \right) \right)\end{aligned} \tag{C.20}$$

$$\begin{aligned}\text{MASSE}_{\text{single}} &= \frac{1}{\ln 2} \sum_{l=0}^{L-1} \frac{T_{\text{TS}} - \text{E}_l[T_{\text{G}}]}{T_{\text{TS}}} \frac{p_N(l)}{p_1(l)} \\ &\times \left[\ln(1+\gamma_{\text{th},l}) \cdot e^{-\frac{\gamma_{\text{th},l}}{\overline{\gamma}}} - \ln(1+\gamma_{\text{th},l+1}) \cdot e^{-\frac{\gamma_{\text{th},l+1}}{\overline{\gamma}}} \right] \\ &+ \frac{1}{\ln 2} \sum_{l=0}^{L-1} \frac{T_{\text{TS}} - \text{E}_l[T_{\text{G}}]}{T_{\text{TS}}} \frac{p_N(l)}{p_1(l)} \\ &\times \left[e^{\frac{1}{\overline{\gamma}}} \left(E_1\left(\frac{(\gamma_{\text{th},l}+1)}{\overline{\gamma}} \right) - E_1\left(\frac{(\gamma_{\text{th},l+1}+1)}{\overline{\gamma}} \right) \right) \right]\end{aligned} \tag{C.22}$$

where $\text{E}_l[T_{\text{G}}]$ is the guard time contribution from a trial with threshold $\gamma_{\text{th},l}$ and $p_1(l)$ is the probability that a random user is in the successful interval l. By using a similar derivation as in Appendix 2, we obtain the MASSE for a Rayleigh fading channel given in (C.22).

The two random single user feedback protocols have different values of $\text{E}_l[T_{\text{G}}]$ which make their MASSE different. For the Ranked Single-User Feedback protocol, $\text{E}_l[T_{\text{G}}]$ in (C.22) is the same as in (C.19), where the expressions for $T_{\text{G,coll},l}$ and $T_{\text{G,nocoll},l}$ are given by (C.7) and (C.4), respectively.

The expected guard time for the Exponential Backoff protocol, given

success in interval l, can be found by modifying (C.14) as follows:

$$\begin{aligned} \mathrm{E}_l[T_\mathrm{G}] &= (T_\mathrm{FB}+T_\mathrm{ACK})\cdot(L-l) \\ &+ (T_\mathrm{FB}+T_\mathrm{ACK})\cdot\sum_{n=2}^{N}\sum_{i=1}^{\infty}\sum_{j=1}^{i}\frac{(1-q^j)^n}{1-(1-q^j)^n}\cdot\frac{\pi(i,l,n)}{p_N(l)}\cdot\frac{\mathrm{P_{succ}}}{1-(1-q^i)^n} \\ &+ (T_\mathrm{FB}+T_\mathrm{ACK})\cdot\sum_{n=2}^{N}\sum_{i=1}^{\infty} i\cdot\frac{\pi(i,l,n)}{p_N(l)}\cdot\frac{\mathrm{P_{succ}}}{1-(1-q^i)^n}. \end{aligned} \tag{C.23}$$

When $L=1$ all the users are within the successful interval, and the user with the highest rank among the N users will be chosen for the Ranked Single-User Feedback protocol. In this case (C.22) reduces to:

$$\mathrm{MASSE_{RR}} = \frac{T_\mathrm{TS}-T_\mathrm{FB}}{T_\mathrm{TS}}\frac{1}{\ln 2}e^{1/\overline{\gamma}}E_1\left(\frac{1}{\overline{\gamma}}\right), \tag{C.24}$$

where the subscript RR denotes *Round Robin*. The ratio $\frac{T_\mathrm{TS}-T_\mathrm{FB}}{T_\mathrm{TS}}$ arises because the selected user feeds back his CSI estimate so that adaptive modulation and coding can be employed. For $L=1$ the Exponential Backoff protocol avoids the first collision and resolves the contention problem as usual.

6 Performance Evaluation of the Proposed Feedback Protocols: Discussion and Numerical Results

The main emphasis of this section is to evaluate the performance of the five described feedback protocols together with the the Full Feedback protocol and the Round Robin protocol based on the analysis in Section 4 and Section 5. The performance of the protocols will be evaluated by plotting the guard time and the MASSE for different number of thresholds (L) and users (N). Before presenting the numerical results we describe the IEEE 802.11 parameter values chosen for our numerical analysis.

6.1 IEEE 802.11 Parameter Values

To implement our protocols in an IEEE 802.11 network, we describe the following four packet types based on the general frame format defined in the standard [11].

Query (QRY) packet:

- 2 bytes FC (frame control)
- N times 6 bytes RA (receiver address)

- 1 byte Number of thresholds, L
- 4 bytes FCS (frame check sequence)

Feedback (FB) packet:

- 2 bytes FC
- 6 bytes TA (transmitter address)
- 1 byte CNR estimate
- 4 bytes FCS

Reservation (RES) packet:

- 2 bytes FC
- 4 bytes FCS

Acknowledgment (ACK) packet:

- 2 bytes FC
- 1 byte (0,1,e) ACK
- 4 bytes FCS

The FC field identifies the function and the fields of the packet, while the FCS field makes it possible for the receiver to separate packets from noise. In addition to these MAC-layer protocol fields, we also have to take the physical layer protocol fields into account. In IEEE 802.11 the physical layer protocol is called Physical Layer Convergence Protocol (PLCP) [15]. The packet headers of this protocol consists of a preamble and a header. If we assume that Direct Sequence Spread Spectrum (DSSS) is implemented at the physical layer the PLCP preamble consists of 18 bytes and the PLCP header consists of 5 bytes [15]. It should be observed that this implementation of DSSS does only combat interference and does not facilitate that multiple users can access the channel simultaneously.

To be able to calculate the duration of the packets listed above we have assumed that they are transmitted at the base rate 2 Mbps and that the propagation delay and packet processing time has the duration of a Short Interframe Space (SIFS). If we assume that a SIFS equals 10 μs (IEEE 802.11b) then T_{FB} equals 154 μs, T_{RES} equals 128 μs, and T_{ACK} equals 130 μs. For the Full Feedback protocol and the Round Robin protocol, no ACK packets are necessary, so the feedback from each user has the duration T_{FB}. As already mentioned in Section 4, we have also assumed that T_{QRY} has zero duration for all the algorithms.

6.2 Numerical Results for the Guard Time

In Figs. C.2 and C.3 we show plots of how the guard time varies with the number of thresholds for 4 and 12 users, respectively. For 4 users we see that the Ranked Single-User Feedback protocol gives the shortest guard

time, while the Modified Splitting algorithm gives the shortest guard time for 12 users. It should also be noted that the Full Feedback protocol gives a relatively short guard time for 4 users. However, since the guard time is proportional to the number of users for Full Feedback protocol, this protocol will perform the worst for a high number of users.

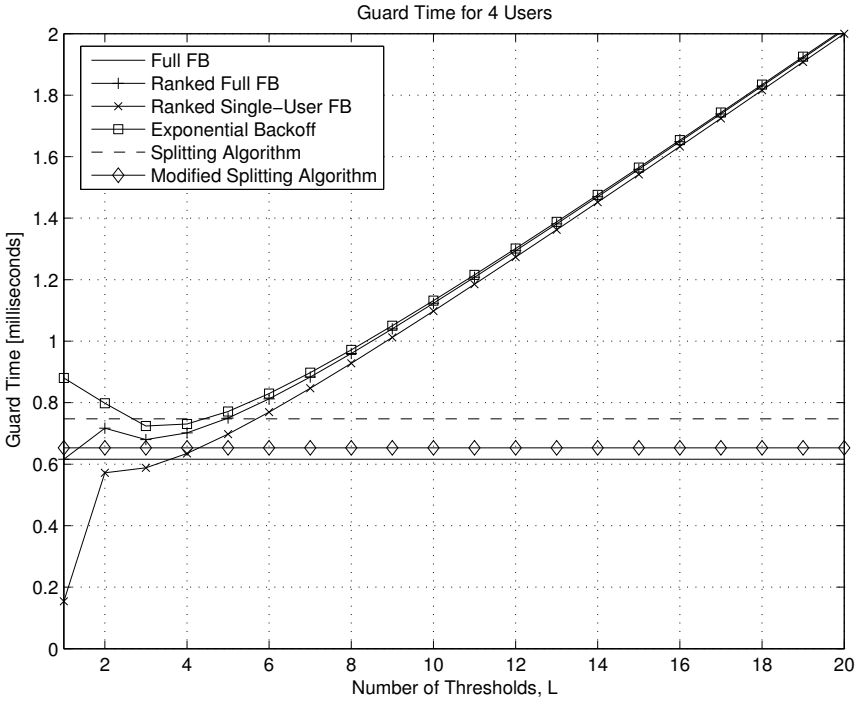

FIGURE C.2: Guard time for Rayleigh fading with $\overline{\gamma} = 15$ dB and 4 users.

6.3 Numerical Results for the MASSE

Figs. C.4 and C.5 show how the MASSE varies with the number of thresholds for short time-slots (T_{TS}=5 ms), for 4 and 12 users, respectively. The corresponding plots for long time-slots (T_{TS}=50 ms) are shown in Figs. C.6 and C.7.

For comparison purposes we have included graphs of the MASSE for *No Guard Time* and *Round Robin*. The former case corresponds to a theoretical system with no guard time and full MUD exploitation. The latter case

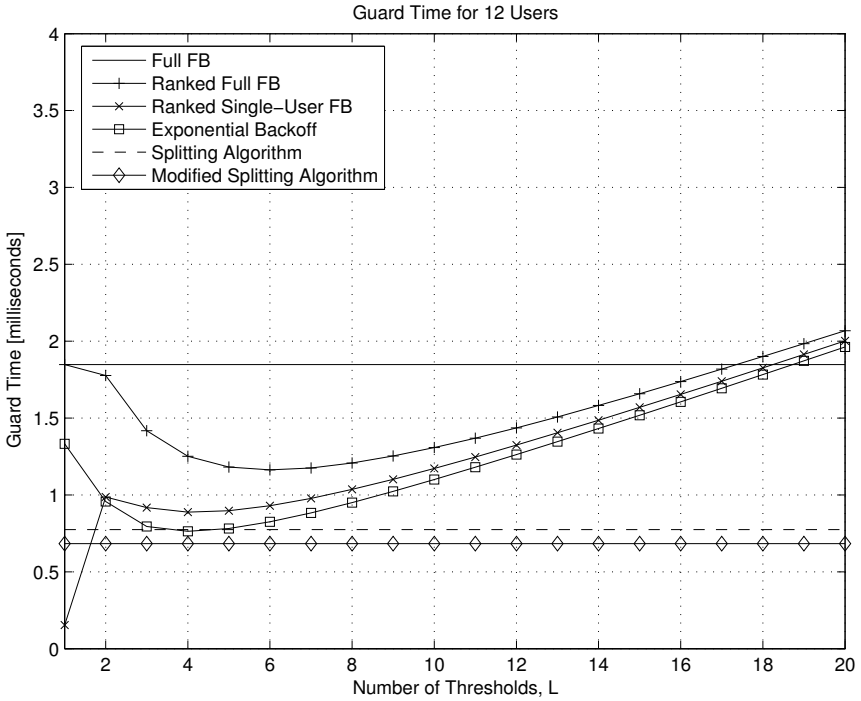

FIGURE C.3: Guard time for Rayleigh fading with $\overline{\gamma} = 15$ dB and 12 users.

corresponds to a system where adaptive coding and modulation are used, while opportunistic scheduling is not implemented. For this latter system, the users are scheduled in a Round Robin fashion. Feedback is still needed from the selected user in order to perform adaptive coding and modulation.

Although the Ranked Single-User Feedback protocol had the shortest guard time for 4 users, the Full Feedback protocol ensures that the MUD gain is maximized, and therefore the Full Feedback protocol yields the best MASSE performance for 4 users. For a higher number of users, the Modified Splitting algorithm shows the best MASSE performance since this protocol ensures full MUD exploitation and has a relatively short guard time.

For long time-slots, we see that the gain from the feedback reducing protocols diminishes. However, for many users the Modified Splitting protocol still gives a small gain over the other feedback protocols.

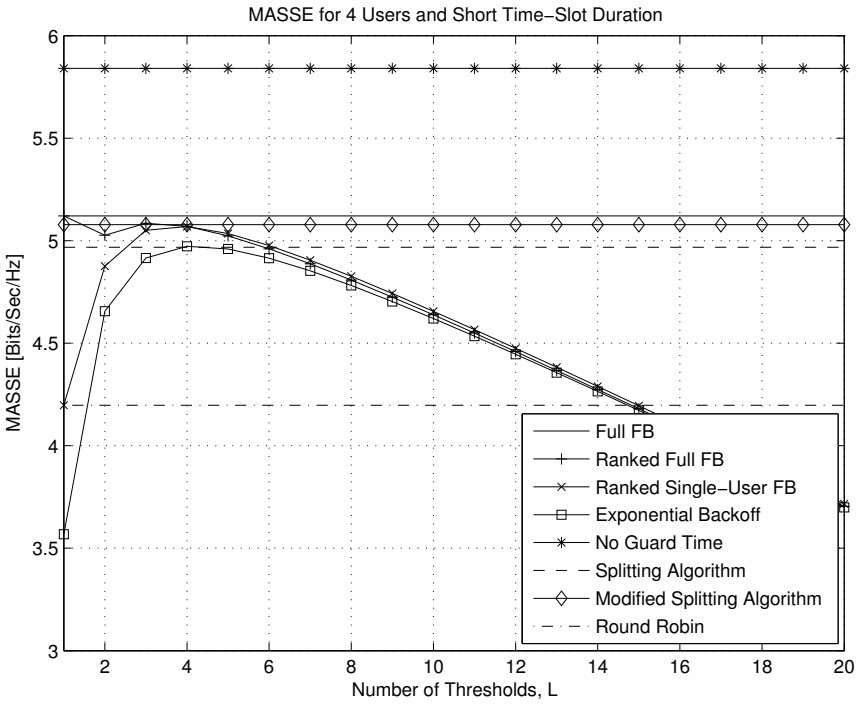

FIGURE C.4: MASSE for Rayleigh fading with $\overline{\gamma} = 15$ dB and 4 users. T_{TS} =5 ms.

7 Conclusions

In this paper we studied feedback protocols for possible use in slotted cellular ALOHA-based networks exploiting MUD. We considered downlink transmission where the base station transmits data to the mobile users. To be able to exploit MUD, the base station wants to schedule the user with the best channel quality for each time-slot. Therefore, the base station needs to collect feedback from the mobile users. In conventional networks that exploit MUD, feedback is collected from all users, which can be a time-consuming process. Consequently, we analyzed feedback protocols aimed at reducing the number of users transmitting feedback, and hence the guard time used to collect feedback.

We proposed three new feedback protocols for ALOHA-based cellular networks, namely, (i) Ranked Full Feedback, (ii) Ranked Single-User Feedback, and (iii) Exponential Backoff. Closed-form expressions were also

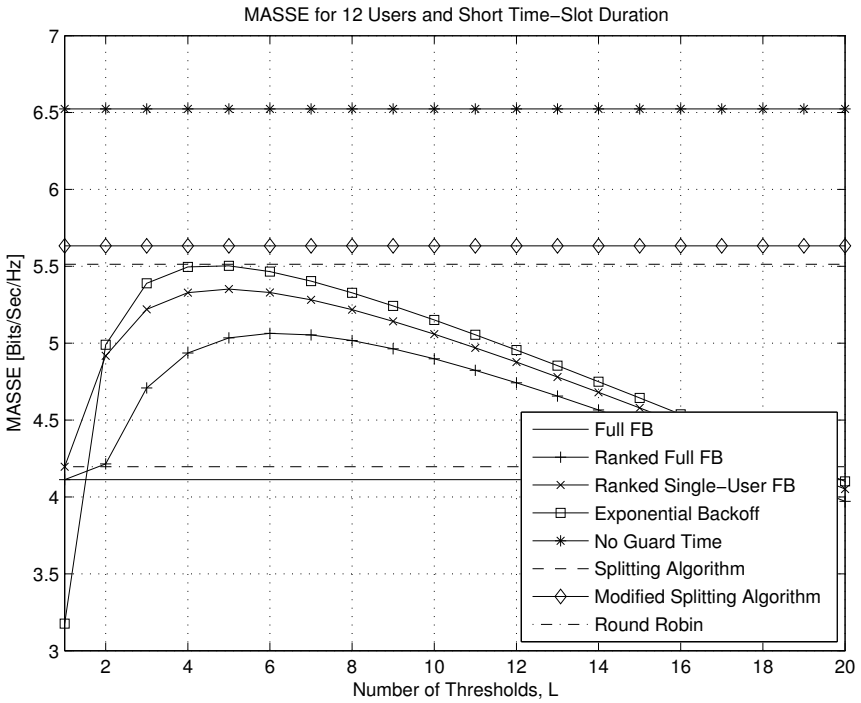

FIGURE C.5: MASSE for Rayleigh fading with $\overline{\gamma} = 15$ dB and 12 users. $T_{\text{TS}} = 5$ ms.

found for the guard time duration and the MASSE for these three protocols. We also investigated the guard time and MASSE performance in an IEEE 802.11-based cellular network for the three new protocols and compared their performance with the Splitting algorithm proposed in [3] and a new and modified version of this algorithm. Our plots showed that the five different feedback protocols all give a feedback reduction for a system with many mobile users, and that the Modified Splitting algorithm showed the best MASSE performance. However, for a low (4) number of users the Full Feedback algorithm surprisingly showed the best MASSE performance.

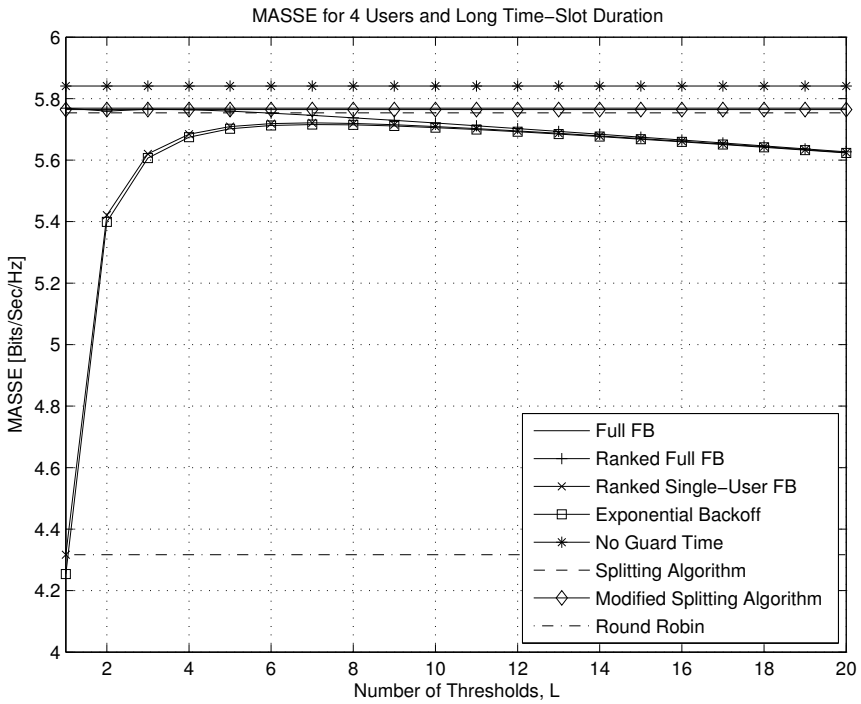

FIGURE C.6: MASSE for Rayleigh fading with $\overline{\gamma} = 15$ dB and 4 users. T_{TS} =50 ms.

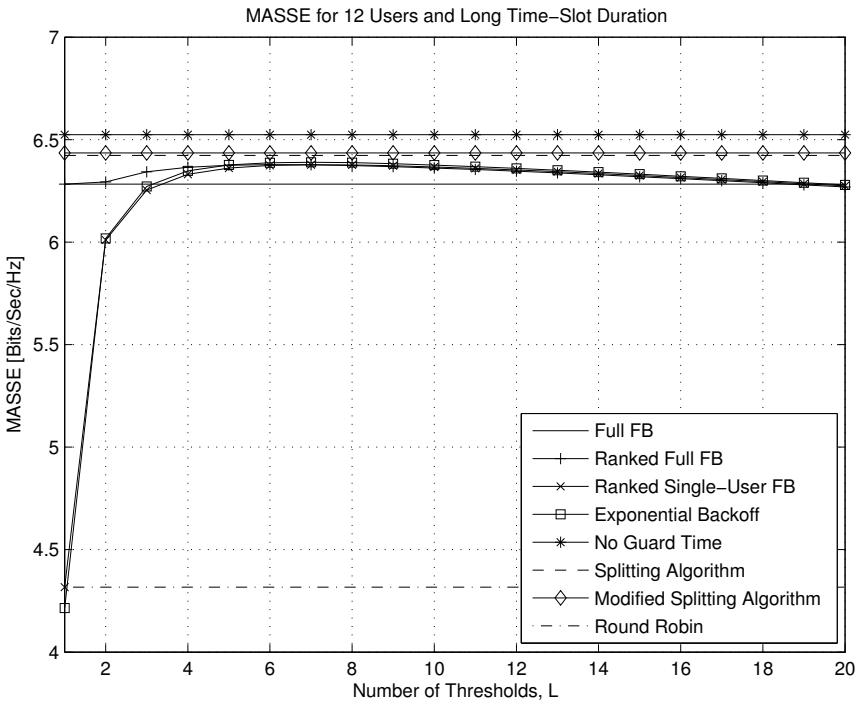

FIGURE C.7: MASSE for Rayleigh fading with $\overline{\gamma} = 15$ dB and 12 users. $T_{\text{TS}} = 50$ ms.

References

[1] P. Viswanath, D. N. C. Tse, and R. Laroia, "Opportunistic beamforming using dumb antennas," *IEEE Trans. Inform. Theory*, vol. 48, pp. 1277–1294, June 2002.

[2] V. Hassel, M.-S. Alouini, D. Gesbert, and G. E. Øien, "Exploiting multiuser diversity using multiple feedback thresholds." In *Proc. IEEE Vehicular Technology Conference (VTC'05-spring)*, Stockholm, Sweden, May 2005.

[3] X. Qin. and R. Berry, "Opportunistic splitting algorithms for wireless networks," in *Proc. IEEE International Conference on Computer Communications (INFOCOM'04)*, (Hong Kong, PR China), pp. 1662 – 1672, Mar. 2004.

[4] R. Knopp and P. A. Humblet, "Information capacity and power control in single cell multiuser communications," in *Proc. IEEE Int. Conference on Communications (ICC'95)*, (Seattle, WA, USA), pp. 331–335, June 1995.

[5] M. Andrews, K. Kumaran, K. Ramanan, A. Stolyar, P. Whiting, and R. Vijayakumar, "Providing quality of service over a shared wireless link," *IEEE Communications Mag.*, vol. 39, pp. 150–154, Feb. 2001.

[6] P. Bender, P. Black, M. Grob, R. Padovani, N. Sindhushyana, and S. Viterbi, "CDMA/HDR: A bandwidth-efficient high-speed wireless data service for nomadic users," *IEEE Communications Mag.*, vol. 38, pp. 70–77, July 2000.

[7] D. Gesbert and M.-S. Alouini, "How much feedback is multi-user diversity really worth?," in *Proc. IEEE Int. Conference on Communications (ICC'04)*, (Paris, France), pp. 234–238, June 2004.

[8] H. Koubaa, V. Hassel, and G. E. Øien, "Multiuser diversity gain enhancement by guard time reduction." In *Proc. IEEE Workshop on Signal Processing Advances in Wireless Communications (SPAWC'05)*, New York, NY, USA, June 2005.

[9] H. Koubaa, V. Hassel, and G. E. Øien, "Contention-less feedback for multiuser diversity scheduling." In *Proc. IEEE Vehicular Technology Conference (VTC'05-fall)*, Dallas, TX, USA, Sept. 2005.

[10] K. J. Hole and G. E. Øien, "Spectral efficiency of adaptive coded modulation in urban microcellular networks," *IEEE Trans. on Veh. Technol.*, vol. 50, pp. 205–222, Jan. 2001.

[11] IEEE Standards Department, "IEEE Std 802.11. Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) specifications," tech. rep., IEEE, September 1999.

[12] D. G. Jeong and W. S. Jeon, "Performance of an exponential backoff scheme for slotted-ALOHA protocol in local wireless environment," *IEEE Trans. on Veh. Technol.*, vol. 44, pp. 470–479, Aug. 1995.

[13] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products*. San Diego, CA, USA: Academic Press, 6th ed., 2000.

[14] F. Florén, O. Edfors, and B.-A. Molin, "The effect of feedback quantization on the throughput of a multiuser diversity scheme," in *Proc. IEEE Global Communications Conference (GLOBECOM'03)*, vol. 1, (San Francisco, CA, USA), pp. 497–501, Dec. 2003.

[15] M. Ergen, "IEEE 802.11 tutorial." University of California Berkeley, `http://www.eecs.berkeley.edu/~ergen/docs/ieee.pdf`, June 2002.

Paper D

Spectral Efficiency and Fairness for Opportunistic Scheduling Algorithms

Vegard Hassel, Geir E. Øien, and Marius Røed Hanssen

Submitted to
IEEE Transactions on Wireless Communications

Abstract

In this paper we analyze spectral efficiency and fairness for four different scheduling algorithms operating in a wireless cellular network where the users' channels have different average carrier-to-noise ratios (CNRs). The four scheduling algorithms investigated are: Round Robin Scheduling, Maximum CNR Scheduling, Normalized CNR Scheduling, and Opportunistic Round Robin Scheduling [1, 2]. We develop closed-form expressions for the system spectral efficiency and for two types of fairness measures for all four algorithms. The differences in spectral efficiency and fairness of the four scheduling algorithms are highlighted by analyzing plots of our closed-form expressions and by analytically investigating the asymptotic fairness behavior.

1 Introduction

For efficient utilization of the scarce radio spectrum available for wireless communication, *opportunistic multiuser scheduling* has recently attracted much attention. Opportunistic scheduling increases the system spectral efficiency by selecting users with favorable channel conditions to transmit or receive data [3, 4]. The scheduling algorithm that maximizes the system spectral efficiency among all time division multiplexing (TDM) based algorithms is the one where the user with the highest carrier-to-noise ratio (CNR) is served at all times [3]. We thus say that this algorithm maximizes the multiuser diversity (MUD) gain [4]. Here, we refer to this algorithm as *Max CNR scheduling* (MCS). However, always selecting the very best user can lead to starvation of the users having the lowest average CNRs. It is therefore necessary to develop scheduling algorithms that take both the channel conditions and the quality-of-service (QoS) demands of the users into account. As a first approach to fulfilling QoS, we can analyze the *fairness* of different scheduling algorithms [5].

The *Opportunistic Round Robin* (ORR) scheduling policy [1, 2] obtains higher fairness by combining the *Round Robin* (RR) scheduling policy and the MCS policy. This algorithm will lead to a fairness in the resource allocation that is close to that of the RR policy, and will at the same time be able to exploit some of the MUD in the system. To have a more fair resource allocation for the scenario where the users have different average CNRs, we can choose to exploit the *relative MUD* that exists between the users. The relative MUD is defined as the MUD which is due to instantaneous channel fluctuations which are independent of the average channel statistics. One algorithm that takes advantage of the relative MUD is the *Normalized CNR Scheduling* (NCS) algorithm [6].

Contributions. In this paper we develop a new closed-form system spectral efficiency expression for a version of the ORR algorithm. In addition, we define new expressions for *time-slot fairness* and *throughput fairness*, and look at the asymptotic behavior of these expressions for a general scheduling algorithm. We also develop closed-form expressions for these two fairness measures for the RR, MCS, NCS, and ORR algorithms when the users' channels are independently and non-identically distributed (i.n.d.) with different average CNRs from user to user. Our analytical expressions make it possible to analyze the mathematical relationships between different system parameters and to obtain numerical results without having to do simulations. Our work also provides an analytical framework for obtaining similar closed-form expressions for other

scheduling algorithms[1].

Organization. The rest of this paper is organized as follows. In Section 2 we present the system model and in Section 3 we give a short description of the four scheduling algorithms in question. Closed-form expressions for system spectral efficiency and fairness are presented in Section 4 and Section 5, respectively. In Section 6, we present our numerical results. Section 7 lists our conclusions.

2 System Model

We consider a single base station that serves the up-links and down-links of N users using TDM. Before performing scheduling, the base station is assumed to receive perfect information about the users' CNRs. For each time-slot the scheduling decision is sent from the base station to the relevant user. It is assumed that the channels of the users are i.n.d. slowly-varying, flat Rayleigh fading channels with average received CNRs $\overline{\gamma}_i$, where $i \in \{1, ..., N\}$ is the user index. The time-slot duration is assumed to be less than one coherence time, i.e., the channels can be regarded more or less as constant during one time-slot. We also assume that the CNR values from time-slot to time-slot are uncorrelated. This means that one user will seldom experience two adjacent time-slots with the same CNR values, and consequently, the opportunistic distribution of time-slots between the users appear to be more fair compared to the corresponding resource allocation in a real-life wireless network. Consequently, the fairness expressions found in this paper gives optimistic bounds on the fairness that can be expected for more realistic channel models. However, we expect the relative ranking of the four algorithms' properties with respect to fairness to be maintained also when we deviate from our simplified channel model.

When the users' channels are independently and identically distributed (i.i.d.) with an average CNR of $\overline{\gamma}$, the expressions for spectral efficiency and fairness can be obtained by simply replacing $\overline{\gamma}_i$ with $\overline{\gamma}$ in the closed-form expressions obtained for i.n.d. channels.

3 Description of the Four Scheduling Algorithms

3.1 The Round Robin (RR) Algorithm

To have a reference algorithm, we start by investigating RR, where the users are assigned time-slots in a sequential fashion. This is a non-opportunistic

[1]Parts of this paper are based on work in [7].

scheduling policy, implemented e.g. in GSM. Since the base station do not need any feedback information from the users to conduct the scheduling this is the most simple algorithm to implement.

3.2 The Max CNR Scheduling (MCS) Algorithm

Another interesting scheduling algorithm is the MCS algorithm where the user with the best channel conditions is selected in any time-slot. This is the most opportunistic of the time-slotted scheduling algorithms.

3.3 The Normalized CNR Scheduling (NCS) Algorithm

Using the MCS algorithm when the users' channels are i.n.d. will in many cases not take advantage of the *relative MUD* that exists between the users. However, by scheduling the users that have the highest ratio $\chi_i(t) = \frac{\gamma_i(t)}{\overline{\gamma}_i}$, the relative MUD will be exploited. This ratio expresses the instantaneous CNR level of user i divided by his average CNR. When always scheduling the users with the highest $\chi_i(t)$, and with many users in the cell, the users will be selected when they are close to their CNR peak values. This scheduling algorithm has a similar performance as the Proportional Fair Scheduling (PFS) algorithm when the time-window t_c in the PFS algorithm is long [4].

3.4 The Opportunistic Round Robin (ORR) Algorithm

For the ORR algorithm, the time-slots are allocated in successive rounds of N competitions, where N is the number of users [1, 2]. All the N users will be assigned one time-slot within a round. However, as opposed to RR, the time-slots are assigned *opportunistically*. For the first competition within a round, the best user out of all the N users are chosen. However, for the next competition this user is not participating and the best out of the remaining $N-1$ users is selected. For each new competition the winner from the last competition is taken out. Consequently, only one user participates in the last competition of a round. The advantage of this algorithm is that it opportunistically takes advantage of the channel conditions of the users and at the same time ensures that the allocated time-slots are evenly distributed among the users after every complete round. For i.n.d. channels, the ORR algorithm will not give the same degree of opportunistic assignment of time-slots. When the users' average CNRs are spread far apart, the ORR algorithm will approach the same spectral efficiency as for ordinary RR scheduling. However, by instead scheduling the user with the highest

normalized CNR $\chi_i(t) = \frac{\gamma_i(t)}{\overline{\gamma}_i}$ in each competition, the relative MUD will be exploited. This algorithm has already been proposed in [8] and we denote the algorithm as Normalized ORR (N-ORR).

4 Spectral Efficiency Analysis

When designing optimal scheduling algorithms, the *maximum average system spectral efficiency* (MASSE) theoretically attainable is a natural and important performance measure. The MASSE [bits/s/Hz] is defined as the maximum average sum of spectral efficiencies within a cell, shared between all users' up-links and down-links. The expression for the MASSE for constant-power, optimal rate adaptation is given as [9]:

$$\text{MASSE} = \sum_{i=1}^{N} p_i \int_0^{\infty} \log_2(1+\gamma) p_{\gamma_i^*}(\gamma)\, d\gamma, \tag{D.1}$$

where p_i is the probability of user i being selected in any time-slot (access probability), and $p_{\gamma_i^*}(\gamma)$ is the probability density function (PDF) of the CNR for the scheduling policy under study when the user with average CNR $\overline{\gamma}_i$ is selected. In this section we will investigate the MASSE of the four scheduling algorithms described in the previous section.

4.1 MASSE for the RR Algorithm

The MASSE of the RR scheduling algorithm simply equals the spectral efficiency averaged over all the users [10, Eq. (34)]:

$$\text{MASSE} = \frac{1}{N \ln 2} \sum_{i=1}^{N} e^{1/\overline{\gamma}_i} E_1\left(\frac{1}{\overline{\gamma}_i}\right), \tag{D.2}$$

where $E_1(x) = \int_1^{\infty} e^{-xt}/t\, dt$ is the exponential integral function.

4.2 MASSE for the MCS Algorithm

The MASSE of the MCS algorithm can be expressed as [6, Eq. (14)]:

$$\text{MASSE} = \frac{1}{\ln 2} \sum_{i=1}^{N} \frac{1}{\overline{\gamma}_i} \sum_{\tau \in T_i^N} \text{sign}(\tau) \frac{e^{\left(\frac{1}{\overline{\gamma}_i}+|\tau|\right)}}{\frac{1}{\overline{\gamma}_i}+|\tau|} E_1\left(\frac{1}{\overline{\gamma}_i}+|\tau|\right), \tag{D.3}$$

where T_i^N denotes a set containing the terms arising from a certain type of expansion of the product $\prod_{\substack{j=1 \\ j\neq i}}^{N}(1-e^{-\gamma/\overline{\gamma}_j})$ [11, Sec. III-D-2].

4.3 MASSE for the NCS Algorithm

The MASSE of the NCS algorithm can be expressed as [6, Eq. (28)]:

$$\text{MASSE} = \frac{1}{\ln 2} \sum_{i=1}^{N} \sum_{j=0}^{N-1} \binom{N-1}{j} \frac{(-1)^j}{1+j} e^{\frac{1+j}{\overline{\gamma}_i}} E_1\left(\frac{1+j}{\overline{\gamma}_i}\right). \tag{D.4}$$

4.4 MASSE for the N-ORR Algorithm

When combining the ORR and NCS algorithms in the way described in Section 3.4, the users will get selected in a competition if they have the highest *normalized* CNR. It can easily be shown that the normalized CNR is independently and identically distributed with unity average value for all the users; thus, the relative multiuser gain is conserved for this algorithm, and each user will experience a multiuser gain as if all the users were i.i.d. with the *same* average CNR as this user. Consequently, user i experiences the following *individual* cumulative distribution function (CDF) if he is scheduled when n ($\leq N$) users are competing:

$$P_{\gamma_i^*}(\gamma) = P_{\gamma_i}^n(\gamma), \tag{D.5}$$

where $P_{\gamma_i}(\gamma)$ is the CDF of the CNR for a single user with average CNR $\overline{\gamma}_i$. Since the a priori probability that a user is selected in an arbitrary competition is $\frac{1}{N}$, the *total* CDF $P_{\gamma^*}(\gamma)$ of the N-ORR algorithm can be expressed as the average over all users, and all competitions within a round of competitions. Differentiating $P_{\gamma^*}(\gamma)$ with respect to γ, we obtain the following PDF for the N-ORR algorithm:

$$p_{\gamma^*}(\gamma) = \frac{1}{N^2} \sum_{n=1}^{N} n \sum_{i=1}^{N} P_{\gamma_i}^{n-1}(\gamma) p_{\gamma_i}(\gamma), \tag{D.6}$$

where $p_{\gamma_i}(\gamma)$ is the PDF of the CNR for a single user with average CNR $\overline{\gamma}_i$. Inserting this expression into (D.1) and using a similar derivation as for [10, Eq. (44)] we obtain the following expression for the MASSE:

$$\text{MASSE} = \frac{1}{N^2 \ln 2} \sum_{n=1}^{N} n \sum_{i=1}^{N} \sum_{j=0}^{n-1} \binom{n-1}{j} (-1)^j \frac{e^{\frac{1+j}{\overline{\gamma}_i}}}{1+j} E_1\left(\frac{1+j}{\overline{\gamma}_i}\right). \tag{D.7}$$

5 Fairness Analysis

Several measures of fairness have been introduced in the preceding literature. In [12], *Jain's Fairness Index* (JFI), which has been used in several recent

papers ([13], [14]), was introduced. JFI measures the fairness experienced by a user on the average. The properties of the JFI make it well suited as a measure for fairness. The index has all the following desired properties of a fairness measure [12]:

- *population size independent*, i.e., JFI is applicable to any number of users.
- *scale independent*, i.e., JFI is only dependent on the resource allocation relative to the expected resource allocation to an arbitrary user given a scheduling algorithm and a number of system parameters.
- *bounded*, i.e., JFI is bounded between zero and one, where zero means total unfairness and one means total fairness. It should be noted that total unfairness will only occur when we have an infinite number of users and one user gets all the resources.
- *continuous*, i.e., continuous changes in the resource allocation influence JFI continuously.

We choose to use the following version of JFI [12, Eq. (2)]:

$$\mathrm{F}(K) = \frac{(\mathrm{E}_K[X])^2}{\mathrm{E}_K[X^2]}, \tag{D.8}$$

where X is a random variable describing the amount of resource allocated to a user, and $\mathrm{E}_K[\cdot]$ is the expectation calculated over the distribution of the resource allocation within a time-window of K time-slots. Since we are interested in investigating the variation in the resource allocation between the users, $\mathrm{E}_K[\cdot]$ is an average over all *users* in the system. As opposed to the traditional definition of JFI which is based on calculating the fairness resulting from an actual allocation, the definition in (D.8) makes it possible to calculate the fairness of a scheduling algorithm based on the statistics of the algorithm and the wireless channel, and thus to judge the fairness effects of various algorithms before the performance of the algorithms is evaluated through simulations or practical experiments.

5.1 Definitions and Asymptotic Analysis of Time-Slot Fairness and Throughput Fairness

5.1.1 Time-Slot Fairness

Scheduling algorithms for TDM-based wireless cells allocate time-slots to the different users. Choosing X in (D.8) to be the number of time-slots allocated to the users, we can define the *time-slot fairness* as:

$$\mathrm{F}_{\mathrm{TS}}(K) = \frac{(\mathrm{E}_K[M])^2}{\mathrm{E}_K[M^2]}, \tag{D.9}$$

where M is a random variable that describes the number of time-slots allocated to an arbitrary user within a window of K time-slots. This random variable is discrete and can take on the values $k = 1, ..., K$. For any scheduling algorithm, the expected number of time-slots allocated to an arbitrary user within a time-window of K time-slots is $\mathrm{E}_K[M] = K/N$. It should also be noted that M will converge to Kp_i, when K grows to infinity. Consequently, we can express the second moment of M as K grows to infinity as:

$$\lim_{K\to\infty} \mathrm{E}[M^2] = \frac{1}{N}\sum_{i=1}^{N}(Kp_i)^2 = \frac{1}{N}\left(\frac{K}{N}\right)^2\sum_{i=1}^{N}(Np_i)^2\,. \tag{D.10}$$

We can now express the asymptotic time-slot fairness when K goes to infinity as:

$$\lim_{K\to\infty} \mathrm{F_{TS}}(K) = \frac{\left(\frac{K}{N}\right)^2}{\frac{1}{N}\left(\frac{K}{N}\right)^2\sum_{i=1}^{N}(Np_i)^2} = \frac{1}{\frac{1}{N}\sum_{i=1}^{N}(Np_i)^2}. \tag{D.11}$$

We see that the time-slot fairness will converge to unity only for the scheduling algorithms that have $p_i = \frac{1}{N}$, for all $i = 1, ..., N$. This means that the time-slot fairness for all the scheduling algorithms investigated in this paper will converge to unity, as K grows large, except for the MCS algorithm when the users' channels are i.n.d..

5.1.2 Throughput Fairness

Although the time-slot fairness might be good for a given scheduler, the users do not necessarily therefore experience the same fair allocation of throughput. This might be because the users have different average CNRs and will thus experience different throughput in their assigned time-slots. Consequently, we chose to investigate the *throughput fairness*, which can be defined as:

$$\mathrm{F_{TP}}(K) = \frac{(\mathrm{E}_K[R])^2}{\mathrm{E}_K[R^2]} = \frac{(\mathrm{E}_K[R])^2}{(\mathrm{E}_K[R])^2 + \mathrm{var}(R)}, \tag{D.12}$$

where $X = R$ [bits per time-window per Hz] is a random variable describing the throughput allocated to an arbitrary user within K time-slots. It should be noted that because of the scale independency of the JFI, this definition of the throughput fairness measures fairness relative to the expected throughput that is allocated to an arbitrary user for a given scheduling algorithm and some given system parameters. Intuitively, this is logical since fairness always has to measure the resource allocation of one user relative to the allocation to the other users. When the users' channels are i.n.d.,

the variance in the resource allocation relative to the expected allocation, will be higher for the throughput allocation compared to the corresponding variance of the time-slot allocation. The JFI will therefore show higher time-slot fairness than throughput fairness in this case. This is also logical, since (i) users with good channel quality will observe a high degree of unfairness in the throughput allocation since they can transmit more bits compared to the average user and (ii) users with bad channel quality will observe a high degree of unfairness in the throughput allocation since they can transmit fewer bits than the average user. The users will not observe the same degree of unfairness in the time-slot allocation. It can therefore be advantageous to observe the throughput fairness instead of the time-slot fairness for many systems. It should however be noted that it is important to compare the results for throughput fairness with the corresponding MASSE for a given scheduling algorithm and some given system parameters. Since the throughput fairness is measured relative to the expected throughput in the system, we have to choose scheduling algorithms that give both a relatively high degree of fairness and a relatively high MASSE.

By using the results from the asymptotic analysis of the time-slot allocation, and by assuming that the throughput allocated in an arbitrary time-slot is independent of M for large values of K, we obtain:

$$\lim_{K\to\infty} \mathrm{E}_K[R] = \frac{1}{N}\sum_{i=1}^{N}\int_0^\infty Kp_i \log_2(1+\gamma) p_{\gamma_i *}(\gamma) d\gamma = \frac{K}{N}\cdot \mathrm{MASSE}. \quad \text{(D.13)}$$

The second moment of the throughput allocation, as K goes to infinity, can be expressed as:

$$\begin{aligned}\lim_{K\to\infty} \mathrm{E}_K[R^2] &= \frac{1}{N}\sum_{i=1}^{N}\int_0^\infty \left(Kp_i \log_2(1+\gamma)\right)^2 p_{\gamma_i *}(\gamma) d\gamma \\ &= N\left(\frac{K}{N}\right)^2 \sum_{i=1}^{N} p_i^2 \int_0^\infty \left(\log_2(1+\gamma)\right)^2 p_{\gamma_i *}(\gamma) d\gamma, \quad \text{(D.14)}\end{aligned}$$

The corresponding expression for the asymptotic throughput fairness can now be expressed as:

$$\lim_{K\to\infty} \mathrm{F_{TP}}(K) = \frac{(\mathrm{MASSE})^2}{N\sum_{i=1}^{N} p_i^2 \int_0^\infty \left(\log_2(1+\gamma)\right)^2 p_{\gamma_i *}(\gamma) d\gamma}. \quad \text{(D.15)}$$

It is not obvious from this expression what values of p_i and distributions of $p_{\gamma_i *}(\gamma)$ will give us total throughput fairness. However, by analyzing the

equivalent expression

$$\lim_{K\to\infty} \mathrm{F}_{\mathrm{TP}}(K) = \lim_{K\to\infty} \frac{(\mathrm{E}_K[R])^2}{(\mathrm{E}_K[R])^2+\mathrm{var}(R)}$$
$$= \frac{\left(\frac{K}{N}\mathrm{MASSE}\right)^2}{\left(\frac{K}{N}\mathrm{MASSE}\right)^2+\frac{1}{N}\sum_{i=1}^{N}\int_0^\infty \left(Kp_i\log_2(1+\gamma)-\frac{K}{N}\mathrm{MASSE}\right)^2 p_{\gamma_i*}(\gamma)d\gamma}, \tag{D.16}$$

we see that the throughput fairness converges to unity only if $p_i = \frac{1}{N}$ and $p_{\gamma_i*}(\gamma)$ equals the *dirac delta function* at $\gamma = \overline{\gamma}$ for all values of i, i.e., the CNR is constant and identical for all the users. Intuitively, this is logical since a constant and identical CNR will lead to zero variance in the rate of the users. Since we assume non-constant (random) CNRs in this paper, none of the investigated scheduling algorithms will lead to full throughput fairness when K goes to infinity.

5.2 Time-Slot Fairness for Different Scheduling Algorithms

5.2.1 Time-Slot Fairness for RR

For RR, the time-slots are assigned non-opportunistically to the users in rounds of N time-slots, where each user is assigned one time-slot in each round. If we assume that the order of the users within each round is random, and hence, the number of allocated time-slots is independent of i, we obtain the following probability mass function (PMF) for M:

$$p_M(k) = \begin{cases} \frac{(k+1)N-K}{N}, & k = \lfloor \frac{K}{N} \rfloor \\ \frac{K-(k-1)N}{N}, & k = \lceil \frac{K}{N} \rceil \\ 0, & \text{otherwise} \end{cases}. \tag{D.17}$$

From this PMF it can be shown that the average number of time-slots assigned to any user is $\mathrm{E}_K[M] = K/N$. Similarly, it can be shown that the second moment of M can be expressed as

$$\mathrm{E}_K[M^2] = \left(\left\lfloor \frac{K}{N} \right\rfloor\right)^2 \left(\frac{(\lfloor \frac{K}{N} \rfloor + 1)N - K}{N}\right), + \left(\left\lceil \frac{K}{N} \right\rceil\right)^2 \left(\frac{K - (\lceil \frac{K}{N} \rceil - 1)N}{N}\right) \tag{D.18}$$

for $K \bmod N \neq 0$, and $\mathrm{E}_K[M^2] = (K/N)^2$ otherwise. Inserting the expressions for $\mathrm{E}_K[M^2]$ and $\mathrm{E}_K[M] = K/N$ into (D.9), we obtain a closed-form expression for the time-slot fairness of the RR algorithm.

5.2.2 Time-Slot Fairness for MCS

For MCS, where the user with the highest CNR is chosen in each time-slot, the distribution of the number of time-slots allocated to user i within K time-slots is dependent on the probabilities of selecting user i in an arbitrary time-slot [6, Eq. (12)]:

$$p_i = \int_0^\infty p_{\gamma_i}(\gamma) \prod_{\substack{j=1 \\ j\neq i}}^{N} P_{\gamma_j}(\gamma) d\gamma = \frac{1}{\overline{\gamma}_i} \sum_{\tau \in T_i^N} \text{sign}(\tau) \frac{1}{\frac{1}{\overline{\gamma}_i} + |\tau|}, \tag{D.19}$$

where $P_{\gamma_j}(\gamma)$ is the CDF of the CNR of a single user with average CNR $\overline{\gamma}_j$.

To calculate the first and second moments of the time-slot distribution between the users, we have to find the PMF of how the time-slots are distributed between the users. This PMF can be expressed by the *multinomial distribution* [15]:

$$p_M(\mathbf{k}) = \begin{cases} \frac{K!}{k_1! \cdot k_2! \cdots k_N!} \prod_{i=1}^{N} p_i^{k_i}, & \text{if } \sum_{i=1}^{N} k_i = K \\ 0, & \text{otherwise,} \end{cases} \tag{D.20}$$

where the vector $\mathbf{k} = [k_1, ..., k_i, ..., k_N]$ denotes the different values of M for each of the N different users.

We can now obtain the expected number of time-slots scheduled to an arbitrary user as:

$$\mathrm{E}_K[M] = \sum_{\mathbf{k}:\sum_{i=1}^{N} k_i = K} \left(\frac{1}{N} \sum_{i=1}^{N} k_i \right) \frac{K!}{k_1! \cdot k_2! \cdots k_N!} \prod_{i=1}^{N} p_i^{k_i} = \frac{K}{N}. \tag{D.21}$$

The corresponding second moment of the number of time-slots allocated to a user can be expressed as:

$$\mathrm{E}_K[M^2] = \sum_{\mathbf{k}:\sum_{i=1}^{N} k_i = K} \left(\frac{1}{N} \sum_{i=1}^{N} k_i^2 \right) \frac{K!}{k_1! \cdot k_2! \cdots k_N!} \prod_{i=1}^{N} p_i^{k_i}. \tag{D.22}$$

Inserting the expressions for these two first moments into (D.9), we subsequently obtain a closed-form expression for the time-slot fairness of the MCS algorithm.

5.2.3 Time-Slot Fairness for NCS

It should be observed that when the relatively best user is chosen in the i.n.d. scenario, the users will have the same probability of being scheduled in an arbitrary time-slot. This means that the PMF of the number of

time-slots allocated to an arbitrary user can be expressed by the *binomial distribution* [16, p. 1179]:

$$p_M(k) = \binom{K}{k} p_i^k (1-p_i)^{K-k}, \tag{D.23}$$

where $p_i = \frac{1}{N}$.

We can now find the expected number of time-slots allocated to any user as:

$$\mathrm{E}_K[M] = \sum_{k=1}^{K} k\, p_M(k) = \frac{K}{N}. \tag{D.24}$$

Similarly, we can find the second moment of the time-slot allocation to be:

$$\mathrm{E}_K[M^2] = \sum_{k=1}^{K} k^2 p_M(k) = \frac{K(N+K-1)}{N^2}. \tag{D.25}$$

Inserting the expressions for $\mathrm{E}_K[M]$ and $\mathrm{E}_K[M^2]$ into (D.9), we obtain a closed-form expression for the time-slot fairness for the NCS algorithm.

5.2.4 Time-Slot Fairness for N-ORR

For the N-ORR algorithm, we will obtain the same random time-slot allocation within a round as for the random RR algorithm. Because of the same randomness of the time-slot allocation for the RR and N-ORR algorithms, these two algorithms will have the same time-slot fairness behavior.

5.3 Throughput Fairness for Different Scheduling Algorithms

5.3.1 Throughput Fairness for RR

The first moment of the throughput allocation for the RR algorithm can be written as follows:

$$\begin{aligned} \mathrm{E}_K[R] &= \frac{1}{N} \sum_{i=1}^{N} \sum_{k=0}^{K} p_M(k) \int_0^{\infty} k \log_2(1+\gamma) p_{\gamma_i}(\gamma) d\gamma \\ &= \frac{K}{N^2 \ln 2} \sum_{i=1}^{N} e^{1/\overline{\gamma}_i} E_1\left(\frac{1}{\overline{\gamma}_i}\right). \end{aligned} \tag{D.26}$$

Furthermore, the second moment of the throughput allocation can be expressed as:

$$\begin{aligned} \mathrm{E}_K[R^2] &= \frac{1}{N}\sum_{i=1}^{N}\sum_{k=0}^{K} p_M(k)\int_0^{\infty}(k\log_2(1+\gamma))^2 p_{\gamma_i}(\gamma)d\gamma \\ &= \frac{\mathrm{E}_K[M^2]}{N(\ln 2)^2}\sum_{i=1}^{N}\frac{1}{\overline{\gamma}_i}\Psi\left(\frac{1}{\overline{\gamma}_i}\right), \end{aligned} \tag{D.27}$$

where $\mathrm{E}_K[M^2]$ is the second moment of the time-slot allocation for RR and $\Psi(\mu)$ is given by

$$\begin{aligned} \Psi(\mu) &= \int_0^{\infty}\ln^2(1+\gamma)e^{-\mu\gamma}d\gamma \\ &= e^{\mu}\left\{\frac{1}{\mu}\left[\frac{\pi^2}{6}+(C+\ln(\mu))^2\right]-2_3F_3(1,1,1;2,2,2;-\mu)\right\}, \end{aligned} \tag{D.28}$$

with $C = 0.57721566490$ being Euler's constant [17, (9.73)], and ${}_pF_q(a_1,\cdots,a_p;b_1,\cdots,b_q;\cdot)$ being the *generalized hypergeometric function* [18]. This expression has been found using the derivation given in Appendix 3. Inserting the obtained expressions for the mean and the variance of the throughput allocation into (D.12), we obtain a closed-form expression for the throughput fairness of the RR algorithm.

5.3.2 Throughput Fairness for MCS

Before we can find the moments of the throughput allocation for MCS we first need to find the PDF of the CNR conditioned on user i being scheduled, $p_{\gamma_i^*}(\gamma)$:

$$\Pr(\gamma_i=\gamma|\gamma_j<\gamma_i,\forall j\neq i)=\frac{\Pr(\gamma_i=\gamma \text{ and } \gamma_j<\gamma_i,\forall j\neq i)}{\Pr(\gamma_j<\gamma_i,\forall j\neq i)}=\frac{p_{\gamma_i}(\gamma)}{p_i}\prod_{\substack{j=1\\ j\neq i}}^{N}P_{\gamma_j}(\gamma). \tag{D.29}$$

The first moment of the throughput allocation for the MCS algorithm

can now be found as follows:

$$\begin{aligned}
\mathrm{E}_K[R] &= \sum_{\mathbf{k}:\sum_{i=1}^{N} k_i = K} \left(\frac{1}{N} \sum_{i=1}^{N} \int_0^{\infty} k_i \log_2(1+\gamma) p_{\gamma_i^*}(\gamma) d\gamma \right) \\
&\times \frac{K!}{k_1! \cdot k_2! \cdots k_N!} \prod_{i=1}^{N} p_i^{k_i} \\
&= \frac{1}{N \ln 2} \sum_{\mathbf{k}:\sum_{i=1}^{N} k_i = K} \left[\sum_{i=1}^{N} \frac{k_i}{\overline{\gamma}_i p_i} \sum_{\tau \in T_i^N} \mathrm{sign}(\tau) \frac{e^{\left(\frac{1}{\overline{\gamma}_i} + |\tau|\right)}}{\frac{1}{\overline{\gamma}_i} + |\tau|} E_1 \left(\frac{1}{\overline{\gamma}_i} + |\tau| \right) \right] \\
&\times \frac{K!}{k_1! \cdot k_2! \cdots k_N!} \prod_{i=1}^{N} p_i^{k_i}. \qquad \text{(D.30)}
\end{aligned}$$

Similarly, we can obtain the second moment of the throughput allocation as:

$$\begin{aligned}
\mathrm{E}_K[R^2] &= \sum_{\mathbf{k}:\sum_{i=1}^{N} k_i = K} \left(\frac{1}{N} \sum_{i=1}^{N} \int_0^{\infty} [k_i \log_2(1+\gamma)]^2 p_{\gamma_i^*}(\gamma) d\gamma \right) \\
&\times \frac{K!}{k_1! \cdot k_2! \cdots k_N!} \prod_{i=1}^{N} p_i^{k_i} \\
&= \frac{1}{N (\ln 2)^2} \sum_{\mathbf{k}:\sum_{i=1}^{N} k_i = K} \left[\sum_{i=1}^{N} \frac{k_i^2}{\overline{\gamma}_i p_i} \sum_{\tau \in T_i^N} \mathrm{sign}(\tau) \Psi \left(\frac{1}{\overline{\gamma}_i} + |\tau| \right) \right] \\
&\times \frac{K!}{k_1! \cdot k_2! \cdots k_N!} \prod_{i=1}^{N} p_i^{k_i}. \qquad \text{(D.31)}
\end{aligned}$$

Inserting the expressions for $\mathrm{E}_K[R]$ and $\mathrm{E}_K[R^2]$ into (D.12), we subsequently obtain a closed-form expression for the throughput fairness of the MCS algorithm.

5.3.3 Throughput Fairness for NCS

For the NCS policy, each user will experience a MUD gain as if all the other users were i.i.d. with the *same* average CNR as this user. The CNR of user i in the time-slots he is scheduled is therefore distributed according to the following CDF [3]:

$$P_{\gamma_i^*}(\gamma) = P_{\gamma_i}^N(\gamma). \qquad \text{(D.32)}$$

Differentiating this expression with respect to γ, we obtain the following PDF for the NCS algorithm:

$$p_{\gamma_i^*}(\gamma) = N P_{\gamma_i}^{N-1}(\gamma) p_{\gamma_i}(\gamma). \qquad \text{(D.33)}$$

We can now use this PDF to find the first moment of the throughput allocation:

$$\begin{aligned} \mathrm{E}_K[R] &= \frac{1}{N}\sum_{i=1}^{N}\sum_{k=0}^{K} p_M(k)\int_0^\infty k\log_2(1+\gamma)p_{\gamma_i^*}(\gamma)d\gamma \\ &= \frac{K}{N\ln 2}\sum_{i=1}^{N}\sum_{j=0}^{N-1}\binom{N-1}{j}\frac{(-1)^j}{1+j}e^{\frac{1+j}{\overline{\gamma}_i}}E_1\left(\frac{1+j}{\overline{\gamma}_i}\right), \end{aligned} \quad \text{(D.34)}$$

where $p_i = \frac{1}{N}$ for all the users.

Similarly, we can obtain the second moment of the throughput allocation as:

$$\begin{aligned} \mathrm{E}_K[R^2] &= \frac{1}{N}\sum_{i=1}^{N}\sum_{k=0}^{K} p_M(k)\int_0^\infty [k\log_2(1+\gamma)]^2 p_{\gamma_i^*}(\gamma)d\gamma \\ &= \frac{K(N+K-1)}{(N\ln 2)^2}\sum_{i=1}^{N}\frac{1}{\overline{\gamma}_i}\sum_{j=0}^{N-1}\binom{N-1}{j}(-1)^j\Psi\left(\frac{1+j}{\overline{\gamma}_i}\right). \end{aligned} \quad \text{(D.35)}$$

As for the RR and MCS algorithms, the closed-form throughput fairness expression for NCS can subsequently be found by inserting these two moments into (D.12).

5.3.4 Throughput Fairness for N-ORR

Since the N-ORR algorithm schedules the users with the highest ratio $\chi_i(t) = \frac{\gamma_i(t)}{\overline{\gamma}_i}$ in each competition and since this ratio has the same distribution for all users, the PMF for the number of time-slots k being allocated to a user is independent of i and can be expressed as in (D.17). The users that get $k = \lfloor \frac{K}{N} \rfloor$ time-slots are only involved in whole rounds of competitions. This means that these users will not participate in the last round of competitions that is finished before all the users are allocated one time-slot each. In this case, user i will experience a CDF of the CNR when he is scheduled that equals the average CDF over one round:

$$P_{\gamma_i^*}\left(\gamma | k = \left\lfloor \frac{K}{N} \right\rfloor\right) = \frac{1}{N}\sum_{n=1}^{N} P_{\gamma_i}^n(\gamma), \quad \text{(D.36)}$$

However, the CDF of the CNR for a user getting $k = \lceil \frac{K}{N} \rceil$ out of K time-slots, when K is not a multiple of N (i.e. $\mathrm{mod}(K, N) \neq 0$), can be expressed

as the average over *all* the rounds, since such a user will also participate in the last unfinished round:

$$P_{\gamma_i^*}\left(\gamma|k=\left\lceil\frac{K}{N}\right\rceil\right)=\frac{(k-1)\sum_{n=1}^{N}P_{\gamma_i}^{n}(\gamma)}{kN}+\frac{\sum_{n=kN-K+1}^{N}P_{\gamma_i}^{n}(\gamma)}{k(K-(k-1)N)}. \quad \text{(D.37)}$$

Differentiating these CDFs with regard to γ, we obtain the corresponding PDFs:

$$p_{\gamma_i^*}\left(\gamma|k=\left\lfloor\frac{K}{N}\right\rfloor\right)=\frac{1}{N}\sum_{n=1}^{N}nP_{\gamma_i}^{n-1}(\gamma)p_{\gamma_i}(\gamma), \quad \text{(D.38)}$$

and

$$\begin{aligned}p_{\gamma_i^*}\left(\gamma|k=\left\lceil\frac{K}{N}\right\rceil\right) &= \frac{(k-1)\sum_{n=1}^{N}nP_{\gamma_i}^{n-1}(\gamma)p_{\gamma_i}(\gamma)}{kN}\\ &+\frac{\sum_{n=kN-K+1}^{N}nP_{\gamma_i}^{n-1}(\gamma)p_{\gamma_i}(\gamma)}{k(K-(k-1)N)}. \quad \text{(D.39)}\end{aligned}$$

It should be noted that when $\mathrm{mod}(K,N)=0$, the PDF of the CNR reduces to (D.38). We can now use these expressions to express the expected value of the throughput allocation when $\mathrm{mod}(K,N)\neq 0$:

$$\begin{aligned}\mathrm{E}_K[R] &= \frac{1}{N}\sum_{i=1}^{N}\sum_{k=0}^{K}p_M(k)\int_0^{\infty}k\log_2(1+\gamma)p_{\gamma_i^*}(\gamma|k)d\gamma\\ &= \frac{1}{N^2}\sum_{i=1}^{N}\left(\left\lfloor\frac{K}{N}\right\rfloor\sum_{n=1}^{N}A_i(n)+\sum_{n=\left\lceil\frac{K}{N}\right\rceil N-K+1}^{N}A_i(n)\right), \quad \text{(D.40)}\end{aligned}$$

where $A_i(n)$ is given by:

$$A_i(n)=\frac{n}{\ln 2}\sum_{j=0}^{n-1}\binom{n-1}{j}(-1)^j\frac{e^{\frac{1+j}{\overline{\gamma}_i}}}{1+j}E_1\left(\frac{1+j}{\overline{\gamma}_i}\right). \quad \text{(D.41)}$$

Similarly, we can express the second moment of the throughput allocation when $\mathrm{mod}(K,N)\neq 0$ as:

$$\begin{aligned}\mathrm{E}_K[R^2] &= \frac{1}{N}\sum_{i=1}^{N}\sum_{k=0}^{K}p_M(k)\int_0^{\infty}(k\log_2(1+\gamma))^2p_{\gamma_i^*}(\gamma|k)d\gamma\\ &= \frac{1}{N^2}\sum_{i=1}^{N}\left(\frac{K}{N}\left\lfloor\frac{K}{N}\right\rfloor\sum_{n=1}^{N}B_i(n)+\left\lceil\frac{K}{N}\right\rceil\sum_{n=\left\lceil\frac{K}{N}\right\rceil N-K+1}^{N}B_i(n)\right),\end{aligned} \quad \text{(D.42)}$$

where $B_i(n)$ is given by:

$$B_i(n) = \frac{n}{\overline{\gamma}_i(\ln 2)^2} \sum_{j=0}^{n-1} \binom{n-1}{j} (-1)^j \Psi\left(\frac{1+j}{\overline{\gamma}_i}\right). \tag{D.43}$$

To obtain the expression above, we have inserted the Rayleigh PDF and CDF and used the binomial expansion formula [17, Eq. (1.111)]. The resulting integral has been solved by using the derivation in the Appendix.

When $\mathrm{mod}(K, N) = 0$, the two first moments of the throughput allocation reduces to:

$$\mathrm{E}_K[R] = \frac{K}{N^3} \sum_{i=1}^{N} \sum_{n=1}^{N} A_i(n), \tag{D.44}$$

and

$$\mathrm{E}_K[R^2] = \frac{K^2}{N^4} \sum_{i=1}^{N} \sum_{n=1}^{N} B_i(n). \tag{D.45}$$

Inserting the expressions for $\mathrm{E}_K[R]$ and $\mathrm{E}_K[R^2]$ into (D.12), we finally obtain a closed-form expression for the throughput fairness of the N-ORR algorithm.

6 Numerical Results

6.1 MASSE Plots

Fig. D.1 shows a plot of the MASSE as a function of N for users with i.i.d. Rayleigh channels with $\overline{\gamma} = 15$ dB. As expected, the NCS algorithm does achieve the same MASSE as the MCS algorithm when the users' channels are i.i.d.. Since the ORR algorithm is a combination of RR and MCS, we see that the MASSE of ORR lies between the MASSE of MCS and the MASSE of RR. Also note that the relative difference between the MASSE of MCS and ORR is decreasing for an increasing number of users.

The plot in Fig. D.2 shows the MASSE for users with i.n.d. channels where the users have average CNRs that have been chosen deterministically in an interval around 15 dB. The lowest value of this interval is 5 dB, and the highest is 17.79 dB. We have looked at the case where the users have average CNRs that are evenly spread within the interval, and the case where half of the users have an average CNRs of 5 dB and the other half have an average CNRs of 17.79 dB. For the clustered case, one user has an average CNR of 15 dB if we have an odd number of users.

Since the NCS algorithm does not schedule the user with the *absolute* best channel for each time-slot, NCS will always have lower MASSE than

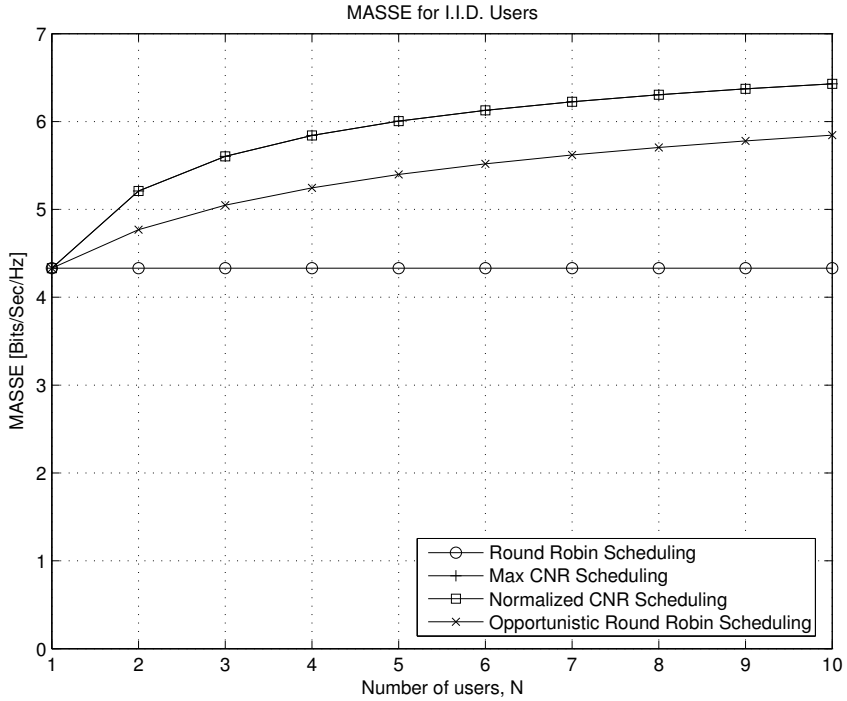

FIGURE D.1: MASSE for users with i.i.d. Rayleigh fading channels with an average CNR of 15 dB.

MCS for the i.n.d. scenario. It should also be observed that the MASSE of the RR, NCS and N-ORR scheduling algorithms will suffer the more the users' average CNRs deviate from the overall average CNR value. Mathematically, this can be understood in the following way. All these three scheduling algorithms will have access probability $p_i = \frac{1}{N}$ for all the users, and the sum in (D.1) can therefore be moved into the integral. The expression for the MASSE now consists of N terms within the integral. For the i.n.d. scenario, the value of γ from term to term will have a higher variance compared to the i.i.d. scenario. Since $\log_2(1+\gamma)$ is a concave function, it can now be argued that the sum of N terms inside the integral is higher for the i.i.d. scenario compared to the i.n.d. scenario. Consequently, the resulting MASSE for the i.n.d. scenario will always be lower than the corresponding MASSE for the i.i.d. scenario for algorithms having the same access probability $p_i = \frac{1}{N}$ for all the users.

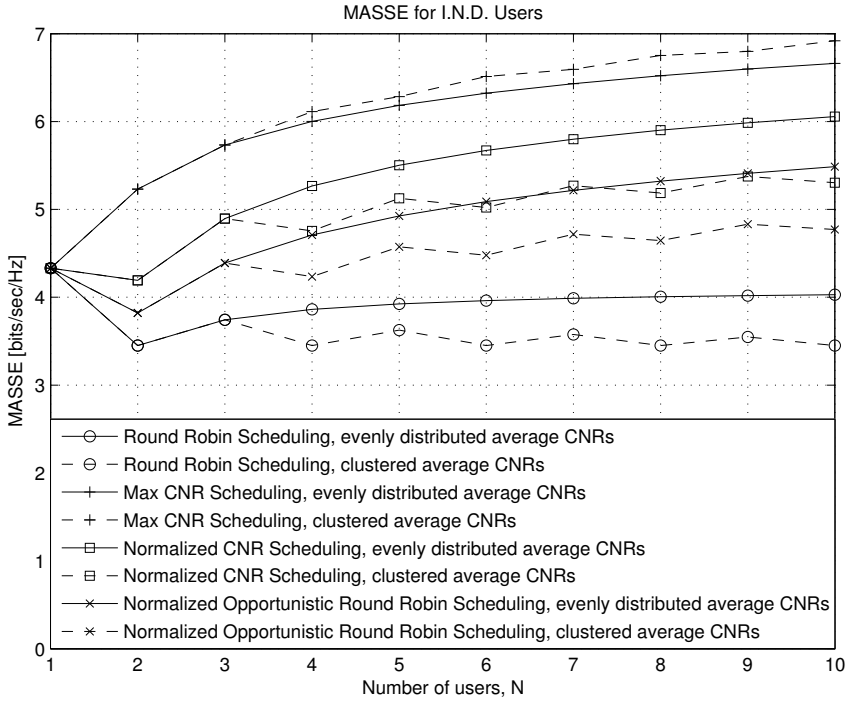

FIGURE D.2: MASSE for users with i.n.d. Rayleigh fading channels with a total average CNR of 15 dB. The average CNR of the user(s) with the worst channel quality is 5 db, while the average CNR of the user(s) with the best channel quality is 17.79 dB.

6.2 Fairness Plots

Figs. D.3, D.4, and D.5 show the time-slot fairness as a function of K for four users with Rayleigh fading channels with a total average CNR of 15 dB. For the plot in Fig. D.3, the users' average CNRs are i.i.d., while for the plot in Fig. D.4 the users' deterministic average CNRs are evenly spread between 5 dB and 17.79 dB. For the plot in Fig. D.5 the average CNRs are clustered at 5 dB for two of the users, and at 17.79 dB for the other two users. In all cases we see that the RR and (N-)ORR algorithms converge relatively fast to unity time-slot fairness. It should also be observed that the time-slot fairness of these two algorithms is perfect when the last competition of a round is finished.

For the plot where the users' channels are i.i.d., all the algorithms con-

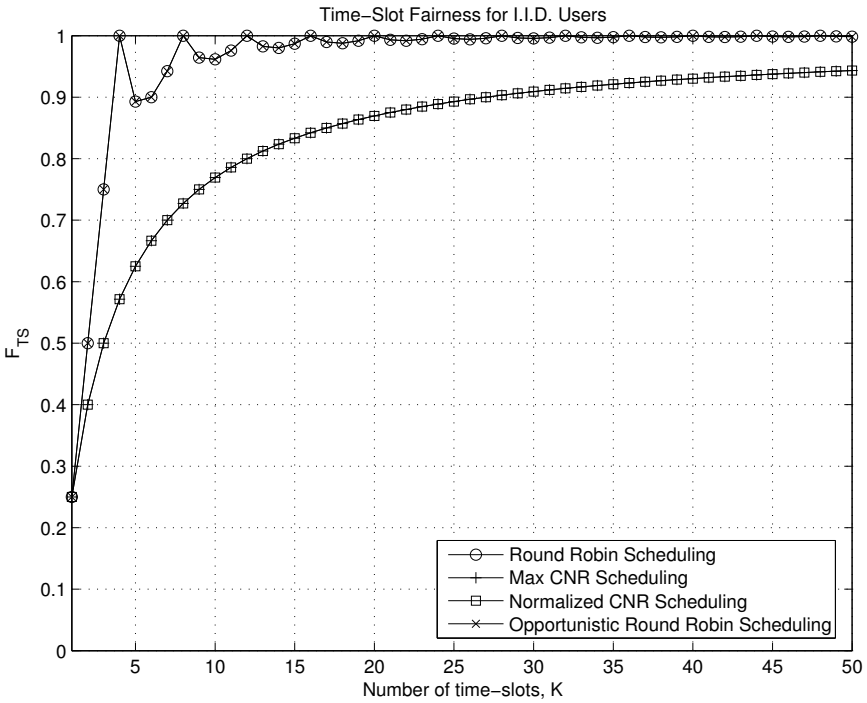

FIGURE D.3: Time-slot fairness for four users with i.i.d. Rayleigh fading channels with an average CNR of 15 dB.

verge to unity time-slot fairness (See Sec. 5.1.1). When the users' average CNRs are i.n.d., the time-slot fairness of the RR, NCS, and N-ORR algorithms still converge to unity. However, the time-slot fairness of the MCS algorithm converges to 0.6220 for large values of K when the average CNRs are evenly distributed between 5 and 17.79 dB and to 0.5083 when the users' average CNRs are clustered at 5 dB for two users and 17.19 for the two other users. We see that the time-slot fairness converges slower when the users' average CNRs deviate much from the total average CNR. Therefore, we can conclude that the scheduling algorithms will yield higher fairness the closer the users' average CNRs are to the total average CNR in the cell.

Plots of $F_{TP}(K)$ are shown in Figs. D.6, D.7, and D.8 for the average CNR values used in Figs. D.3, D.4, and D.5, respectively. We can observe that the curves have similar shapes as the corresponding curves in the time-slot fairness plots. However, since the throughput fairness is affected by the variance of the number of bits allocated in each time-slot, the throughput

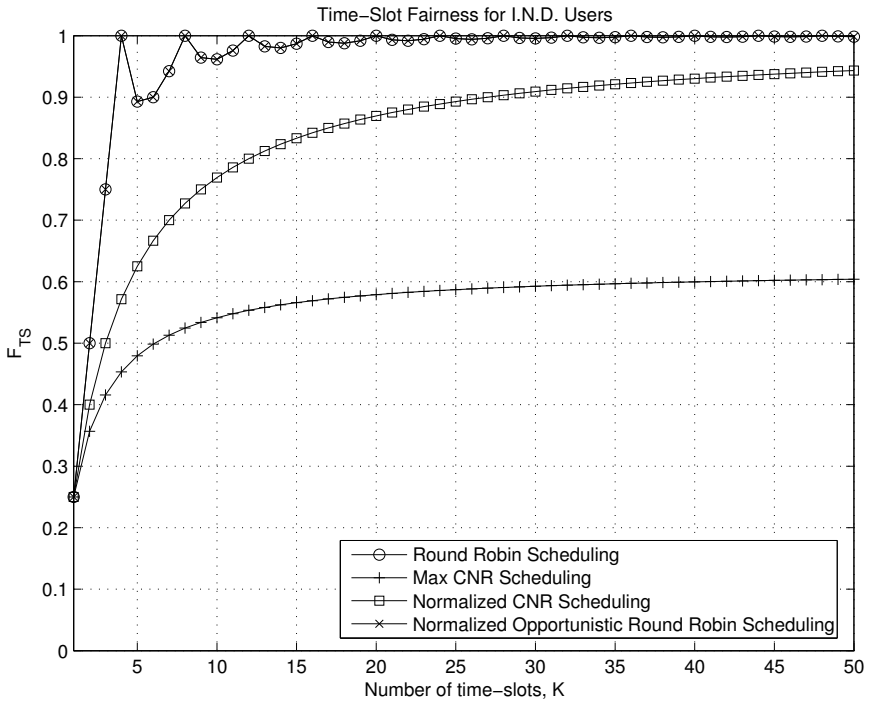

FIGURE D.4: Time-slot fairness for four users with i.n.d. Rayleigh fading channels with a total average CNR of 15 dB. The average CNR of the users are evenly distributed from 5 dB to 17.79 dB.

fairness will always be lower than the corresponding time-slot fairness. It is also interesting to note that since the throughput fairness is measured relative to the expected throughput of the scheduling algorithm, the ORR algorithm will obtain a higher throughput fairness than the RR algorithm, since the ORR algorithm obtains a higher spectral efficiency than the RR algorithm (See Fig. D.2).

By using similar derivations as in Section 5.2, and inserting the results into (D.15), we can obtain closed-form expression for the asymptotic throughput fairness values as K goes to infinity. From these expressions we have calculated the asymptotic throughput fairness for our i.i.d. scenario to be 0.8855, 0.9804, 0.9804, and 0.9436, for RR, MCS, NCS, and ORR, respectively. The asymptotic throughput fairness values for the scenario where the users have evenly distributed average CNRs are 0.7940, 0.5862, 0.9028, and 0.8604 for RR, MCS, NCS, and N-ORR, respectively; while the

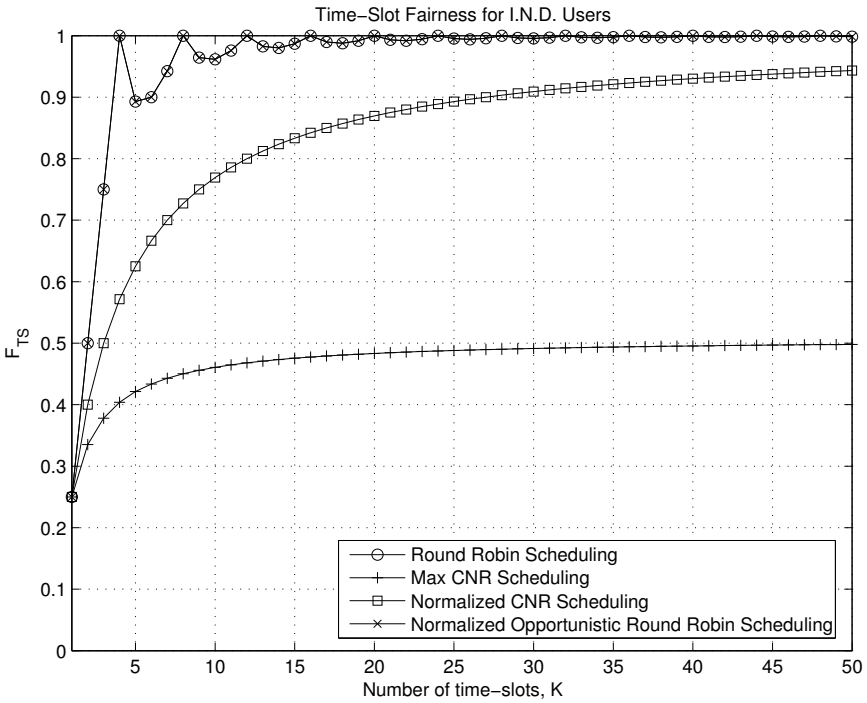

FIGURE D.5: Time-slot fairness for four users with i.n.d. Rayleigh fading channels with a total average CNR of 15 dB. The average CNR of two users are clustered at 5 dB, while the average CNR of the other two users are clustered at 17.17 dB.

corresponding values for the scenario where the users' average CNRs are clustered are 0.7101, 0.4883, 0.8317, and 0.7843, for RR, MCS, NCS, and N-ORR, respectively. These asymptotic values correspond well with the throughput fairness plots shown in Figs. D.6, D.7, and D.8 for large values of K. It is interesting to note that the NCS algorithm converges to the highest throughput fairness for large values of K. This means that for applications needing short-term fairness, the ORR algorithm will give the best performance, while for applications needing long-term fairness, the NCS algorithm will give the best performance.

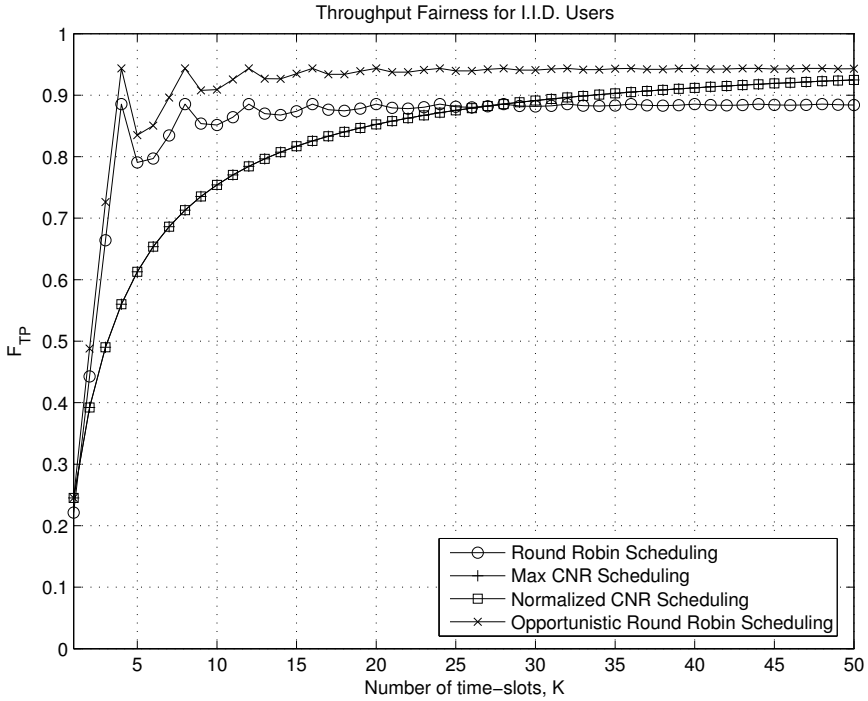

FIGURE D.6: Throughput fairness for four users with i.i.d. Rayleigh fading channels with an average CNR of 15 dB.

7 Conclusion

In this paper we have conducted an analytical evaluation of different scheduling algorithms when the users' channels have different average CNRs. Closed-form expressions for the system spectral efficiency, the time-slot fairness, and the throughput fairness have been found for the RR, MCS, NCS, ORR, and N-ORR scheduling algorithms. In addition, we have developed expressions for the asymptotic time-slot fairness when and throughput fairness when the length of the time-window over which the fairness is calculated goes to infinity. Our numerical results show that while the MCS and RR algorithms either have high MASSE or high fairness, the NCS, ORR, and N-ORR algorithms can obtain both relatively high MASSE and high fairness at the same time.

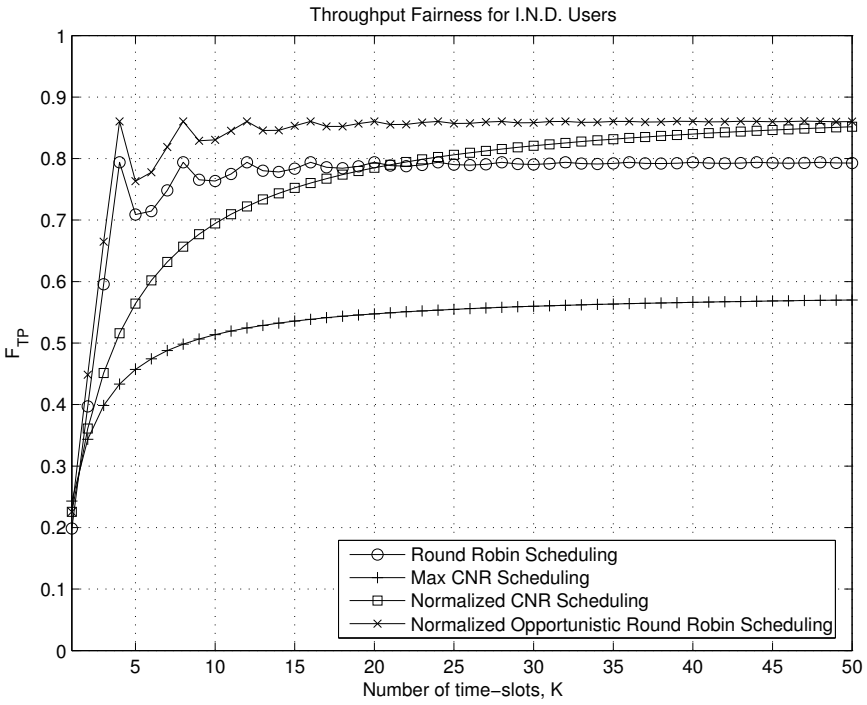

FIGURE D.7: Throughput fairness for four users with i.n.d. Rayleigh fading channels with a total average CNR of 15 dB. The average CNR of the users are evenly distributed from 5 dB to 17.79 dB.

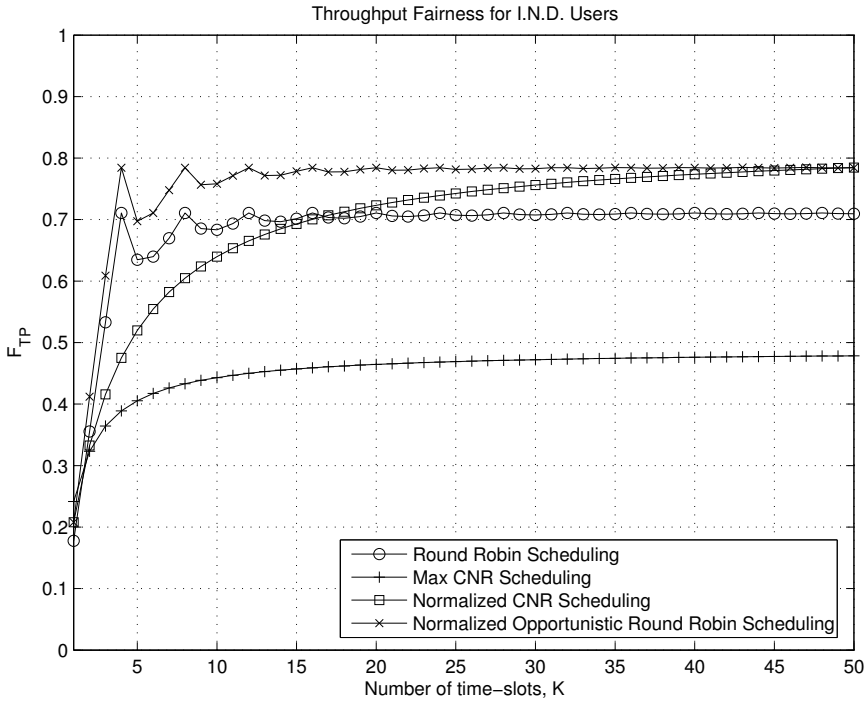

FIGURE D.8: Throughput fairness for four users with i.n.d. Rayleigh fading channels with a total average CNR of 15 dB. The average CNR of two users are clustered at 5 dB, while the average CNR of the other two users are clustered at 17.17 dB.

References

[1] S. S. Kulkarni and C. Rosenberg, "Opportunistic scheduling policies for wireless systems with short term fairness constraints," in *Proc. IEEE Global Communications Conf. (GLOBECOM'03)*, (San Francisco, CA, USA), pp. 533–537, Dec. 2003.

[2] M. Johansson, "Diversity-enhanced equal access - considerable throughput gains with 1-bit feedback," in *Proc. IEEE Workshop on Signal Processing Advances in Wireless Communications (SPAWC'04)*, (Lisbon, Portugal), July 2004.

[3] R. Knopp and P. A. Humblet, "Information capacity and power control in single cell multiuser communications," in *Proc. IEEE Int. Conf. on Communications (ICC'95)*, (Seattle, WA, USA), pp. 331–335, June 1995.

[4] P. Viswanath, D. N. C. Tse, and R. Laroia, "Opportunistic beamforming using dumb antennas," *IEEE Trans. Inform. Theory*, vol. 48, pp. 1277–1294, June 2002.

[5] Y. Liu and E. Knightly, "Opportunistic fair scheduling over multiple wireless channels," in *Proc. IEEE Joint Conference of the Computer and Communications Societies (INFOCOM'03)*, vol. 2, (San Francisco, CA, USA), pp. 1106–1115, Mar. 2003.

[6] L. Yang and M.-S. Alouini, "Performance analysis of multiuser selection diversity," in *Proc. IEEE Int. Conf. on Communications (ICC'04)*, (Paris, France), pp. 3066–3070, June 2004.

[7] V. Hassel, M. R. Hanssen, and G. E. Øien, "Spectral efficiency and fairness for opportunistic round robin scheduling." in *Proc. IEEE Int. Conf. Comm. (ICC'06)*, (Istanbul, Turkey), June 2006.

[8] Z. Ji, Y. Yang, J. Zhou, M.Takai, and R. Bagrodia, "Exploiting medium access diversity in rate adaptive wireless LANs," in *Proc. ACM International Conference on Mobile Computing and Networking (MOBICOM'04)*, (Philadephia, PA, USA), pp. 345–359, Sept.-Oct. 2004.

[9] A. J. Goldsmith and P. P. Varaiya, "Capacity of fading channels with channel side information," *IEEE Trans. Inform. Theory*, vol. IT-43, pp. 1896–1992, Nov. 1997.

[10] M.-S. Alouini and A. J. Goldsmith, "Capacity of Rayleigh fading channels under different adaptive transmission and diversity-combining techniques," *IEEE Trans. on Veh. Technol.*, vol. 48, pp. 1165–1181, July 1999.

[11] T. Eng, N. Kong, and L. B. Milstein, "Comparison of diversity combining techniques for Rayleigh-fading channels," *IEEE Trans. Commun.*, vol. 44, pp. 1117–1129, Sept. 1996.

[12] R. Jain, D. Chiu, and W. Hawe, "A quantitative measure of fairness and discrimination for resource allocation in shared computer systems," DEC Research Report TR-301, Digital Equipment Corporation, Maynard, MA, USA, Sept. 1984.

[13] H. Sirisena, A. Haider, M. Hassan, and K. Fawlikowski, "Transient fairness of optimized end-to-end window control," in *Proc. IEEE Global Communications Conference (GLOBECOM'03)*, (San Francisco, CA, USA), pp. 3979 – 3983, Dec. 2003.

[14] G. Berger-Sabbatel, A. Duda, O. Gaudoin, M. Heusse, and F. Rousseau, "Fairness and its impact on delay in 802.11 networks," in *Proc. IEEE Global Communications Conference (GLOBECOM'04)*, (Dallas, TX, USA), pp. 2967–2973, Dec. 2004.

[15] Wolfram Research Inc., "Multinomial distribution." `http://mathworld.wolfram.com/MultinomialDistribution.html`.

[16] E. Kreyszig, *Advanced Engineering Mathematics*. New York: John Wiley & Sons, Inc., 7th ed., 1993.

[17] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products*. San Diego, CA, USA: Academic Press, 6th ed., 2000.

[18] Wolfram Research Inc., "Generalized hypergeometric function: Primary definition." `http://functions.wolfram.com/07.31.02.0001.01`.

Paper E

Throughput Guarantees for Wireless Networks with Opportunistic Scheduling: A Comparative Study

Vegard Hassel, Geir E. Øien, and David Gesbert

Accepted for publication in
IEEE Transactions on Wireless Communications

Abstract

In this letter we develop an expression for the approximate *throughput guarantee violation probability* (TGVP) for users in time-slotted networks for any scheduling algorithm with a given mean and variance of the bit-rate in a time-slot, and a given distribution for the number of time-slots allocated within a time-window. Based on this general result, we evaluate closed-form expressions for the TGVPs for four well-known scheduling algorithms. Through simulations we also show that our TGVP approximation is tight for a realistic network with moving users with correlated channels and realistic throughput guarantees.

1 Introduction

In modern cellular network standards like HSPA, 1xEVDO, and Mobile WiMAX, the rate of a user is adapted to the channel quality [1]. By giving priority to users with high channel quality, the system capacity can be increased significantly [2, 3]. However, fulfilling the users' (quality-of-service) QoS requirements in such a system can be difficult since the users with the lowest channel quality will often be starved. Consequently, it is necessary to implement scheduling algorithms that take both the channel quality and the QoS demands of the users into account.

Many previous publications have concentrated on analyzing how *fair* the resource allocation is in the network [4, 5]. However, it can be difficult to quantify fairness and the concept of fairness can often be difficult to understand both for the operators and the mobile users. In commercial networks it is more useful to look at a more precise notion of QoS, namely *throughput guarantees*. The advantage of being able to quantify throughput guarantees will make it easier for the network operator to offer a service that is tailor-made to the applications that are going to be transmitted. In addition, the network operators do not have to over-dimension the wireless networks to satisfy the QoS demands of the customers.

There are two types of throughput guarantees that can be offered to customers, namely *hard* or *deterministic throughput guarantees*, and *soft* or *stochastic throughput guarantees*. The hard throughput guarantees promise, with unit probability, a certain throughput to the users within a given time-window, while soft throughput guarantees promise that each user will have a specified throughput within a given time-window, with a probability that is high, but less than unity. For telecommunications networks in general, and for wireless networks in particular, soft throughput guarantees are more suited for specifying QoS than hard throughput guarantees. This is because such networks often have a varying number of users and varying loads from the applications of these users. For wireless networks, the varying quality of the radio channel will further add uncertainty to the size of the throughput that can be guaranteed for short time-spans. In addition, will opportunistic scheduling give priorities to the users with the best channel conditions (subject to various constraints), and the waiting period between each time a user is scheduled can therefore vary significantly. This makes soft throughput guarantees suited as QoS metrics for modern wireless networks.

Obtaining analytical expressions for what soft throughput guarantees that can be offered in a wireless network makes it possible to calculate the QoS of the users in a very efficient way for a set of instantaneous system

parameters. Such analytical expressions can therefore be used directly in adaptive radio resource algorithms for wireless networks where the users move around with high speed and where real-time applications constitute the dominating traffic load.

Contributions. Quantifying the soft throughput guarantees that can be given for a certain scheduling algorithm without conducting experimental investigations has, to the best of our knowledge, not been looked into before. We obtain a general expression for a tight approximation of the *throughput guarantee violation probability* (TGVP), for a given mean and variance of the number of bits transmitted in a time-slot, and a given distribution for the number of time-slots allocated to a user within a time-window. We also investigate the tightness of this approximation for a realistic scenario with users that have correlated channels[1].

Organization. The rest of this letter is organized as follows. In Section 2 we present the system model. We develop a general expression for the approximate TGVP in Section 3. In Section 4 we plot closed-form expressions for the approximate TGVP for four different scheduling algorithms and analyze the tightness of the approximation by comparing the analytical results with simulations for a realistic scenario. Our conclusions are presented in Section 5.

2 System Model

We consider a single base station that serves N users using time-division multiplexing (TDM). The analytical results will be valid for the downlink, however also for the uplink if reciprocity can be assumed between the downlink and uplink. In any case we assume that the total bandwidth available for the users is W [Hz] and that the transmit power is constant for all transmitters. Each user measures his own CNR perfectly, and before performing scheduling, the base station is assumed to receive these measurements from all the users. For each time-slot the base station takes a scheduling decision and broadcasts this decision to the selected user before transmission starts. We assume that the channels of the users are flat Rayleigh block-fading channels with a constant average received CNR $\overline{\gamma}_i$ for user i. The variations in average CNR in real-life networks is often on the time-scale of several seconds, while realistic throughput guarantees are calculated for time-scales under 100 milliseconds. Consequently, it is realistic to assume that the average CNRs are constant over the time-window for which the throughput guarantees are calculated.

[1]Parts of this letter are based on work in [6] and [7].

The block or time-slot duration, T_{TS} [seconds], is assumed to be less than one coherence time, i.e., the channels can be regarded more or less as constant during one time-slot. To obtain our analytical results we also assume that the CNR values from time-slot to time-slot are uncorrelated. This means that one user will very seldom experience two adjacent time-slots with the same CNR values, and consequently, the opportunistic distribution of time-slots between the users appear to be more fair. This will influence our analytical results to some extent since it is easier to fulfill the throughput guarantees within a given time-window when such a channel model is assumed.

Another important assumption is that the users always have data to send or transmit. For real-time applications this is often a realistic assumption because the packet flow from the applications is relatively constant in this case.

3 How to Quantify the Throughput Guarantees

A soft throughput guarantee can be expressed as the probability of *not* fulfilling a given throughput guarantee, i.e., the *throughput guarantee violation probability*, TGVP. Defining the desired throughput guarantee as guaranteeing a throughput of B [bits] over a time-window T_W [seconds] for all N users with probability at least $1-\epsilon$, we can analytically define the problem as attempting to constrain the TGVP to be less than or equal to ϵ [8]:

$$\Pr(b_i < B) \leq \epsilon, \ i = 1, 2, \cdots, N, \tag{E.1}$$

i.e., the probability of the number of bits b_i being transmitted to or from user i within a time-window T_W being below B, should be less than or equal to ϵ.

3.1 Computing Throughput Guarantee Violation Probabilities

To be able to obtain an exact TGVP we would have to find a probability mass function (PMF) for the sum of bits that a user can transmit in the M time-slots he is allocated. From [9] and several other publications, we conclude that finding an exact closed-form expression for the value of the TGVP $\Pr(b_i < B)$ is a complex problem that has not yet been solved, and may very well not be solvable in closed form. We will therefore instead look at how we can *approximate* the TGVP.

We now formulate a proposition that can be used as a tool to specify an achievable soft throughput guarantee of B bits over a time-window T_W

constituting K time-slots. For users transmitting over a time-slotted block fading channel, with $b_{i,j}$ bits being transmitted to or from user i in the jth time-slot he is scheduled, and the probability that user i gets $M = k$ out of K time-slots denoted as $p_M(k|i)$, the following holds:

Proposition: The probability that the throughput constraint B is violated over K time-slots for user i can be approximated as:

$$\Pr(b_i < B) \approx p_M(0|i) + \frac{1}{2}\sum_{k=1}^{K} p_M(k|i)\,\mathrm{erfc}\left(-\frac{B/k - \mu_{\bar{b}_{i,k}}}{\sqrt{2}\sigma_{\bar{b}_{i,k}}}\right), \tag{E.2}$$

where $\bar{b}_{i,k} = \frac{1}{k}\sum_{j=1}^{k} b_{i,j}$ is the average number of bits being transmitted to or from user i when he is allocated $M = k$ time-slots, and $\mu_{\bar{b}_{i,k}}$ and $\sigma^2_{\bar{b}_{i,k}}$ is the mean and variance of $\bar{b}_{i,k}$, respectively, and $\mathrm{erfc}(x) = \frac{2}{\sqrt{\pi}}\int_x^\infty e^{-t^2}dt$ is the *complementary error function*.

Proof: The allocation of different number of time-slots to a user constitute mutually exclusive events. The TGVP for user i over K time-slots can therefore be expressed as follows, using the law of total probability:

$$\begin{aligned}\Pr(b_i < B) &= \Pr(b_i < B|0)\cdot p_M(0|i)\\ &+ \Pr(b_i < B|1)\cdot p_M(1|i)\\ &\ldots\\ &+ \Pr(b_i < B|K)\cdot p_M(K|i),\end{aligned} \tag{E.3}$$

where $\Pr(b_i < B|k)$ denotes the TGVP when user i is assigned $M = k$ time-slots and $p_M(k|i)$ denotes the probability that user i gets $M = k$ time-slots within the interval of K time-slots.

To be able to discuss a total throughput guarantee B within K time-slots, we first consider the number of bits transmitted to or from user i within the jth time-slot he is scheduled, and denote this number by $b_{i,j}$. For a system using constant transmit power and capacity-achieving codes which operate at the Shannon capacity limit we will have $b_{i,j} = T_{\mathrm{TS}} W \log_2(1 + \gamma_{i,j})$, where $\gamma_{i,j}$ is the CNR in the jth time-slot user i is scheduled.

We can now express the probability for violating the throughput guarantee B when k out of K time-slots are scheduled to user i as:

$$\begin{aligned}\Pr(b_i < B|k) &= \Pr\left(\sum_{j=1}^{k} b_{i,j} < B\right)\\ &= \Pr\left(\bar{b}_{i,k} < \frac{B}{k}\right)\\ &\approx \frac{1}{2}\,\mathrm{erfc}\left(-\frac{B/k - \mu_{\bar{b}_{i,k}}}{\sqrt{2}\sigma_{\bar{b}_{i,k}}}\right),\end{aligned} \tag{E.4}$$

where $\text{erfc}\left(-\frac{x-\mu}{\sqrt{2}\sigma}\right) = \Pr(X \leq x)$ is the cumulative distribution function (CDF) of a Gaussian distributed random variable X with mean μ and variance σ^2. In the expression in (E.4) we have $\mu_{\bar{b}_{i,k}} = \mu_{b_{i,j}}$ and $\sigma^2_{\bar{b}_{i,k}} = \sigma^2_{b_{i,j}}/k$ where $\mu_{b_{i,j}}$ and $\sigma^2_{b_{i,j}}$ are the mean and variance of the number of bits transmitted to or from user i in the jth time-slot he is scheduled. The approximation above has been obtained by using the *Central Limit Theorem* (CLT) [10, p. 1231].

By inserting (E.4) into (E.3), we see that the expression for the total throughput guarantee can be expressed as in (E.2). □

4 Numerical Results

In this section we plot and compare the expressions for the approximate TGVPs for four different scheduling algorithms. We also evaluate the accurateness of these expressions. However, before evaluating the plots, we choose to comment on the system parameters used in this section.

4.1 Realistic System Parameters for Cellular Networks

For the wireless standards 1xEVDO, HSDPA, and Mobile WiMAX, the time-slot length for the downlink is respectively 1.67, 2, and 5 ms [1]. The European IST research project WINNER I has suggested a time-slot duration of 0.34 ms for a future wireless system [11]. According to [12], the maximum one-way delay over a wireless HSDPA link should lie between 80 and 150 ms for voice over IP (VoIP) conversations to achieve good speech quality. If we assume that $T_W = 80$ ms, K equals 235, 48, 40, and 16 time-slots for WINNER I, 1xEVDO, HSDPA, and Mobile WiMAX, respectively.

The raw throughput needed for one-way, telephone-quality speech varies from about 5 kbit/s up to 64 kbit/s [13]. The corresponding raw throughput needed for one-way videoconferencing varies from 64 kbit/s up to 500 kbit/s. In addition a minimum of 4 percent protocol overhead has to be added. From these throughput demands and the value of T_W, realistic values for B can be calculated for each application session for a given set of system parameters.

4.2 Comparison of the TGVP of Different Scheduling Algorithms

Figs. E.1 and E.2 show the TGVP-performance of different scheduling algorithms for 10 users requesting B bits within a time-window $T_W = 80$ ms for a system with the time-slot length of Mobile WiMAX and WINNER I, respectively. We have plotted the TGVP performance for four algorithms,

namely Round Robin Scheduling (RR), Maximum CNR Scheduling (MCS), Normalized CNR Scheduling (NCS) and Normalized Opportunistic Round Robin Scheduling (ORR). By using the expressions in Table E.1 and inserting $p_M(k|i)$, $\mu_{\bar{b}_{i,k}} = \mathrm{E}[b_{i,j}]$ and $\sigma^2_{\bar{b}_{i,k}} = (\mathrm{E}[b^2_{i,j}] - (\mathrm{E}[b_{i,j}])^2)/k$ into (E.2), we obtain the TGVP approximations for these four scheduling algorithms. For the RR policy, the time-slots are allocated to the users in a sequential manner, i.e. totally non-opportunistically. The most opportunistic algorithm is the MCS policy because it always schedules the user with the highest CNR, and hence the highest rate. The NCS policy is a more fair policy because it schedules the users with the highest CNR relative to their own average CNR [14]. The ORR policy was introduced in [15] and is a combination of the RR and MCS policies. For this algorithm, the time-slots are allocated in rounds of N competitions where the users are guaranteed to be assigned one time-slot in each round. For the first competition the best user is chosen. This user is than taken out from the rest of the competitions in the round, and for the second time-slot the best of the remaining users are chosen. For each competition a new user is taken out and for the last time-slot in a round the channel is assigned to the remaining user. If the users average CNRs are spread far apart, the ORR algorithm will have the same spectral efficiency as conventional RR Scheduling. To have a more efficient ORR algorithm for this scenario, we have modified this algorithm such that the user with the highest normalized CNR is chosen in each competition. We refer to this algorithm as the Normalized-ORR (N-ORR) algorithm.

Figs. E.1 and E.2 are plotted for a user with $\overline{\gamma}_i = 5$, where the all the users' channels are Rayleigh distributed with constant average CNRs that have a total average of 15 dB. The user with the worst channel has an average CNR of $\overline{\gamma}_i = 5$ dB and the user with the best channel has $\overline{\gamma}_i = 17.79$ dB. We have chosen to plot the TGVP for the user with the worst channel because this user will have the lowest TGVP values of all the users in the system. The most interesting parts the figures are where the TGVP is close to zero, since for these low TGVP values it is a high probability that the throughput guarantee is fulfilled. We can observe that for both the Mobile WiMAX and the WINNER I systems, the N-ORR algorithm shows the best TGVP-performance. This algorithm can support close to hard throughput guarantees up to about 0.5 bits/sec/Hz for Mobile WiMAX, while the corresponding throughput guarantee limit for the WINNER I system is over 2 bits/sec/Hz. The reason why this value more than quadruples from $K = 16$ time-slots to $K = 235$ time-slots is that the more time-slots we have within the time-window, the higher is the likelihood that all the users will be assigned some time-slots with good channel conditions. Hence, it

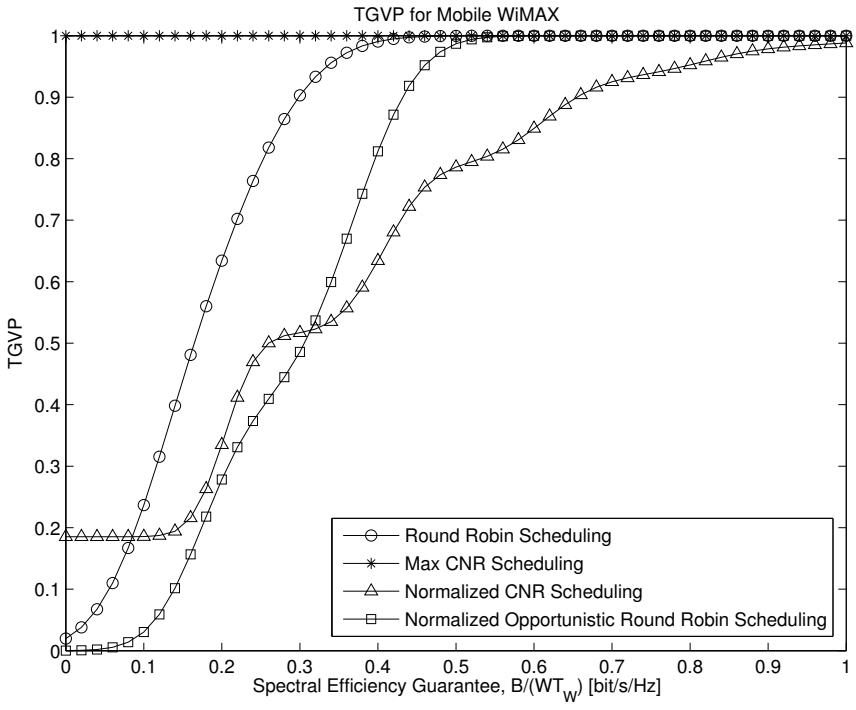

FIGURE E.1: Approximated Throughput Guarantee Violation Probability for a specific user i experiencing Rayleigh fading with $\overline{\gamma}_i = 5$ dB. There are 9 other users in the cell. Plotted for the Mobile WiMAX time-slot length of 5 ms and a time-window of $T_W = 80$ ms, corresponding to $K = 16$ time-slots.

will be easier to obtain a low TGVP for large values of K.

Also the RR algorithm shows a relatively good TGVP-performance for $K = 16$ time-slots. This is because this algorithm can promise that all the user get at least one time-slot within a time-window of N time-slots. The MCS algorithm is not very useful to guarantee any throughput for the user with the worst channel. This is because the user with the highest CNR is chosen at all times and there is therefore a low probability that the user with $\overline{\gamma}_i = 5$ dB is chosen.

In this paper we have assumed that only one user is scheduled in each time-slot. Since both Mobile WiMAX and WINNER I are based on orthogonal frequency-division multiplexing (OFDM) with respectively 720 and 1664 sub-carriers for user data, it is possible to schedule more user within

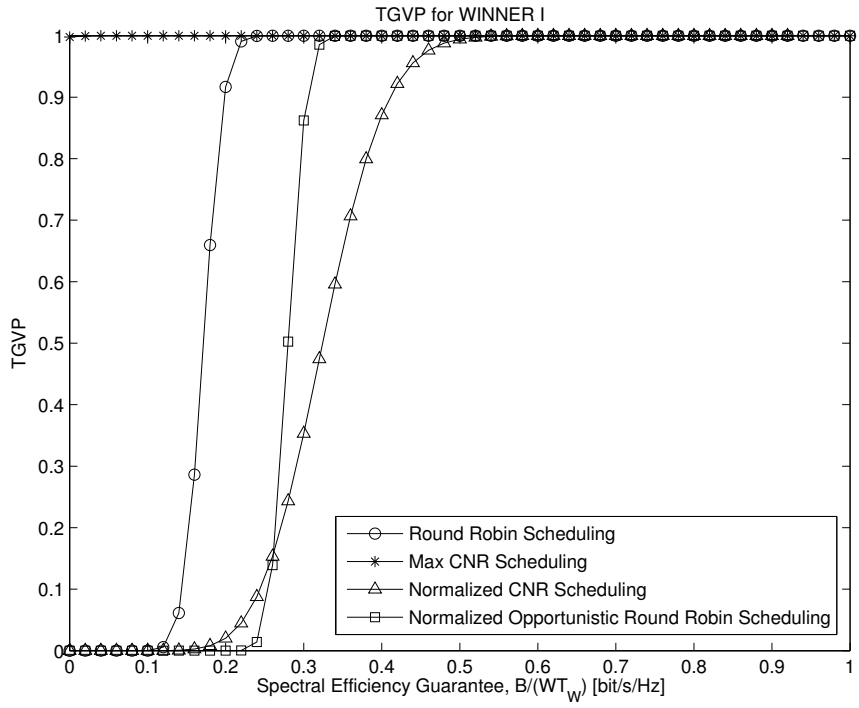

FIGURE E.2: Approximated Throughput Guarantee Violation Probability for a specific user i experiencing Rayleigh fading with $\overline{\gamma}_i = 5$ dB. There are 9 other users in the cell. Plotted for the WINNER I time-slot length of 0.34 ms and a time-window of $T_W = 80$ ms, corresponding to $K = 235$ time-slots.

the same time-slot for these systems, if we assume that channel estimates of each sub-carrier are available at the base station [1, 11]. Consequently, the corresponding TGVP performance for OFDM-based systems will be higher than the results shown in this paper. How much the TGVP performance will increase for a OFDM-based system model depends on the CNR correlation between the sub-carriers. Our closed-form expressions can also be used to obtain TGVP approximations for this system model by replacing K with $K \cdot N_{SC}$ and W with W_{SC}, where N_{SC} is the number of sub-carriers and W_{SC} is the bandwidth of each sub-carrier.

4.3 On the Accuracy of the Approximate TGVP

Figs. E.3 and E.4 show the TGVP approximations for N-ORR together with the corresponding Monte Carlo simulated TGVPs for respectively Mobile WiMAX and WINNER I. The approximate results are based on the assumption that the time-slots are uncorrelated, while the Monte Carlo simulations are for users that have a correlated CNR from time-slot to time-slot. We have used Jakes' correlation model with carrier frequency of $f_c = 1$ GHz and a user speed of $v = 30$ m/s. The channel gain is modeled as a sum of sinusoids with correlation coefficient $f_D T_{\mathrm{TS}} = \frac{v f_c}{c}$, where f_D is the Doppler frequency shift and c is the speed of light [16].

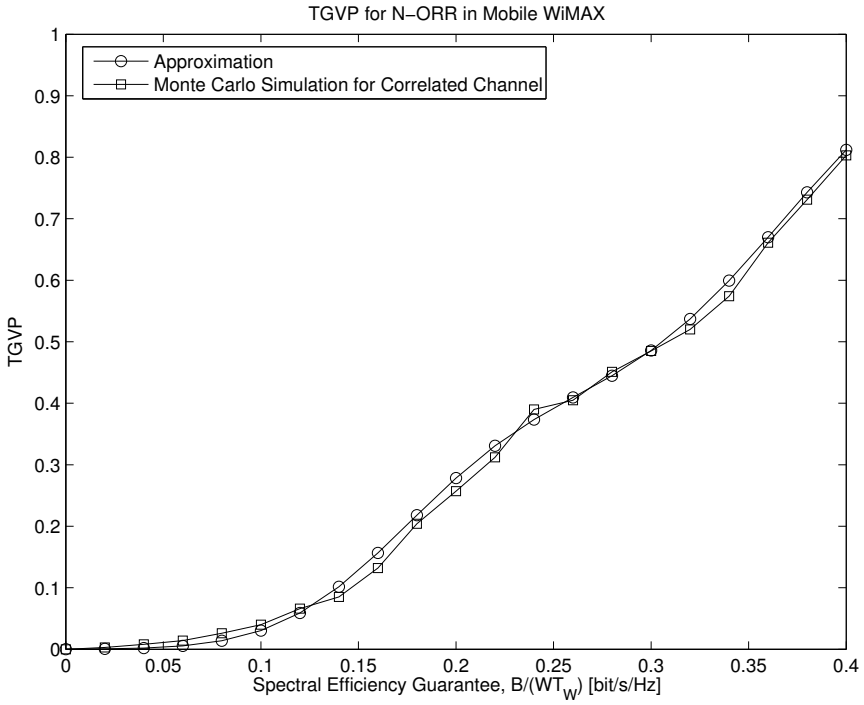

FIGURE E.3: Approximated TGVP vs. Monte Carlo simulated TGVP for a user with $\overline{\gamma}_i = 5$. There are 9 other users in the cell and N-ORR scheduling is used. Plotted for the Mobile WiMAX time-slot length of 5 ms and a time-window of $T_W = 80$ ms, corresponding to $K = 16$ time-slots. Each value in the simulated graph is an average over 1000 Monte Carlo simulations.

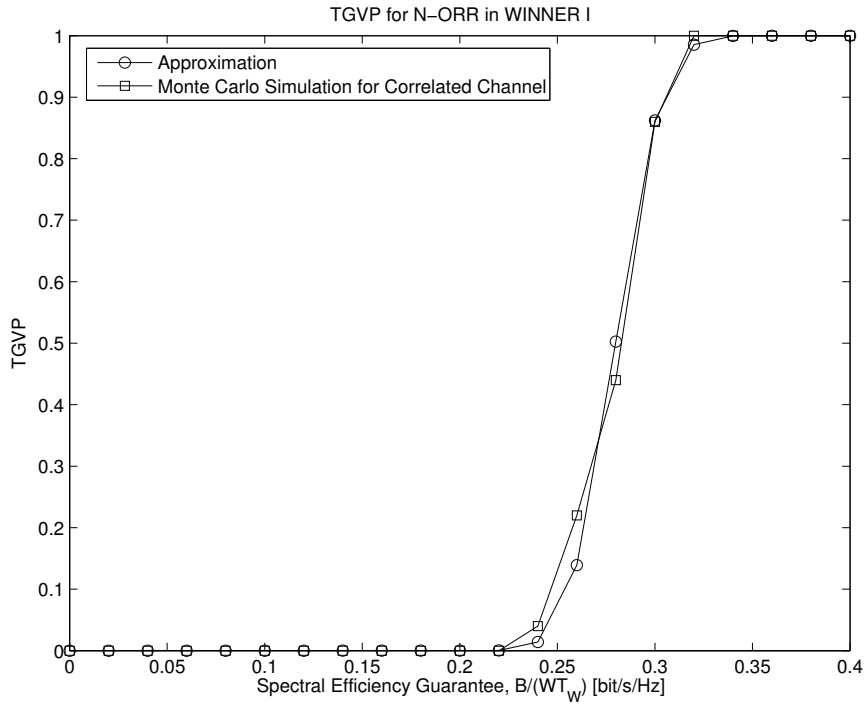

FIGURE E.4: Approximated TGVP vs. Monte Carlo simulated TGVP for a user with $\overline{\gamma}_i = 5$. There are 9 other users in the cell and N-ORR scheduling is used. Plotted for the WINNER I time-slot length of 0.34 ms and a time window of $T_W = 80$ ms, corresponding to $K = 235$ time-slots. Each value in the simulated graph is an average over 50 Monte Carlo simulations.

The tightness of the approximation is both influenced by K and T_{TS}. Since the CLT is used to obtain the formula for the approximative TGVPs, we therefore need to calculate the TGVP for a relatively large number of time-slots K to obtain a tight approximation. However, if we have shorter time-slots, we will also experience a higher correlation between the time-slots. Since we have assumed uncorrelated time-slots to obtain our TGVP approximation, we will therefore have a less tight approximation for short time-slots. For both Mobile WiMAX ($K = 16$ time-slots) and WINNER I ($K = 235$ time-slots) we see that our approximate results are too optimistic for TGVPs close to zero. However, we see that the TGVP approximation close to TGVP$= 0$ is slightly better for WINNER I and we can there-

fore conclude that the number of time-slots K within the time-window T_W will affect the tightness of TGVP-approximation more than the fact that the shorter time-slots are more correlated.

For long values of T_W, the value of K is higher and the correlation over the time-window is smaller. We can therefore conclude that long time-windows will lead to more tight TGVP approximations.

5 Conclusion

In this letter we have developed a general approximation for the TGVP which can be obtained in a time-slotted wireless network with any scheduling policy with (i) a given set of system parameters, (ii) known first two moments of the bits transmitted to or from the scheduled user in a time-slot, and (iii) a given distribution of the number of time-slots allocated to a user within a time-window. We have evaluated closed-form expressions for the corresponding TGVP approximations for four well-known scheduling algorithms, namely Round Robin, Maximum CNR Scheduling, Normalized CNR Scheduling and Normalized Opportunistic Round Robin. Our TGVP approximations were also compared to Monte Carlo simulations for users with correlated channels. From our numerical investigations, it can be concluded that correlated time-slots have a small effect on the tightness of the approximations. It can also be concluded that the TGVP approximations are tighter for relatively long time-windows T_W.

TABLE E.1: Closed-Form Expressions for $p_M(k|i)$, $\mathrm{E}[b_{i,j}]$ and $\mathrm{E}[b_{i,j}^2]$. Derivations are found in Appendix 4.

RR	$p_M(k\|i) = \begin{cases} \frac{(k+1)N-K}{N}, & k = \lfloor \frac{K}{N} \rfloor \\ \frac{K-(k-1)N}{N}, & k = \lceil \frac{K}{N} \rceil \\ 0, & \text{otherwise} \end{cases}$
	$\mathrm{E}[b_{i,j}] = \frac{WT_{\mathrm{TS}}}{\ln 2} e^{1/\overline{\gamma}_i} E_1\left(\frac{1}{\overline{\gamma}_i}\right)$
	$\mathrm{E}[b_{i,j}^2] = \frac{(WT_{\mathrm{TS}})^2}{\overline{\gamma}_i (\ln 2)^2} \Psi\left(\frac{1}{\overline{\gamma}_i}\right)$
MCS	$p_M(k\|i) = \binom{K}{k} p_i^k (1-p_i)^{K-k}$, $p_i = \frac{1}{\overline{\gamma}_i} \sum_{\tau \in T_i^N} \mathrm{sign}(\tau) \frac{1}{\frac{1}{\overline{\gamma}_i} + \|\tau\|}$
	$\mathrm{E}[b_{i,j}] = \frac{WT_{\mathrm{TS}}}{p_i \overline{\gamma}_i \ln 2} \sum_{\tau \in T_i^N} \mathrm{sign}(\tau) \frac{e^{\left(\frac{1}{\overline{\gamma}_i} + \|\tau\|\right)}}{\frac{1}{\overline{\gamma}_i} + \|\tau\|} E_1\left(\frac{1}{\overline{\gamma}_i} + \|\tau\|\right)$
	$\mathrm{E}[b_{i,j}^2] = \frac{(WT_{\mathrm{TS}})^2}{p_i \overline{\gamma}_i (\ln 2)^2} \sum_{\tau \in T_i^N} \mathrm{sign}(\tau) \Psi\left(\frac{1}{\overline{\gamma}_i} + \|\tau\|\right)$
NCS	$p_M(k\|i) = \binom{K}{k} \frac{1}{N}^k (1 - \frac{1}{N})^{K-k}$
	$\mathrm{E}[b_{i,j}] = \frac{NWT_{\mathrm{TS}}}{\ln 2} \sum_{j=0}^{N-1} \binom{N-1}{j} \frac{(-1)^j}{1+j} e^{\frac{1+j}{\overline{\gamma}_i}} E_1\left(\frac{1+j}{\overline{\gamma}_i}\right)$
	$\mathrm{E}[b_{i,j}^2] = \frac{N(WT_{\mathrm{TS}})^2}{\overline{\gamma}_i (\ln 2)^2} \sum_{j=0}^{N-1} \binom{N-1}{j} (-1)^j \Psi\left(\frac{1+j}{\overline{\gamma}_i}\right)$
N-ORR	$p_M(k) = \begin{cases} \frac{(k+1)N-K}{N}, & k = \lfloor \frac{K}{N} \rfloor \\ \frac{K-(k-1)N}{N}, & k = \lceil \frac{K}{N} \rceil \\ 0, & \text{otherwise} \end{cases}$
	$\mathrm{E}[b_{i,j}] = \begin{cases} \frac{WT_{\mathrm{TS}}}{N} \sum_{n=1}^{N} A_i(n), & k = \lfloor \frac{K}{N} \rfloor \\ WT_{\mathrm{TS}} \left(\frac{(k-1)\sum_{n=1}^{N} A_i(n)}{kN} + \frac{\sum_{n=kN-K+1}^{N} A_i(n)}{k(K-(k-1)N)} \right), & k = \lceil \frac{K}{N} \rceil \end{cases}$
	$\mathrm{E}[b_{i,j}^2] = \begin{cases} \frac{WT_{\mathrm{TS}}}{N} \sum_{n=1}^{N} B_i(n), & k = \lfloor \frac{K}{N} \rfloor \\ WT_{\mathrm{TS}} \left(\frac{(k-1)\sum_{n=1}^{N} B_i(n)}{kN} + \frac{\sum_{n=kN-K+1}^{N} B_i(n)}{k(K-(k-1)N)} \right), & k = \lceil \frac{K}{N} \rceil \end{cases}$

References

[1] WiMAX Forum, "Mobile WiMAX – Part II: A Comparative Analysis." `http://www.wimaxforum.org/news/downloads/Mobile_WiMAX_Part2_Comparative_Analysis.pdf`, May 2006.

[2] R. Knopp and P. A. Humblet, "Information capacity and power control in single cell multiuser communications," in *Proc. IEEE Int. Conf. on Communications (ICC'95)*, (Seattle, WA, USA), pp. 331–335, June 1995.

[3] P. Viswanath, D. N. C. Tse, and R. Laroia, "Opportunistic beamforming using dumb antennas," *IEEE Trans. Inform. Theory*, vol. 48, pp. 1277–1294, June 2002.

[4] Y. Liu and E. Knightly, "Opportunistic fair scheduling over multiple wireless channels," in *Proc. IEEE Joint Conference of the Computer and Communications Societies (INFOCOM'03)*, vol. 2, (San Francisco, CA, USA), pp. 1106–1115, Mar. 2003.

[5] V. Hassel, M. R. Hanssen, and G. E. Øien, "Spectral efficiency and fairness for opportunistic round robin scheduling." Presented at the *IEEE Int. Conf. Comm. (ICC'06)*, (Istanbul, Turkey), June 2006.

[6] V. Hassel, G. E. Øien, and D. Gesbert, "Throughput guarantees for wireless networks with opportunistic scheduling." Presented at *IEEE Global Communications Conf. (GLOBECOM'06)*, (San Fancisco, CA, USA), Nov.-Dec. 2006.

[7] V. Hassel, G. E. Øien, and D. Gesbert, "Throughput guarantees for opportunistic scheduling: A comparative study." Presented at *International Telecommunications Symposium (ITS'06)*, (Fortaleza, Brazil), Sept. 2006.

[8] D. Ferrari, "Client requirements for real-time communication services," *IEEE Communications Mag.*, vol. 28, pp. 65–72, Nov. 1990.

[9] N. C. Beaulieu, "An infinite series for the computation of the complementary probability distribution function of a sum of independent random variables and its application to the sum of Rayleigh random variables," *IEEE Trans. Commun.*, vol. 38, pp. 1463–1474, Sept. 1990.

[10] E. Kreyszig, *Advanced Engineering Mathematics*.
New York: John Wiley & Sons, Inc., 7th ed., 1993.

[11] M. Sternad, T. Svenson, and G. Klang, "WINNER MAC for cellular transmission," in *Proc. of IST Mobile Summit*, (Mykonos, Greece), June 2006.

[12] B. Wang, K. I. Pedersen, T. E. Kolding, and P. E. Mogensen, "Performance of VoIP on HSDPA," in *Proc. IEEE Vehicular Technology Conference (VTC'05-spring)*, (Stockholm, Sweden), May-June 2005.

[13] F. Fluckiger, *Understanding Networked Multimedia: Applications and Technology*.
Englewood Cliffs, New Jersey: Prentice-Hall, Inc., 1995.

[14] L. Yang and M.-S. Alouini, "Performance analysis of multiuser selection diversity," in *Proc. IEEE Int. Conf. on Communications (ICC'04)*, (Paris, France), pp. 3066–3070, June 2004.

[15] S. S. Kulkarni and C. Rosenberg, "Opportunistic scheduling policies for wireless systems with short term fairness constraints," in *Proc. IEEE Global Communications Conf. (GLOBECOM'03)*, (San Francisco, CA, USA), pp. 533–537, Dec. 2003.

[16] H. J. Bang, "Advanced scheduling techniques for wireless data networks." Master Thesis, Department of Physics, University of Oslo, Feb. 2005.

Paper F

Scheduling Algorithms for Increased Throughput Guarantees in Wireless Networks

Vegard Hassel, Sébastien de la Kethulle de Ryhove, and Geir E. Øien

Submitted to
Workshop on Resource Allocation in Wireless Networks (RAWNET'07),
Limassol, Cyprus, April 2007.

Abstract

For cellular wireless networks carrying real-time traffic, it is in the interest of both network operators and customers that throughput guarantees can be offered. In this paper, we formulate an optimization problem which aims at maximizing the throughput that can be guaranteed to the mobile users. By building on results obtained by Borst and Whiting and by assuming that the distributions of the users' carrier-to-noise ratios are known, we find the solution to this problem for users with different channel quality distributions, both for the scenario where all the users have the same throughput guarantee, and for the scenario where all the users have different throughput guarantees. Based on these solutions, we propose an adaptive scheduling algorithm that performs significantly better than other well-known scheduling algorithms.

1 Introduction

In modern wireless networks, *opportunistic multiuser scheduling* has been implemented in order to obtain a more efficient utilization of the scarcely available radio spectrum. For wireless cellular standards such as 1xEVDO, HSDPA and Mobile WiMAX [1], the scheduling algorithms are often not specified in the standardization documents. The scheduling algorithms implemented might therefore vary from vendor to vendor. Selecting the most efficient scheduling algorithms will be critical for having the most efficient utilization of a wireless network; consequently, the vendors that implement the most suited scheduling algorithms will have a competitive advantage.

Opportunistic multiuser scheduling will give higher throughput in a wireless cell than non-opportunistic algorithms like Round Robin because priority is given to the users with the most favorable channel conditions [2, 3]. However, always selecting the users with the best channel quality may lead to starvation of other users. Consequently, the quality-of-service (QoS) demands of the users also have to be taken into account when designing practical wireless scheduling algorithms. A common approach to obtain higher QoS in the network is to have a fairer resource allocation among the users [4, 5]. One widely adopted fair scheduling policy is the Proportional Fair Scheduling (PFS) algorithm [6]. When there are many users in a cell, this algorithm ensures both that the users are scheduled close to their own peak carrier-to-noise ratio (CNR) and that they have the same probability of being scheduled in a randomly picked time-slot [7].

With real-time traffic transmitted over wireless networks, the need for more exact QoS measures is in the interests of both network operators and customers. The customers want to know what they have bought and the operators would rather not give away more network capacity to the customers than they have paid for. A QoS measure that is well suited to quantify QoS guarantees exactly is a *throughput guarantee*, i.e., how many bits a user is guaranteed to transmit or receive within a time-window. Throughput guarantees can in principle be either *hard* or *deterministic*, and *soft* or *statistical*. Hard throughput guarantees promise with unit probability that a guarantee will be fulfilled, while the corresponding soft throughput guarantees promise with a lower than unity – but preferably high – probability that the specified throughput guarantee will be fulfilled. For telecommunications networks in general, and for wireless networks in particular, soft throughput guarantees are more suitable for specifying QoS than hard throughput guarantees. This is because such networks often have a varying number of users and varying loads from the applications of these users. For wireless networks, the varying quality of the radio channel will further

add uncertainty to the size of the throughput that can be guaranteed during short time-spans.

In [8] Andrews et al. propose scheduling algorithms that aim at fulfilling throughput guarantees by giving different priorities to the users depending on how far they are from their maximum and minimum throughput guarantees. One of the problems with this algorithm is that it takes action only when a throughput guarantee has been violated. Andrews et al. have therefore shown in [8] how time parameters of their algorithm can be set shorter than the actual time window of interest to alleviate this issue. In this paper we propose an alternative scheduling algorithm that tries to fulfill the throughput guarantees before they are violated.

Borst and Whiting have elegantly proved that a certain scheduling policy provides the highest throughput guarantee for wireless networks [9]. However, they briefly argue that the rate distributions of the users are unknown and they have therefore not shown how this optimal scheduling policy can be found for users with differently distributed CNRs. They have also not designed algorithms that will give the lowest short-term *throughput guarantee violation probability* (TGVP), which we define as the probability of not fulfilling a throughput guarantee within a specified time-window, averaged over all the users in the system. In the next section, we argue that for many scenarios the CNR distributions of the users can in fact be estimated, and that we hence can use these distributions to develop efficient scheduling algorithms for providing short-term throughput guarantees.

Our proposed scheduling algorithms do not only aim at fulfilling the throughput guarantees that are promised to the mobile users in a wireless network (see Section 5), but our analysis can also be used to estimate the expected TGVP of all the users if a new user is admitted into the system. Such real-time TGVP estimates can be useful when performing admission control.

It should be noted that our analysis involves several idealistic assumptions (see Section 2). For example, we assume that the CNR can be estimated perfectly and fed back with infinite precision and no delay, that ideal adaptive modulation and coding can be performed, that the CNR distributions of the users can be estimated perfectly, and that the population of backlogged users is constant over the time-window the throughput guarantees are calculated. How realistic these assumptions are for real-life networks is a subject for further research.

Contributions. We formulate an optimization problem aimed at finding an optimal scheduling algorithm that obtains maximum throughput guarantees in a wireless network. By building on the results in [9] and by assuming that the distributions of the users' CNRs are known, we show

how the solution to this optimization problem can be obtained numerically both when the throughput guarantees are (i) the same and (ii) different for all the mobile users. We also propose an adaptive algorithm that improves the performance of the optimal algorithm for short time-windows.

Organization. The rest of this paper is organized as follows. In Section 2 we present the system model, and in Section 3 we formulate the optimization problem for obtaining the highest possible throughput guarantee over a time-window. In Section 4 we show how the solution to this problem can be found when all the users have the same throughput guarantees. The corresponding solution for heterogeneous throughput guarantees is discussed in Section 5, while we describe the novel adaptive algorithm in Section 6. In Section 7 we discuss some practical considerations before presenting our numerical results in Section 8. We list our conclusions in Section 9.

2 System Model

We consider a single base station that serves N backlogged users using time-division multiplexing (TDM). The analysis conducted in this paper is valid both for the uplink and the downlink; in either case we assume that the total available bandwidth for the users is W [Hz] and that the users have constant transmit power. Each user estimates his own CNR perfectly, and before performing downlink scheduling the base station is assumed to receive these measurements from all the users. The base station also performs uplink scheduling based on perfect channel estimates, and for each time-slot, the base station takes a scheduling decision and distributes this decision to the selected user before uplink transmission starts.

It is assumed that the communication channel between the base station and every one of the users can be modelled by a flat, block-fading channel subject to additive white Gaussian noise (AWGN); and moreover that the communication channels corresponding to the different users fade independently. The block duration equals one time-slot and is denoted T_{TS} [seconds]. We also assume that the CNR values corresponding to different time-slots are correlated. The correlation model used in our simulations will be described in detail in Section 7.

The average CNR of user i is denoted by $\overline{\gamma}_i$. Without loss of generality, we assume that the user indices are assigned in a manner which is such that user 1 has the lowest average CNR, user 2 has the second lowest average CNR, and so on, down to user N, which has the highest average CNR. Assuming constant average CNR values for the time-window over

which the throughput guarantees are calculated can be realistic for a real-life wireless network. This is because the average CNR of the users' CNR distributions normally changes on a time-scale of several seconds while the throughput-guarantees are often calculated over time-windows of less than one hundred milliseconds.

We also assume that the probability distributions of the CNRs of each of the users are perfectly known (a known joint CNR distribution is *not* required). In modern cellular standards like 1xEVDO, HSDPA and Mobile WiMAX [1], much of the information needed for obtaining precise probability distribution estimates is already available. To conduct adaptive coding and modulation, modern cellular networks have precise, real-time CNR estimates of the users. These channel quality estimates can therefore be used to obtain estimates of the probability distributions of the CNRs of each one of the users. Such probability distribution estimates can be obtained from some hundred CNR estimates by using e.g. order statistic filter banks [10]. To further improve the estimates of the probability distributions, we can adapt the estimation techniques to the types of terrain the users operate in and to the speed of the users. For example, for a channel with many reflectors, with no line-of-sight (LOS) component, and with a relatively high speed of the users, a Rayleigh channel model will give a good estimate of the distribution of the channel gain. When we have a LOS component, a Rice channel can be assumed.

Another important assumption is that the population of backlogged users is constant and equal to N. According to [9] this assumption is realistic since the separation of time-scales makes the population of backlogged users nearly static, i.e., the population of backlogged users changes much slower than the time-window over which the throughput guarantees are calculated.

3 The Optimization Problem

The goal of this section is to formulate an optimization problem aimed at obtaining the maximal throughput guarantee B [bits], which can be achieved within a time-window of T_W [seconds]. A similar optimization problem has been formulated in [9]. Here, we assume that the same throughput guarantee is promised to all the users, i.e.,

$$T_i \bar{R}_i = B, \tag{F.1}$$

for all $i = 1, ..., N$, where T_i [seconds] is the accumulated time allocated to user i over the time-window and $\bar{R}_i$ [bits/s] is the average rate for user i

when he is transmitting or receiving. By virtue of the TDM assumption, the sum of the T_is satisfies

$$\sum_{i=1}^{N} T_i = T_W. \tag{F.2}$$

Inserting (F.1) into (F.2), the following yields:

$$B = \frac{T_W}{\sum_{i=1}^{N} \frac{1}{\bar{R}_i}}. \tag{F.3}$$

A slight modification of (F.1) also gives

$$p(i) T_W \bar{R}_i = B, \tag{F.4}$$

where $p(i)$ is the access probability for user i within the duration of the time-window T_W. Setting (F.4) equal to (F.3), we obtain

$$p(i) = \frac{1}{\bar{R}_i \sum_{j=1}^{N} \frac{1}{\bar{R}_j}}, \tag{F.5}$$

which links the access probability to the average rates. Assuming that T_i is long enough and contains enough time-slots for the channel to reveal its ergodic properties, and assuming that the Shannon capacity can be achieved, the average rate $\bar{R}_i$ for user i when he is transmitting or receiving, can be written

$$\bar{R}_i = W \int_0^{\infty} \log_2(1+\gamma) p_{\gamma *}(\gamma | i) d\gamma, \tag{F.6}$$

where $p_{\gamma *}(\gamma | i)$ is the probability density function (PDF) of the CNR of user i when this user is scheduled.

From the equations above, our objective is to find a *scheduling policy* that gives the maximum B that can be promised to all the users over the time-window T_W, meaning that (F.3) has to be maximized subject to the constraints (F.6), for $i = 1, \ldots, N$. We show in the next section how to obtain this optimal scheduling policy.

4 Solution to the Optimization Problem

It was shown in [9] that the following scheduling algorithm gives the solution to the optimization problem described in the previous section:

$$i^*(t_k) = \underset{1 \le i \le N}{\operatorname{argmax}} \left(\frac{r_i(t_k)}{\alpha_i} \right), \tag{F.7}$$

where $i^*(t_k)$ is the index of the user that is going to be scheduled in time-slot k, $r_i(t_k)$ is the instantaneous rate of user i in time-slot k, and α_i is a constant. However, in [9] it is not shown how the optimal α_is can be found. If we assume that the PDFs of the users' channel gains are known and that we have an ideal link adaptation protocol and block-fading, we can use this result to obtain a solution to the optimization problem in the previous section. To obtain this solution, we define the random variable $S_i \triangleq \frac{R_i}{\alpha_i}$, where R_i is the random variable describing the rate of user i. S_i is the scheduling metric of the algorithm, i.e. the metric that decides which user is going to be scheduled. For flat, block-fading channels, the maximal value of the metric S_i for user i within a time-slot (block) with CNR γ can be expressed as

$$S_i(\gamma) = \frac{W \log_2(1+\gamma)}{\alpha_i}. \tag{F.8}$$

In real-life systems we can come close to this maximum value of S_i by using efficient link adaptation and capacity-achieving codes. Assuming that the users have Rayleigh faded channel gains, and denoting by $p_{\gamma_i}(\gamma)$ the PDF of the CNR of user i, the PDF for the normalized rate $S_i = s$ for user i can be written

$$p_{S_i}(s) = \left.\frac{p_{\gamma_i}(\gamma)}{\frac{dS_i(\gamma)}{d\gamma}}\right|_{\gamma=2^{\frac{s\cdot\alpha_i}{W}}-1} = \frac{\alpha_i \ln(2)}{W\overline{\gamma}_i} 2^{\frac{s\cdot\alpha_i}{W}} e^{-\frac{2^{\frac{s\cdot\alpha_i}{W}}-1}{\overline{\gamma}_i}}, \tag{F.9}$$

where $\overline{\gamma}_i$ is the average CNR of user i. The corresponding cumulative distribution function (CDF) can be expressed as

$$P_{S_i}(s) = \int_0^s p_{S_i}(x)dx = 1 - e^{-\frac{2^{\frac{s\cdot\alpha_i}{W}}-1}{\overline{\gamma}_i}}. \tag{F.10}$$

Note that in principle we could have used different CNR distributions (e.g. Rayleigh, Rice, Nakagami) for the different users. However, in this paper, we have for simplicity reasons assumed that all the users have Rayleigh distributed channel gains.

We can now express the access probability of user i as

$$p(i) = \int_0^\infty p_{S_i}(s) \prod_{\substack{j=1 \\ j\neq i}}^{N} P_{S_j}(s)\, ds. \tag{F.11}$$

Furthermore, the PDF of S_i when user i is scheduled can be found by using *Bayes' rule*:

$$p_{S_i}(s|i) = \frac{p_{S_i}(s)}{p(i)} \prod_{\substack{j=1 \\ j \neq i}}^{N} P_{S_j}(s). \tag{F.12}$$

We can also express the expected value of S_i conditioned on user i being scheduled, as

$$\mathrm{E}[S_i|i] = \frac{\mathrm{E}[R_i|i]}{\alpha_i} = \frac{\bar{R}_i}{\alpha_i} = \int_0^\infty s\, p_{S_i}(s|i)\, ds. \tag{F.13}$$

Combining (F.5), (F.11), and (F.13) we obtain $3N$ equations in $3N$ unknowns, and can thus find the values for the $p(i)$s, the $\bar{R}_i$s, and the α_is. However, since it is possible to express the $p(i)$s and the $\bar{R}_i$s as functions of the α_is, and since multiplying all the α_is with an arbitrary constant does not change the behavior of the algorithm, we can set e.g. $\alpha_1 = 1$, finding a solution to our optimization problem boils down to solving a set of $N-1$ independent equations in $N-1$ unknowns. A solution can be found by using numerical integration together with an algorithm for solving sets of nonlinear equations. This can for example be achieved in *Matlab* by using the functions *quad* and *fsolve*. It should be noted that it has not been proved that the solution to this set of equations is unique.

TABLE F.1: Example of parameters for 10 Rayleigh-distributed users.

i	$\overline{\gamma}_i$ [dB]	$\overline{\gamma}_i$	$p(i)$	$\bar{R}_i$ [bit/s]	α_i
1	5.0000	3.1623	0.180356	3.146640	2.751868
2	9.7712	9.4868	0.120250	4.719458	4.510686
3	11.9897	15.8114	0.104241	5.444237	5.354891
4	13.4510	22.1359	0.095904	5.917567	5.916338
5	14.5424	28.4605	0.090533	6.268622	6.337987
6	15.4139	34.7851	0.086660	6.548775	6.675953
7	16.1394	41.1096	0.083694	6.780829	6.958115
8	16.7609	47.4342	0.081318	6.978923	7.200381
9	17.3045	53.7587	0.079353	7.151803	7.412701
10	17.7875	60.0833	0.077691	7.304769	7.601701

In Table I we give an example of the parameters of 10 users with a total average CNR of 15 dB (averaged over all the users). From (F.4), we can see

that $B_{opt}/(WT_W) = p(i)\bar{R}_i/W$ for the optimized values of $p(i)$ and $\bar{R}_i$. It is easily seen by using the values in Table I to calculate the product $p(i)\bar{R}_i$ for $1 \leq i \leq 10$ that $B_{opt}/(WT_W) = 0.5675$ bits/s/Hz for all the users for this particular example.

Since this scheduling algorithm maximizes B, we would expect that this algorithm will yield higher values of B than any of the other classical scheduling algorithms. However, one should remember that it is implicitly assumed in (F.1) that the average rate of the users equals their expected throughput. This will only be true when the time-window T_W can be considered infinitely long and contains infinitely many time-slots. The solution is consequently suboptimal for short time-windows containing only a small amount of time-slots. In Section 6 we therefore propose an adaptive algorithm that shows good performance for short T_Ws with few time-slots.

5 Optimization for Heterogeneous Throughput Guarantees

When the throughput guarantees are different from user to user, we can again use the scheduling policy corresponding to (F.7), but with a different set of α_is to obtain the optimal bit allocation. By using B_i [bits] to denote the throughput guarantee for user i during the time-window T_W and by rewriting (F.13), we obtain

$$\frac{B_i}{T_W} = \bar{R}_i p(i) = \alpha_i \int_0^\infty s\, p_{S_i}(s) \prod_{\substack{j=1 \\ j\neq i}}^{N} P_{S_j}(s)\, ds. \tag{F.14}$$

We can now fix the throughput guarantees B_i of up to $N-1$ users and maximize the remaining throughput guarantees by solving the set of $3N$ equations resulting from (F.5), (F.11), and (F.14). To be able to solve this optimization problem, we can for example additionally constrain the users with non-fixed B_is to have equal throughput guarantees. It is also important to note that setting fixed throughput guarantees that are too high will yield an optimisation problem with no solution – meaning that such throughput guarantees are not achievable by the system. One hence should only set fixed throughput guarantees that are achievable by the system.

6 Adapting Weights to Increase Short-Term Performance

The values of α_i found in the previous sections aim at providing throughput guarantees within *any* time-window T_W. This means that these parameters are optimized in a manner which is such that the throughput guarantees should be fulfilled independently of the time instants at which T_W starts or ends. In this section we instead develop an algorithm that will only aim at fulfilling the throughput guarantees for a scenario where the placement of the window T_W is fixed. That is, for every new time-window, the algorithm starts over again and tries to achieve the throughput guarantees. This means that the throughput guarantees cannot be promised within time-windows with a different duration or a different placement than that used by the algorithm. The consequence of this approach is that we may have to adjust the time-window T_W to the bit-streams from different speech and video codecs.

As already mentioned, the scheduling algorithms obtained in the previous sections are only efficient when the throughput guarantees are promised over a long time-window T_W containing many time-slots. To fulfill throughput guarantees for shorter time-windows with fewer time-slots, it is useful to *adapt the values of the parameters α_i to the actual resource allocation that has already been done* within the finite time-window T_W. This adaptation can be optimally done during each time slot by using the approach of the previous section with B_i/T_W replaced by $B_i'/T_W' = (B_i - B_{ik})/(T_W - T_k)$ in (F.14), where B_{ik} is the number of bits assigned to user i after k time-slots within the time-window T_W, and $T_k = kT_{\mathrm{TS}}$. The adaptation of the parameters α_i should in many cases be performed in time intervals of less than a millisecond. Since it can be difficult to conduct the optimal optimization described above in such a short time, we propose the following simple adaptive scheduling algorithm as an alternative:

$$i^*(t_k) = \underset{1\le i\le N}{\operatorname{argmax}} \left(\rho_i(t_{k-1}) \frac{r_i(t_k)}{\alpha_i} \right), \tag{F.15}$$

where $\rho_i(t_k)$ is the ratio

$$\rho_i(t_k) = \frac{\max(0, B_i - B_{ik})}{T_W - T_k} \frac{T_W}{B_i}. \tag{F.16}$$

The rationale behind this scheduling algorithm is as follows: The value of $\rho_i(t_k)$ expresses the normalized share of the throughput guarantee that is to be fulfilled in the remaining $K - k$ time-slots of the time-window T_W. If the

rate guarantee is already fulfilled, the value of $\rho_i(t_k)$ is zero, which means that the user in question is not selected in the remaining $K - k$ time-slots. If a user has been allocated exactly $\frac{B_i T_k}{T_W}$ bits after k time-slots, the value of $\rho_i(t_k)$ will be unity, which means that this user will be scheduled with the same weights as for the non-adaptive policy. For the case where the number of allocated bits after k time slots is lower than $\frac{B_i T_k}{T_W}$ bits, the value of $\rho_i(t_k)$ will be above unity, which means that the user is given higher priority compared to the non-adaptive optimal scheduling policy. Likewise, a user is given lower priority if he has been allocated more than $\frac{B_i T_k}{T_W}$ bits after k time-slots. The priority is determined by the urgency of fulfilling the throughput guarantee within the remainder of the time-window.

7 Practical Considerations

7.1 Real-Life Values of $T_{\mathbf{TS}}$ and T_W

For the wireless standards 1xEVDO, HSDPA and Mobile WiMAX, the time-slot length for the downlink is respectively 1.67, 2, and 5 ms [1]. The European IST research project WINNER I has suggested a time-slot duration of 0.34 ms for a future wireless system [11]. According to [12], the maximum one-way delay over a wireless HSDPA link should lie between 80 and 150 ms for voice over IP (VoIP) conversations to achieve good speech quality. If we assume that $T_W = 80$ ms, T_W contains 235, 48, 40, and 16 time-slots for WINNER I, 1xEVDO, HSDPA and Mobile WiMAX, respectively.

7.2 Real-Life Values of the B_is

Different classes of traffic will need different values for B_i. For a one-way telephony speech connection, B_i/T_W can vary between 5 and 64 kbit/s [13]. The corresponding B_i/T_W for a videoconferencing connection will vary between 64 and 500 kbit/s. It should be remembered that these are raw throughput guarantees and that protocol overhead of typically 4 percent has to be added. For a real-life network we can assume that the B_is correspond to the sum of all the throughput guarantees promised to the different real-time sessions of a user. Hence, for each new videoconferencing or speech connection, the network has to update the B_is and do the optimization of the scheduling algorithm over again.

7.3 What if the CNR Distributions of the Users Change?

The optimal algorithm is based on the assumption that the CNR distributions of the users are known. If the average CNR of one or more users change or the CNR distribution of one or more users change, e.g., from Rayleigh to Rice, the whole optimization problem has to be solved again to obtain new values for the α_is, which is a feasible task. It should be noted that the adaptive factor $\rho_i(t_k)$ is independent of the CNR distributions.

7.4 The Effects of Correlated Time-Slots

When the CNR of a user is assumed to be correlated from time-slot to time-slot, one user can be allocated many consecutive time-slots. It is therefore more difficult to fulfill throughput guarantees for all the users in a system that has strongly temporally correlated channels. The temporal correlation of the channel is both dependent on the speed v of the users and on the carrier frequency f_c of the channel. For the simulations in the next section we assume Jakes' correlation model with $f_c = 1$ GHz and a user speed of 30 m/s. The channel gain can in this case be modeled as a sum of sinusoids correlated according to $f_D T_{\text{TS}}$, where $f_D = \frac{v f_c}{c}$ is the Doppler frequency shift and c is the speed of light [14].

8 Numerical Results

Figs. F.1, F.2, and F.3 show the theoretical TGVP performance in networks that are respectively based on WINNER I, HSDPA, and Mobile WiMAX. For these plots we have assumed that only one user can be scheduled in a time-slot. We focus on the TGVP (see Section 1) because a throughput guarantee in most cases cannot be given with absolute certainty. The guaranteed number of bits B within the time-window T_W should however be promised to the users with high probability. This means that the TGVP performance of the algorithms close to TGVP$= 0$ is the most interesting. We have considered the case where all the users are promised identical throughput guarantees B/T_W, where $T_W = 80$ ms. Unfortunately, we have not included plots displaying the TGVP for a network where the users have heterogeneous throughput guarantees. Analyzing the TGVP for such a scenario for all the users in the network would require many plots and is therefore left out due to space limitations. The results are shown for 10 users having Rayleigh fading channels with average CNRs given in Table I, and with the correlation between the different time-slot CNRs being described by

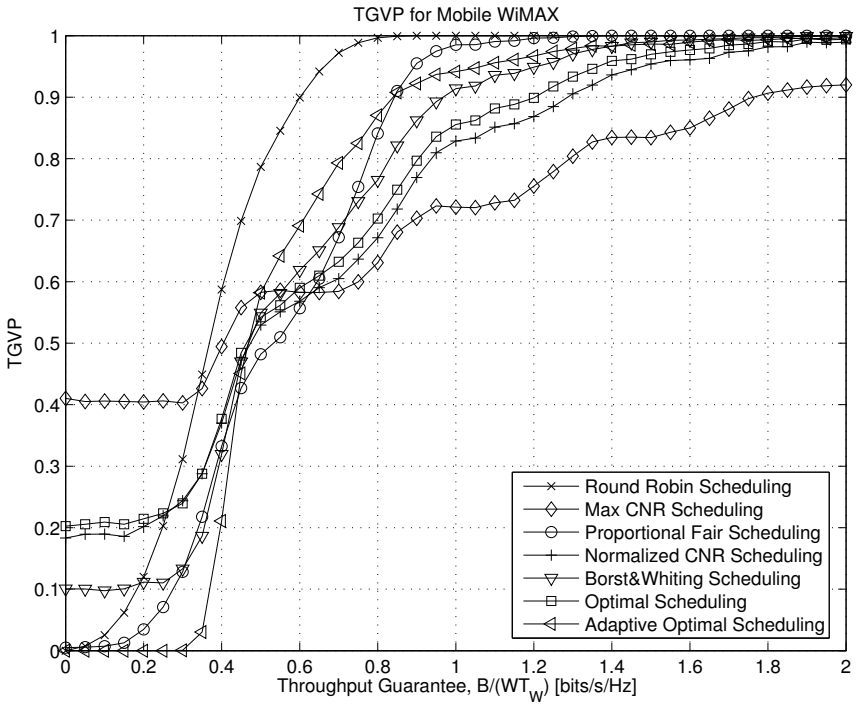

FIGURE F.1: Throughput guarantee violation probability for 10 users in a Mobile WiMAX network. Plotted for a time-window $T_W = 80$ ms that contains 16 time-slots. Each value in the plot is an average over 1000 Monte Carlo simulations.

Jakes' model (see Sec. VII D). It should be noted that this correlation will be stronger for short time-slots than for long time-slots.

We compare the new scheduling policies to five other algorithms, namely Round Robin Scheduling (RR), Maximum CNR Scheduling (MCS), Normalized CNR Scheduling (NCS), Proportional Fair Scheduling (PFS) and the *adaptive* scheduling algorithm proposed by Borst and Whiting in [9]. For the RR policy, the time-slots are allocated to the users in a sequential manner, i.e. totally non-opportunistically. The most opportunistic algorithm is the MCS policy because it always schedules the user with the highest CNR, and hence the highest rate. The NCS policy is a fairer policy because it schedules the users with the highest CNR-to-average-CNR ratio. A similar policy, the PFS algorithm, schedules the user with the highest instantaneous rate divided by a weighted sum of the rate allocated in the

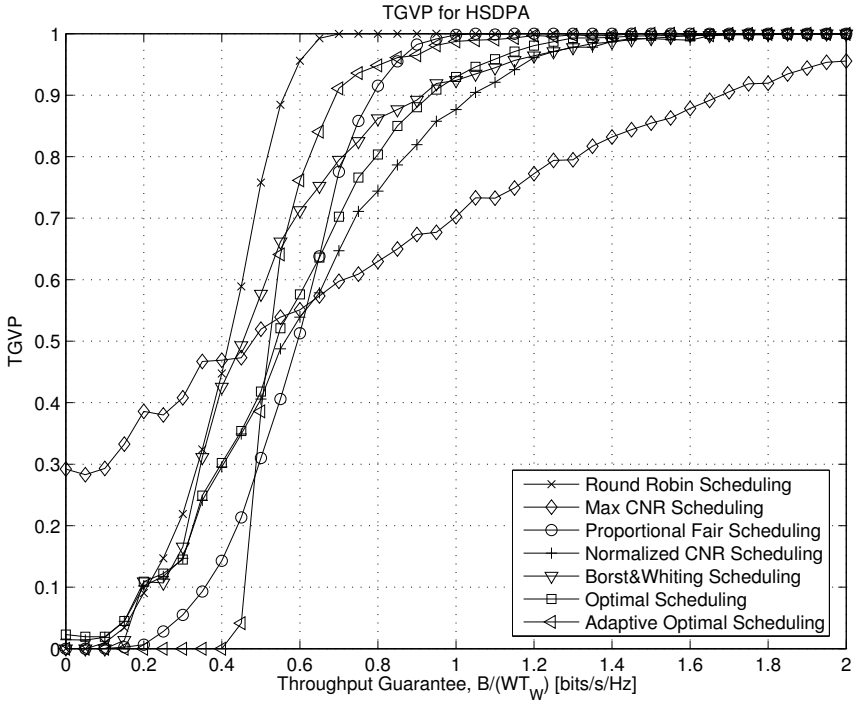

FIGURE F.2: Throughput guarantee violation probability for 10 users in a HSDPA network. Plotted for a time-window $T_W = 80$ ms that contains 40 time-slots. Each value in the plot is an average over 200 Monte Carlo simulations.

previous time-slots [3]. For our simulations, we have implemented the PFS algorithm as described in [3], with the time-constant $t_c = T_W$ and with the initial average rate for each user equal to the theoretical average rate for this user. The adaptive Borst and Whiting algorithm is implemented as described in [9, p. 575] with $\delta(k) = 0.5 * 0.9^k$, where k denotes the kth "reset". The "price updates" of this algorithm are done every $10 * n$th time-slot, where n denotes the nth "price update". To investigate the performance of the adaptive updating of the weights for this algorithm, we have used the optimal weights as initial weights.

Figs. F.1, F.2, and F.3 show the TGVP as a function of $B/(WT_W)$ for a time-window of respectively 16, 40, and 235 time-slots. We see that our novel adaptive algorithm performs better than all the other algorithms for all cases. It should also be noted that since the WINNER I system has many

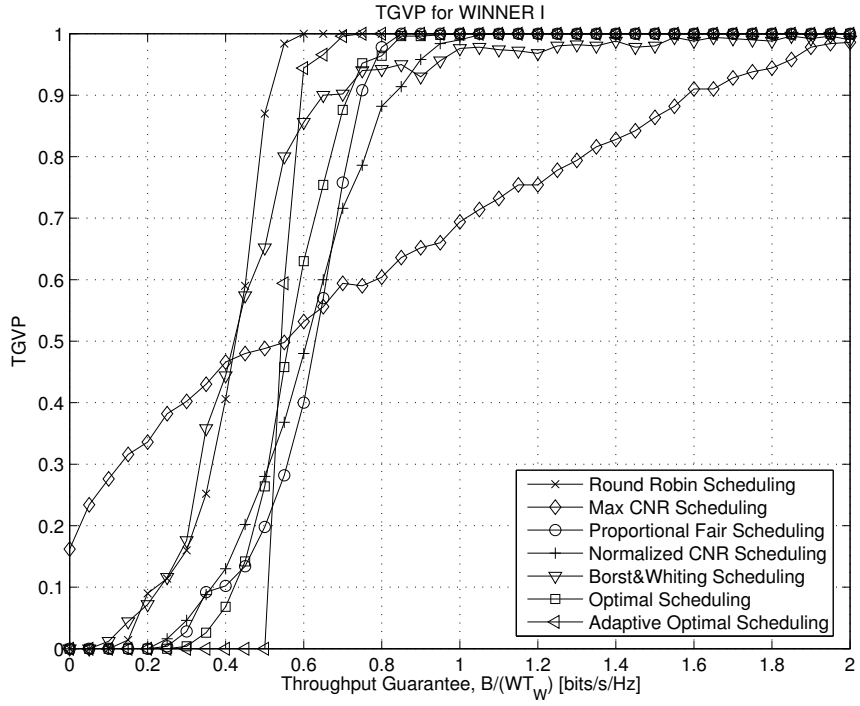

FIGURE F.3: Throughput guarantee violation probability for 10 users in a WINNER I network. Plotted for a time-window $T_W = 80$ ms that contains 235 time-slots. Each value in the plot is an average over 50 Monte Carlo simulations.

time-slots within the time-window of 80 ms, our adaptive algorithm obtains a throughput guarantee that is very close to the optimal throughput guarantee of 0.5675 bits/s/Hz for this system. It is also interesting to observe that the throughput guarantee that can be promised with close to unity probability with our adaptive algorithm is more than twice as large as for the PFS algorithm for all the three systems. Our non-adaptive optimal algorithm only performs better than all the other well-known algorithms for the case where the time-window contains 235 time-slots (WINNER I). The reason for this is that the non-adaptive algorithm is designed for long time-windows containing many time-slots.

9 Conclusion

For wireless networks carrying real-time traffic, providing throughput guarantees is interesting both from the customers' and the network providers' point of view. In order to have the most efficient utilization of the network, a scheduler in such a network should try to distribute the amount of bits that can be received or transmitted by each user according to given throughput guarantees. In this paper, we formulate an optimization problem which aims at finding the maximum number of bits that can be guaranteed to the users within a time-window for a given set of system parameters. By building on the results in [9] and by assuming that the distributions of the users' CNRs are known, we find an optimal scheduling algorithm, both for the case where the throughput guarantees are different from user to user and for the case where the users have the same throughput guarantees. To further improve the short-term performance of this algorithm, we propose an adaptive version of the optimal algorithm. Results from our simulations show that the proposed adaptive algorithm performs significantly better than any of the other well-known scheduling algorithms in networks based on Mobile WiMAX, HSDPA, and WINNER I. For systems that have many time-slots within the time-window, e.g. for WINNER I, our adaptive algorithm approaches the limit of what is theoretically attainable. The disadvantage is however that our adaptive scheduling algorithm can only obtain the simulated TGVP values when the placement of the time-window T_W is fixed.

References

[1] WiMAX Forum, "Mobile WiMAX – Part II: A Comparative Analysis." http://www.wimaxforum.org/news/downloads/Mobile_WiMAX_Part2_Comparative_Analysis.pdf, May 2006.

[2] R. Knopp and P. A. Humblet, "Information capacity and power control in single cell multiuser communications," in *Proc. IEEE Int. Conf. on Communications (ICC'95)*, (Seattle, WA, USA), pp. 331–335, June 1995.

[3] P. Viswanath, D. N. C. Tse, and R. Laroia, "Opportunistic beamforming using dumb antennas," *IEEE Trans. Inform. Theory*, vol. 48, pp. 1277–1294, June 2002.

[4] Y. Liu and E. Knightly, "Opportunistic fair scheduling over multiple wireless channels," in *Proc. IEEE Joint Conference of the Computer and Communications Societies (INFOCOM'03)*, vol. 2, (San Francisco, CA, USA), pp. 1106–1115, Mar. 2003.

[5] V. Hassel, M. R. Hanssen, and G. E. Øien, "Spectral efficiency and fairness for opportunistic round robin scheduling." Presented at the *IEEE Int. Conf. Comm. (ICC'06)*, (Istanbul, Turkey), June 2006.

[6] J. M. Holtzman, "Asymptotic analysis of proportional fair algorithm," in *Proc. IEEE Symp. on Personal, Indoor and Mobile Radio Communications (PIMRC'01)*, vol. 2, (San Diego, CA, USA), pp. F–33–F–37, Sept. 2001.

[7] D. Avidor, S. Mukherjee, J. Ling, and C. Papadias, "On some properties of the proportional fair scheduling policy," in *Proc. IEEE Symp. on Personal, Indoor and Mobile Radio Communications (PIMRC'04)*, vol. 2, (Barcelona, Spain), pp. 853–858, Sept. 2004.

[8] M. Andrews, L. Qian, and A. Stolyar, "Optimal utility based multi-user throughput allocation subject to throughput constraints," in

Proc. of the 24th annual joint conference of the IEEE Conference Computer and Communications Societies (INFOCOM'05), vol. 4, (Miami, FL, USA), pp. 2415–2424, Mar. 2005.

[9] S. C. Borst and P. Whiting, "Dynamic channel-sensitive scheduling algorithms for wireless data throughput optimization," *IEEE Trans. on Veh. Technol.*, vol. 52, pp. 569–586, May 2003.

[10] R. Suoranta, K.-P. Estola, S. Rantala, and H. Vaataja, "PDF estimation using order statistic filter bank," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Proc. (ICASSP'94)*, vol. 3, (Adelaide, Australia), pp. III–625–III–628, Apr. 1994.

[11] M. Sternad, T. Svenson, and G. Klang, "WINNER MAC for cellular transmission," in *Proc. IST Mobile Summit*, (Mykonos, Greece), June 2006.

[12] B. Wang, K. I. Pedersen, T. E. Kolding, and P. E. Mogensen, "Performance of VoIP on HSDPA," in *Proc. IEEE Vehicular Technology Conference (VTC'05-spring)*, (Stockholm, Sweden), May-June 2005.

[13] F. Fluckiger, *Understanding Networked Multimedia: Applications and Technology*.
Englewood Cliffs, New Jersey: Prentice-Hall, Inc., 1995.

[14] H. J. Bang, "Advanced scheduling techniques for wireless data networks." Master Thesis, Department of Physics, University of Oslo, Feb. 2005.

Part III

Appendices

Appendix 1

Derivation of the Last Term in (C.7)

For the Ranked Single-User Feedback protocol, we sequentially investigate if the users are within the successful interval based on their rank. We denote the number of users investigated *before* a user within the successful interval is found as X. The probability of finding one of the n users within the successful interval for the first user investigated is :

$$\Pr(X=0) = \frac{n}{N}. \tag{1.1}$$

If the search is not successful for the first user, the user with the second highest rank will have to be investigated. Now, we have already investigated one user. Consequently, the probability of finding a user within the successful interval is now given as:

$$\Pr(X=1) = \left(1-\frac{n}{N}\right)\frac{n}{N-1} = \frac{N-n}{N}\frac{n}{N-1}. \tag{1.2}$$

Correspondingly, the probability of finding a successful user for the third user is:

$$\Pr(X=2) = \left(1-\frac{n}{N}\right)\left(1-\frac{n}{N-1}\right)\frac{n}{N-2} = \frac{N-n}{N}\frac{N-n-1}{N-1}\frac{n}{N-2}. \tag{1.3}$$

Generalizing (1.1), (1.2) and (1.3), we obtain the expression for success for the $(k+1)$th user:

$$\begin{aligned}\Pr(X=k) &= \frac{N-n}{N}\frac{N-n-1}{N-1}\cdots\frac{N-n-k+1}{N-k+1}\frac{n}{N-k} \\ &= \frac{n(N-n)!(N-k-1)!}{N!(N-n-k)!}.\end{aligned} \tag{1.4}$$

The expected number of users investigated before success is now given as:

$$\mathrm{E}[X] = \sum_{k=0}^{N-n} k \frac{n(N-n)!(N-k-1)!}{N!(N-n-k)!}. \tag{1.5}$$

We know from Section 4 that the probability of having n users in interval l is given by:

$$p(l,n) = \binom{N}{n} (P_\gamma(\gamma_{\mathrm{th},l+1}) - P_\gamma(\gamma_{\mathrm{th},l}))^n (P_\gamma(\gamma_{\mathrm{th},l}))^{N-n}. \tag{1.6}$$

To obtain the time contribution from interval l, the expected number of users that are investigated before a user within the successful interval is found, are weighted by the probability of being in this interval:

$$\begin{aligned} &\sum_{k=0}^{N-n} k \frac{n(N-n)!(N-k-1)!}{N!(N-n-k)!} \binom{N}{n} (P_\gamma(\gamma_{\mathrm{th},l+1}) - P_\gamma(\gamma_{\mathrm{th},l}))^n (P_\gamma(\gamma_{\mathrm{th},l}))^{N-n} \\ = &\sum_{k=0}^{N-n} k \binom{N-k-1}{n-1} (P_\gamma(\gamma_{\mathrm{th},l+1}) - P_\gamma(\gamma_{\mathrm{th},l}))^n (P_\gamma(\gamma_{\mathrm{th},l}))^{N-n}. \end{aligned} \tag{1.7}$$

Summing this expression over all values of n gives the same expression as the last term in (C.7).

Appendix 2

Derivation of (C.20)

The CDF of the CNR of the user with the highest CNR can be found from *order statistics* [1]:

$$P_{\gamma^*}(\gamma) = P_{\gamma}^{N}(\gamma), \tag{2.1}$$

where $P_{\gamma}(\gamma)$ is the of the CNR for a single user. To find the MASSE for such a scenario, the PDF of the highest CNR between all the users has to be found. This PDF can be obtained by differentiating (2.1) with respect to γ [1, (5.85)]:

$$p_{\gamma^*}(\gamma) = N \cdot P_{\gamma}^{N-1}(\gamma) \cdot p_{\gamma}(\gamma), \tag{2.2}$$

where $p_{\gamma}(\gamma)$ is the PDF for a single user. Inserting the CDF and PDF for Rayleigh fading channels ($p_{\gamma}(\gamma) = (1/\overline{\gamma})e^{-\gamma/\overline{\gamma}}$) and using binomial expansion [2, (1.111)], we obtain:

$$p_{\gamma^*}(\gamma) = \frac{N}{\overline{\gamma}} \sum_{n=0}^{N-1} \binom{N-1}{n} (-1)^n e^{-(1+n)\gamma/\overline{\gamma}}. \tag{2.3}$$

Inserting (2.3) into the expression for the spectral efficiency ([Bit/Sec/Hz]) for optimal rate adaptation [3]:

$$\text{MASSE} = \int_0^{\infty} \log_2(1+\gamma) p_{\gamma^*}(\gamma)\, d\gamma, \tag{2.4}$$

we get the following expression for the MASSE:

$$\text{MASSE}_{\text{best}} = \frac{N}{\overline{\gamma} \ln 2} \sum_{n=0}^{N-1} \binom{N-1}{n} (-1)^n \int_0^{\infty} \ln(1+\gamma) e^{-(1+n)\gamma/\overline{\gamma}}\, d\gamma. \tag{2.5}$$

The expression for MASSE has to be weighted by the factor $(T_{\text{TS}} - \text{E}_l[T_{\text{G}}])/T_{\text{TS}}$. This factor is dependent on l, and consequently the integral

in the expression above has to be split into L parts before the weighting operation can take place. This leads to the following expression:

$$\begin{aligned} \mathrm{MASSE}_{\mathrm{best}} &= \frac{N}{\ln 2} \sum_{l=0}^{L-1} \frac{T_{\mathrm{TS}} - \mathrm{E}_l[T_{\mathrm{G}}]}{T_{\mathrm{TS}}} \sum_{n=0}^{N-1} \binom{N-1}{n} (-1)^n \\ &\times \int_{\gamma_{\mathrm{th},l}}^{\gamma_{\mathrm{th},l+1}} \ln(1+\gamma) e^{-(1+n)\gamma/\overline{\gamma}}\, d\gamma, \end{aligned} \tag{2.6}$$

To solve this integral we can use *integration by parts*:

$$\int_{\gamma=a}^{\gamma=b} u dv = \lim_{\gamma \to b} uv - \lim_{\gamma \to a} uv - \int_{\gamma=a}^{\gamma=b} v du, \tag{2.7}$$

where both u and v are functions of γ. Setting $u = \ln(1+\gamma)$ and $v = \frac{-\overline{\gamma}}{1+n} e^{-(1+n)\gamma/\overline{\gamma}}$, we can write the integral in (2.6) as:

$$\begin{aligned} &\int_{\gamma_{\mathrm{th},l}}^{\gamma_{\mathrm{th},l+1}} \ln(1+\gamma) e^{-(1+n)\gamma/\overline{\gamma}}\, d\gamma \\ &= \frac{\overline{\gamma}}{1+n} \left[\ln(1+\gamma_{\mathrm{th},l}) \cdot e^{-\frac{(1+n)\gamma_{\mathrm{th},l}}{\overline{\gamma}}} - \ln(1+\gamma_{\mathrm{th},l+1}) \cdot e^{-\frac{(1+n)\gamma_{\mathrm{th},l+1}}{\overline{\gamma}}} \right] \\ &\quad + \frac{\overline{\gamma}}{1+n} \int_{\gamma_{\mathrm{th},l}}^{\gamma_{\mathrm{th},l+1}} \frac{e^{-(1+n)\gamma/\overline{\gamma}}}{\gamma}\, d\gamma, \end{aligned} \tag{2.8}$$

using [2, (3.352.2)], to solve the integral in (2.8) and inserting the result in (2.6), gives the expression in (C.20).

References

[1] G. L. Stüber, *Principles of Mobile Communications*. Norwell, MA, USA: Kluwer, 1996.

[2] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products*. San Diego, CA, USA: Academic Press, 6th ed., 2000.

[3] A. J. Goldsmith and P. P. Varaiya, "Capacity of fading channels with channel side information," *IEEE Trans. Inform. Theory*, vol. IT-43, pp. 1896–1992, Nov. 1997.

Appendix 3

Derivation of $\Psi(\mu)$ in (D.28)

In this appendix we want to evaluate the integral $\Psi(\mu) = \int_0^\infty \ln^2(1+\gamma)e^{-\mu\gamma}d\gamma$, where μ is a constant. Changing the variable from γ to $x = \gamma + 1$, we obtain:

$$\Psi(\mu) = \int_0^\infty \ln^2(1+\gamma)e^{-\mu\gamma}d\gamma = e^{\mu}\int_1^\infty \ln^2(x)e^{-\mu x}dx. \tag{3.1}$$

The integral on the right-hand side of (3.1) can be solved by using [1, (4.358.1)]:

$$\begin{aligned} e^{-\mu}\Psi(\mu) &= \int_1^\infty \ln^2(x)e^{-\mu x}dx = \frac{\partial^2}{\partial\nu^2}\mu^{-\nu}\,\Gamma(\nu,\mu)|_{\nu=1} \\ &= \ln^2(\mu)\frac{1}{\mu}\Gamma(1,\mu) - 2\ln(\mu)\frac{1}{\mu}\frac{\partial}{\partial\nu}\,\Gamma(\nu,\mu)|_{\nu=1} + \frac{1}{\mu}\frac{\partial^2}{\partial\nu^2}\,\Gamma(\nu,\mu)|_{\nu=1}\,, \end{aligned} \tag{3.2}$$

where $\Gamma(\nu,\mu)$ is the *incomplete gamma function*. Inserting the first and second derivatives of $\Gamma(\nu,\mu)$ from [2] and setting $\nu = 1$, we obtain:

$$\begin{aligned} e^{-\mu}\Psi(\mu) &= \frac{1}{\mu}\Big\{\ln^2(\mu)\Gamma(1,\mu) - 2\ln(\mu)\Gamma(1)^2\mu\,{}_2\tilde{F}_2(1,1;2,2;-\mu) \\ &+ 2\Gamma(1,0,\mu)\ln^2(\mu) - 2\ln(\mu)\Gamma(1)\psi(1) + \Gamma(1,\mu)\ln^2(\mu) \\ &+ \Gamma(1)(\psi^2(1) + \psi'(1) - \ln^2(\mu)) \\ &- 2\mu\,{}_3F_3(1,1,1;2,2,2;-\mu) + 2\mu\ln(\mu)\,{}_2F_2(1,1;2,2;-\mu)\Big\}\,, \end{aligned} \tag{3.3}$$

where $\Gamma(x)$ is the *gamma function* [1, (8.310.1)], $\Gamma(x,y,z)$ is the *generalized incomplete gamma function* [3], $\psi(x) = \frac{d}{dx}\ln\Gamma(x)$ is the *psi function* [1, (8.360.1)],

and ${}_p\tilde{F}_q(a_1,\cdots,a_p;b_1,\cdots,b_q;\cdot)$ is the *regularized generalized hypergeometric function* [4].

Inserting $\Gamma(1)=\Gamma(2)=1$ [1, (8.338.1)], $\Gamma(1,0,\mu)=\Gamma(1)-\Gamma(1,\mu)$ [5], $\psi(1)=-C$ [1, (8.366.1)], $\psi'(1)=\frac{\pi^2}{6}$ [1, (8.366.8)], and ${}_2\tilde{F}_2(1,1;2,2;-\mu)={}_2F_2(1,1;2,2;-\mu)$ [6] into (3.3) we obtain (D.28).

References

[1] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products*. San Diego, CA, USA: Academic Press, 6th ed., 2000.

[2] Wolfram Research Inc., "Incomplete gamma function: Differentiation (subsection 20/01/01)." http://functions.wolfram.com/GammaBetaErf/Gamma2/20/01/01/.

[3] Wolfram Research Inc., "Generalized incomplete gamma function: Primary definition." http://functions.wolfram.com/06.07.02.0001.01.

[4] Wolfram Research Inc., "Regularized generalized hypergeometric function: Primary definition." http://functions.wolfram.com/07.32.02.0001.01.

[5] Wolfram Research Inc., "Generalized incomplete gamma function: Representation through equvivalent functions (subsection 27/02)." http://functions.wolfram.com/06.07.27.0002.01.

[6] Wolfram Research Inc., "Generalized hypergeometric function: Representations through more general functions (subsection 26/01/01)." http://functions.wolfram.com/07.31.26.0001.01.

Appendix 4

Derivations of Expressions in Table E.1

In this appendix we derive the expressions in Table E.1.

1 Round Robin (RR) Scheduling

If it assumed that the order of the users within each round is arbitrary, the number of allocated time-slots is independent of the user index i and we obtain the following probability mass function (PMF) for M:

$$p_M(k) = \begin{cases} \frac{(k+1)N-K}{N}, & k = \lfloor \frac{K}{N} \rfloor \\ \frac{K-(k-1)N}{N}, & k = \lceil \frac{K}{N} \rceil \\ 0, & \text{otherwise} \end{cases} . \tag{4.1}$$

It should be noted that after N time slots are allocated with the RR algorithm all the users have been allocated one time-slot. Hence, $p_M(0)$ is zero for $K \geq N$.

The mean number of bits transmitted per time-slot in k time-slots equals the expected number of bits transmitted in one time-slot [1, Eq. (34)]:

$$\mathrm{E}[b_{i,j}] = WT_{\mathrm{TS}} \int_0^\infty \log_2(1+\gamma) p_{\gamma_i}(\gamma) d\gamma = \frac{WT_{\mathrm{TS}}}{\ln 2} e^{1/\overline{\gamma}_i} E_1\left(\frac{1}{\overline{\gamma}_i}\right), \tag{4.2}$$

where $p_{\gamma_i}(\gamma)$ is the probability density function (PDF) of the carrier-to-noise ratio (CNR) of one user with average CNR $\overline{\gamma}_i$ and $E_1(x) = \int_1^\infty e^{-xt} dt$ is the exponential integral function.

The second moment of the number of bits $b_{i,j}$ transmitted to or from user i in the jth time-slot he is scheduled, can be stated as:

$$\mathrm{E}[b_{i,j}^2] = (WT_{\mathrm{TS}})^2 \int_0^\infty (\log_2(1+\gamma))^2 p_{\gamma_i}(\gamma) d\gamma = \frac{(WT_{\mathrm{TS}})^2}{\overline{\gamma}_i (\ln 2)^2} \Psi\left(\frac{1}{\overline{\gamma}_i}\right), \quad (4.3)$$

where $\Psi(\mu)$ is given by

$$\Psi(\mu) = e^{\mu}\left\{\frac{1}{\mu}\left[\frac{\pi^2}{6} + (C + \ln(\mu))^2\right] - {}_2F_3(1,1,1;2,2,2;-\mu)\right\}, \quad (4.4)$$

with $C = 0.57721566490$ being Euler's constant [2, (9.73)] and ${}_pF_q(a_1, \cdots, a_p; b_1, \cdots, b_q; \cdot)$ being the *generalized hypergeometric function* [3]. This expression has been found by using the derivation given in Appendix 3.

2 Max CNR Scheduling (MCS)

For MCS, where the user with the highest CNR is chosen in each time-slot, the number of time-slots allocated to user i within K time-slots is distributed according to the *binomial distribution* [4, p. 1179]:

$$p_M(k|i) = \binom{K}{k} p_i^k (1-p_i)^{K-k}, \quad (4.5)$$

where p_i is the probability of selecting user i in a time-slot, and can be expressed as [5, Eq. (12)]:

$$p_i = \int_0^\infty p_{\gamma_i}(\gamma) \prod_{\substack{j=1 \\ j\neq i}}^{N} P_{\gamma_j}(\gamma) d\gamma = \frac{1}{\overline{\gamma}_i} \sum_{\tau \in T_i^N} \mathrm{sign}(\tau) \frac{1}{\frac{1}{\overline{\gamma}_i} + |\tau|}, \quad (4.6)$$

where $P_{\gamma_j}(\gamma)$ is the CDF of the CNR of a single user with average CNR $\overline{\gamma}_j$, and T_i^N denotes set containing the terms that arise from an expansion of the product $\prod_{\substack{j=1 \\ j\neq i}}^{N} P_{\gamma_j}(\gamma) = \prod_{\substack{j=1 \\ j\neq i}}^{N} (1 - e^{-\gamma/\overline{\gamma}_j})$ [6, Sec. III-D-2]. As an example we now show this expansion for $i = 1$ and $N = 4$:

$$\begin{aligned} \prod_{\substack{j=1 \\ j\neq i}}^{N} (1 - e^{-\gamma/\overline{\gamma}_j}) &= 1 - \exp\left(-\frac{\gamma}{\overline{\gamma}_2}\right) - \exp\left(-\frac{\gamma}{\overline{\gamma}_3}\right) - \exp\left(-\frac{\gamma}{\overline{\gamma}_4}\right) \\ &+ \exp\left(-\frac{\gamma}{\overline{\gamma}_2} - \frac{\gamma}{\overline{\gamma}_3}\right) + \exp\left(-\frac{\gamma}{\overline{\gamma}_2} - \frac{\gamma}{\overline{\gamma}_4}\right) \\ &+ \exp\left(-\frac{\gamma}{\overline{\gamma}_3} - \frac{\gamma}{\overline{\gamma}_4}\right) - \exp\left(-\frac{\gamma}{\overline{\gamma}_2} - \frac{\gamma}{\overline{\gamma}_3} - \frac{\gamma}{\overline{\gamma}_4}\right). \quad (4.7) \end{aligned}$$

The absolute values of the elements in T_i^N are found by taking the natural logarithm of the terms arising from this expansion and multiplying the results with -1. For $i = 1$ and $N = 4$ the elements of T_1^4 thus have the following absolute values:

$$\left\{0, \frac{1}{\overline{\gamma}_2}, \frac{1}{\overline{\gamma}_3}, \frac{1}{\overline{\gamma}_4}, \left(\frac{1}{\overline{\gamma}_2}+\frac{1}{\overline{\gamma}_3}\right), \left(\frac{1}{\overline{\gamma}_2}+\frac{1}{\overline{\gamma}_4}\right), \left(\frac{1}{\overline{\gamma}_3}+\frac{1}{\overline{\gamma}_4}\right), \left(\frac{1}{\overline{\gamma}_2}+\frac{1}{\overline{\gamma}_3}+\frac{1}{\overline{\gamma}_4}\right)\right\}. \tag{4.8}$$

The signs of the elements in T_1^4 are the signs in the expanded product in (4.7):

$$\text{sign}(T_1^4) = \{+1, -1, -1, -1, +1, +1, +1, -1\}. \tag{4.9}$$

We now find the PDF of the CNR, conditioned on user i being scheduled. This distribution can be found by evaluating the following probability:

$$\begin{aligned}
\Pr(\gamma_i = \gamma | \gamma_j < \gamma_i, \forall j \neq i) &= \frac{\Pr(\gamma_i = \gamma \text{ and } \gamma_j < \gamma_i, \forall j \neq i)}{\Pr(\gamma_j < \gamma_i, \forall j \neq i)} \\
&= \frac{\Pr(\gamma_i = \gamma) \prod_{\substack{j=1 \\ j\neq i}}^{N} \Pr(\gamma_j < \gamma)}{\int_0^\infty p_{\gamma_i}(\gamma) \prod_{\substack{j=1 \\ j\neq i}}^{N} P_{\gamma_j}(\gamma) d\gamma} \\
&= \frac{p_{\gamma_i}(\gamma)}{p_i} \prod_{\substack{j=1 \\ j\neq i}}^{N} P_{\gamma_j}(\gamma).
\end{aligned} \tag{4.10}$$

This probability equals the PDF of the CNR when user i is scheduled, $p_{\gamma_i^*}(\gamma)$. We can easily verify that when the users have the same average CNRs, this PDF reduces to the conventional PDF for MCS found in [7]. We can now use this PDF to obtain the mean value for $b_{i,j}$ for MCS, using a similar derivation as for [5, Eq. (14)]:

$$\begin{aligned}
\text{E}[b_{i,j}] &= WT_{\text{TS}} \int_0^\infty \log_2(1+\gamma) p_{\gamma_i^*}(\gamma) d\gamma \\
&= \frac{WT_{\text{TS}}}{p_i \overline{\gamma}_i \ln 2} \sum_{\tau \in T_i^N} \text{sign}(\tau) \frac{e^{\left(\frac{1}{\overline{\gamma}_i} + |\tau|\right)}}{\frac{1}{\overline{\gamma}_i} + |\tau|} E_1\left(\frac{1}{\overline{\gamma}_i} + |\tau|\right).
\end{aligned} \tag{4.11}$$

Similarly, we can obtain the second moment of the number of bits $b_{i,j}$ transmitted to or from user i in the jth time-slot he is scheduled, by using the derivation in Appendix 3:

$$\begin{aligned} \mathrm{E}[b_{i,j}^2] &= (WT_{\mathrm{TS}})^2 \int_0^\infty (\log_2(1+\gamma))^2 p_{\gamma_i^*}(\gamma) d\gamma \\ &= \frac{(WT_{\mathrm{TS}})^2}{\overline{\gamma}_i p_i (\ln 2)^2} \sum_{\tau \in T_i^N} \mathrm{sign}(\tau) \Psi\left(\frac{1}{\overline{\gamma}_i} + |\tau|\right). \end{aligned} \tag{4.12}$$

3 Normalized CNR Scheduling (NCS)

Because all the users in our system model will have the same distribution for their relative CNRs [5], and because the relatively best users are scheduled, the probability of scheduling a user in a time-slot is the same for all the users. Consequently, the number of time-slots allocated to a user within K time-slots is also distributed according to the *binomial distribution* expressed in (4.5), with $p_i = \frac{1}{N}$ [8]. Because this distribution is the same for all the users, we may set $p_M(k|i) = p_M(k)$.

For the NCS policy, each user will experience a MUD gain as if all the other users were i.i.d. with the *same* average CNR as this user [9]. The CNR of user i in the time-slots he is scheduled can therefore be expressed with the following CDF [7]:

$$P_{\gamma_i^*}(\gamma) = P_{\gamma_i}^N(\gamma). \tag{4.13}$$

Differentiating this expression with respect to γ, we obtain the following PDF for the NCS algorithm:

$$p_{\gamma_i^*}(\gamma) = N P_{\gamma_i}^{N-1}(\gamma) p_{\gamma_i}(\gamma). \tag{4.14}$$

We can now use this PDF to obtain the mean value for $b_{i,k}$ for NCS, using a similar derivation as for [1, Eq. (44)]:

$$\begin{aligned} \mathrm{E}[b_{i,j}] &= WT_{\mathrm{TS}} \int_0^\infty \log_2(1+\gamma) p_{\gamma_i^*}(\gamma) d\gamma \\ &= \frac{NWT_{\mathrm{TS}}}{\ln 2} \sum_{j=0}^{N-1} \binom{N-1}{j} \frac{(-1)^j}{1+j} e^{\frac{1+j}{\overline{\gamma}_i}} E_1\left(\frac{1+j}{\overline{\gamma}_i}\right). \end{aligned} \tag{4.15}$$

Similarly, we can obtain the second moment of the number of bits $b_{i,j}$ transmitted to or from user i in the jth time-slot he is scheduled by using the

derivation in Appendix 3:

$$\begin{aligned} \mathrm{E}[b_{i,j}^2] &= (WT_{\mathrm{TS}})^2 \int_0^\infty (\log_2(1+\gamma))^2 p_{\gamma_i^*}(\gamma) d\gamma \\ &= \frac{N(WT_{\mathrm{TS}})^2}{\overline{\gamma}_i (\ln 2)^2} \sum_{j=0}^{N-1} \binom{N-1}{j} (-1)^j \Psi\left(\frac{1+j}{\overline{\gamma}_i}\right). \end{aligned} \tag{4.16}$$

4 Normalized Opportunistic Round Robin (N-ORR) Scheduling

For the N-ORR we schedule the user with the highest ratio $\chi_i(t) = \frac{\gamma_i(t)}{\overline{\gamma}_i}$ in each competition. As for the NCS algorithm, it can be shown that $\chi_i(t)$ is i.i.d. with unit average for all the users [5], and thus, all the participants in a competition have the same probability of winning. Consequently, the PMF for the number of time-slots M being allocated to a user is independent of i and can be expressed as in (4.1). The users that get $k = \lfloor \frac{K}{N} \rfloor$ time-slots are only involved in whole rounds of competitions. This means that these users will not participate in the last round of competitions that is finished before all the users are allocated one time-slot each. In this case, user i will experience a CDF of the CNR when he is scheduled that equals the average CDF over one round:

$$P_{\gamma_i^*}\left(\gamma | k = \left\lfloor \frac{K}{N} \right\rfloor\right) = \frac{1}{N} \sum_{n=1}^{N} P_{\gamma_i}^n(\gamma), \tag{4.17}$$

However, the CDF of the CNR for a user getting $k = \lceil \frac{K}{N} \rceil$ out of K time-slots, when K is not a multiple of N, can be expressed as the average over *all* the rounds, since such a user will also participate in the last unfinished round:

$$P_{\gamma_i^*}\left(\gamma | k = \left\lceil \frac{K}{N} \right\rceil\right) = \frac{(k-1)\sum_{n=1}^{N} P_{\gamma_i}^n(\gamma)}{kN} + \frac{\sum_{n=kN-K+1}^{N} P_{\gamma_i}^n(\gamma)}{k(K-(k-1)N)}. \tag{4.18}$$

Differentiating these CDFs with regard to γ, we obtain the corresponding PDFs:

$$p_{\gamma_i^*}\left(\gamma | k = \left\lfloor \frac{K}{N} \right\rfloor\right) = \frac{1}{N} \sum_{n=1}^{N} n P_{\gamma_i}^{n-1}(\gamma) p_{\gamma_i}(\gamma), \tag{4.19}$$

and

$$p_{\gamma_i^*}\left(\gamma|k=\left\lceil\frac{K}{N}\right\rceil\right) = \frac{(k-1)\sum_{n=1}^{N} nP_{\gamma_i}^{n-1}(\gamma)p_{\gamma_i}(\gamma)}{kN} + \frac{\sum_{n=kN-K+1}^{N} nP_{\gamma_i}^{n-1}(\gamma)p_{\gamma_i}(\gamma)}{k(K-(k-1)N)}. \quad (4.20)$$

We can now express the first moment of $b_{i,j}$ as:

$$\mathrm{E}[b_{i,j}] = WT_{\mathrm{TS}}\int_0^{\infty}\log_2(1+\gamma)p_{\gamma_i^*}(\gamma|k)p_M(k)d\gamma. \quad (4.21)$$

For $k=\lfloor\frac{K}{N}\rfloor$ we have

$$\mathrm{E}[b_{i,j}] = \frac{WT_{\mathrm{TS}}}{N}\sum_{n=1}^{N}A_i(n), \quad (4.22)$$

while for $k=\lceil\frac{K}{N}\rceil$ we have

$$\mathrm{E}[b_{i,j}] = WT_{\mathrm{TS}}\left(\frac{(k-1)\sum_{n=1}^{N}A_i(n)}{kN} + \frac{\sum_{n=kN-K+1}^{N}A_i(n)}{k(K-(k-1)N)}\right), \quad (4.23)$$

where $A_i(n)$ is given by [9, Eq. (20)]:

$$A_i(n) = \frac{n}{\ln 2}\sum_{j=0}^{n-1}\binom{n-1}{j}\frac{(-1)^j}{1+j}e^{\frac{1+j}{\overline{\gamma}_i}}E_1\left(\frac{1+j}{\overline{\gamma}_i}\right). \quad (4.24)$$

Similarly, we can express the second moment of the number of bits $b_{i,j}$ transmitted to or from user i in the jth time-slot he is scheduled, as:

$$\mathrm{E}[b_{i,j}^2] = (WT_{\mathrm{TS}})^2\int_0^{\infty}(\log_2(1+\gamma))^2 p_{\gamma_i^*}(\gamma|k)p_M(k)d\gamma. \quad (4.25)$$

For $k=\lfloor\frac{K}{N}\rfloor$ we have:

$$\mathrm{E}[b_{i,j}^2] = \frac{WT_{\mathrm{TS}}}{N}\sum_{n=1}^{N}B_i(n), \quad (4.26)$$

while

$$\mathrm{E}[b_{i,j}^2] = WT_{\mathrm{TS}}\left(\frac{(k-1)\sum_{n=1}^{N}B_i(n)}{kN} + \frac{\sum_{n=kN-K+1}^{N}B_i(n)}{k(K-(k-1)N)}\right), \quad (4.27)$$

for $k=\lceil\frac{K}{N}\rceil$, where $B_i(n)$ is found from using the derivation in Appendix 3:

$$B_i(n) = \frac{n}{\overline{\gamma}_i(\ln 2)^2}\sum_{j=0}^{n-1}\binom{n-1}{j}(-1)^j\Psi\left(\frac{1+j}{\overline{\gamma}_i}\right). \quad (4.28)$$

References

[1] M.-S. Alouini and A. J. Goldsmith, "Capacity of Rayleigh fading channels under different adaptive transmission and diversity-combining techniques," *IEEE Trans. on Veh. Technol.*, vol. 48, pp. 1165–1181, July 1999.

[2] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products*. San Diego, CA, USA: Academic Press, 6th ed., 2000.

[3] Wolfram Research Inc., "Generalized hypergeometric function: Primary definition." `http://functions.wolfram.com/07.31.02.0001.01`.

[4] E. Kreyszig, *Advanced Engineering Mathematics*. New York: John Wiley & Sons, Inc., 7th ed., 1993.

[5] L. Yang and M.-S. Alouini, "Performance analysis of multiuser selection diversity," in *Proc. IEEE Int. Conf. on Communications (ICC'04)*, (Paris, France), pp. 3066–3070, June 2004.

[6] T. Eng, N. Kong, and L. B. Milstein, "Comparison of diversity combining techniques for Rayleigh-fading channels," *IEEE Trans. Commun.*, vol. 44, pp. 1117–1129, Sept. 1996.

[7] R. Knopp and P. A. Humblet, "Information capacity and power control in single cell multiuser communications," in *IEEE Int. Conf. on Communications (ICC'95)*, (Seattle, WA, USA), pp. 331–335, June 1995.

[8] D. Avidor, S. Mukherjee, J. Ling, and C. Papadias, "On some properties of the proportional fair scheduling policy," in *Proc. IEEE Symp. on Personal, Indoor and Mobile Radio Communications (PIMRC'04)*, vol. 2, (Barcelona, Spain), pp. 853–858, Sept. 2004.

[9] V. Hassel, M. R. Hanssen, and G. E. Øien, "Spectral efficiency and fairness for opportunistic round robin scheduling." Presented at the *IEEE Int. Conf. Comm. (ICC'06)*, (Istanbul, Turkey), June 2006.