ABSTRACT

Title of dissertation: RESOURCE ALLOCATION SCHEMES
FOR OFDMA BASED WIRELESS SYSTEMS
WITH QUALITY OF SERVICE CONSTRAINTS

Tolga Girici, Doctor of Philosophy, 2007

Dissertation directed by: Professor Anthony Ephremides
Department of Electrical and Computer Engineering

With its capabilities like elimination of intersymbol interference, intercell interference averaging, scalability and high bandwidth efficiency OFDMA is becoming the basis for current wireless communication technologies. In this dissertation we study the problem of multiple access and resource allocation for OFDMA-based cellular systems that support users with various quality of service (QoS) requirements.

In Chapters 2 and 3 of the dissertation, we consider the problem of downlink transmission (from base station to users) for proportional fairness of long term averaged received rates of data users as well as QoS for voice and video sessions. Delay requirements of real time sessions are converted into rate requirements at each frame. The base station allocates available power and bandwidth to individual users based on received rates, rate constraints and channel conditions. We formulate and solve the underlying constrained optimization problem and propose an algorithm that achieves the optimal allocation. In Chapter 3, we obtain a resource allocation scheme that is simpler but achieves a performance comparable to the optimal algorithm proposed in Chapter 2. The algorithms that

we propose are especially intended for broadband networks supporting mobile users as the subchannelization scheme we assume averages out the fading in subchannels and performs better under fast fading environment. This also leads to algorithms that are simpler than the ones available in the literature.

In Chapter 4 of the dissertation we include relay stations to the previous model. The use of low-cost relay stations in OFDM based broadband networks receives increasing attention as they help to improve the cell coverage. For a network supporting heterogeneous traffic we study TDMA based subframe allocation for base and relay stations as well as joint power/bandwidth allocation for individual sessions. We propose an algorithm again using the constrained optimization framework. Our numerical results prove that our multihop relay scheme indeed improves the network coverage and satisfy QoS requirements of user at the cell edge.

In the last Chapter, we deviate from the previous chapters and consider an OFDMA based system where the subchannels experience frequency selective fading. We investigate a standard subchannel allocation scheme that exploits multiuser diversity by allocating each subchannel to the user with maximum normalized SNR. Using extreme value theory and generating function approach we did a queueing analysis for this system and estimated the QoS violations through finding the tail distribution of the queue sizes of users. Simulation results show that our estimates are quite accurate and they can be used in admission control and rate control to improve the resource utilization in the system.

RESOURCE ALLOCATION SCHEMES FOR OFDMA BASED WIRELESS SYSTEMS WITH QUALITY OF SERVICE CONSTRAINTS

by

Tolga Girici

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2007

Advisory Committee:
Prof. Anthony Ephremides, Chair/Advisor
Prof. Armand Makowski
Assoc. Prof. Sennur Ulukus
Assoc. Prof. Richard H. La
Prof. Aravind Srinivasan, Dean's Representative

DEDICATION

To my father Muzaffer Girici

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my research advisor Anthony Ephremides for his guidance throughout my graduate study at University of Maryland. When I first came to USA I didn't have an assistantship, but he believed in me. His broad knowledge, interest and intuition in all areas of wireless communications will affect the way I do my research throughout my life. He is a role model in research, teaching and generally as a person.

I would like to thank my thesis committee members, Professor Armand Makowski, Professor Sennur Ulukuş, Professor Richard H. La and Professor Aravind Srinivasan. Armand Makowski and Sennur Ulukuş were also present in my PhD proposal defense. We had a really informative discussion in my defense and their perspective and suggestions will be useful in improving and extending this work.

Between January 2006 and June 2007, I worked as an assistant at the College Park Fujitsu Lab. This was the most fruitful period of my graduate studies. I would like to thank Dr. Chenxi Zhu and Dr. Jonathan Agre and everyone in the Fujitsu Lab for their collaboration and guidance. I am especially indebted to Chenxi; this thesis wouldn't be possible without him.

My life at Maryland was a long and rough road, but when I look back, I remember many good memories that I share with my friends at College Park. I would especially like to thank Tuna Güven and Onur Kaya for being my roommates and closest friends. We

were always together for years and our friendship will continue forever.

I do not know how to thank my family for their endless support encouragement and love. My parents Muzaffer and Neriman Girici and my brother Emre Girici were always there for me in good and bad times. They were always proud of me. My wife Yasemin Girici is the meaning of my life. She always believed in me more than I believe in myself.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Design of wireless systems involve finding solutions to some *link-level* and *system level* challenges. Link-level challenges are primarily caused by physical medium, which are the channel fading (varying with time and frequency) and multiple access interference. A variety of modulation and coding schemes have been previously proposed in order to overcome these challenges. On the other hand system level challenges are caused by some specific properties of the wireless system, e.g. number and types of users using the network, Quality of Service requirements of different types of traffic. In this thesis we will mainly concentrate on resource allocation in wireless multiple access which requires joint consideration of these link and system level challenges.

1.1 Background and Related Work

Resource allocation and scheduling is of paramount importance in wireless networks, where the resources (power, bandwidth, time) are scarce and channel conditions like noise, fading and shadowing are much more severe when compared to their wired counterparts.

What makes this resource allocation problem more challenging and interesting is that the available resources are shared by users, which are subject to statistically different channel conditions and which demand different types of services. This requires to

change the classical layered approach to network design and analysis, and adopt a new design paradigm, which is called *cross-layering*. One of the most common examples is performing medium access control (MAC) layer functions by taking into account the instantaneous and long term channel conditions, which is a physical layer quantity. In fact, using the channel information it is possible to increase throughput by scheduling at each time slot, the user with the best channel conditions. This is referred to as *multiuser-diversity* [1], which increases the throughput gain as the number of users increases. The scheduling schemes that exploit this diversity are referred to as *opportunistic schedulers* [2]. Recent high speed communication technologies 1$\times$ EV-DO and High Speed Packet Data Access (HSPDA) are based on this phenomenon.

Opportunistic scheduling schemes such as 1$\times$EV are initially designed to support data services. Data user with the best channel condition is scheduled at each time slot. This brings up the issue of *fairness* because users located further away from the Base Station have much less chance of having the best channel. To solve this problem *Proportional Fair* (PF) schedules are proposed, which look at the ratio of current achievable rates and long term received rates. This provides a fair balance between spectral efficiency (bits/sec/Hz) and fairness. High Data Rate (HDR) technology [3] for data communications is based on this technique.

Ever growing demand for online multimedia applications requires scheduling schemes that achieve much higher rate and quality of service (QoS) for various types of services. Various applications such as Web Browsing, FTP, VoIP, Video Streaming and even interactive online gaming have much different traffic loads and delay requirements. Most commonly, the transmitter (e.g. Base Station (BS)) allocates separate buffers for incom-

ing traffic belonging to different types of applications. While scheduling, the buffer occupancy level and delay of the head-of-line packet (which are originally Network Layer parameters) are taken into account. This is another example for cross-layering.

In this work we will study scheduling of heterogeneous traffic for multiple access systems that have multichannel transmission capability. By using multichannel transmission techniques a user can get a number of parallel channels depending on its channel condition and rate requirements and transmit without interfering with other users. Multilevel Modulation and Coding Schemes (MCS) are also employed in order to cope with multipath fading and achieve high data rate and low bit error rates (BER). For example WCDMA based systems such as (HSDPA) use multiple orthogonal spreading sequences and OFDMA based system such as WiMax and Long Term Evolution (LTE) use multiple orthogonal subcarriers. In this work we consider OFDMA as the multicarrier transmission scheme. Within OFDMA framework, the resources allocated to the users come in three dimensions: time slots, frequency and power. This requires the scheduler to operate with higher degree of freedom and more flexibility, and potentially higher multiplexing capacity. This also makes the notion of resource fairness obsolete and makes the problem more involved. We plan to develop scheduling algorithms fully taking advantage of the degree of freedom inherent to OFDMA system. Below, we briefly explain OFDMA technology and its recent applications. This will also help to explain the motivation in choosing this transmission scheme in this thesis.

1.1.1 OFDMA Technology and its Advantages

OFDM is a digital modulation scheme in which a wideband signal is split into a number of narrowband signals. Because the symbol duration of a narrowband signal is larger than that of a wideband signal, the amount of time dispersion caused by multipath delay spread is reduced. OFDM is a special case of multicarrier modulation in which multiple user symbols are transmitted in parallel using different subcarriers with overlapping frequency bands that are mutually orthogonal. This technique implements the same number of channels as conventional FDM with a much reduced bandwidth requirement. In conventional FDM, adjacent channels are well separated using a guard interval. In order to realize the overlapping technique, interference between adjacent channels must be reduced. Therefore, orthogonality between subcarriers is required. In OFDM each subcarrier has an integer number of cycles within a given time interval , and the number of cycles by which each adjacent subcarrier differs is exactly one. This property ensures OFDM subcarrier orthogonality. The subcarriers are data modulated and are fed through a serial- to-parallel conversion process. Each symbol is assigned a subcarrier and an inverse DFT (IDFT) performed to produce a time domain signal.

OFDM deals with multipath delay spread by dividing the total bandwidth B into K narrowband channels where K is the number of subcarriers. Orthogonal Frequency Division Multiple Access (OFDMA) is an extension of OFDM, where multiple users can transmit at the same time by sharing the subcarriers. In order to make this resource sharing more practical subcarriers are grouped into *subchannels*. There are various ways to group the subcarriers, i.e. *subchannelization* methods. There are two classes of subcar-

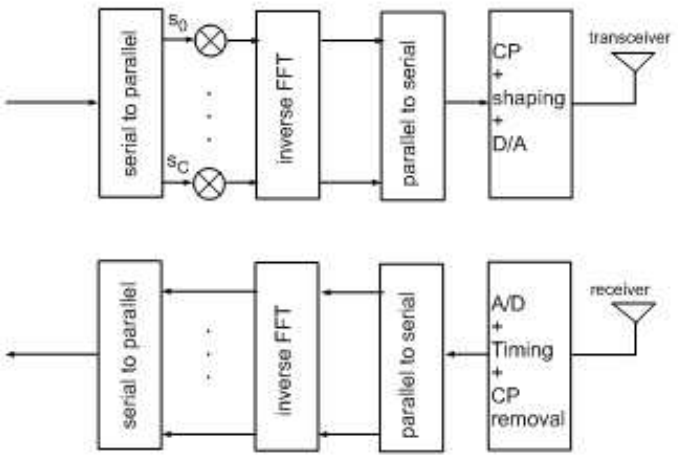

Figure 1.1: OFDM Diagram

rier grouping modes, *distributed* and *adjacent*, which are roughly illustrated in figure 1.2. In general, distributed subcarrier permutations perform very well in mobile applications while adjacent subcarrier permutations can be properly used for fixed, portable, or low mobility environments [4]. *Adjacent* subchannelization (AMC) uses adjacent subcarriers to form subchannels. When used with fast feedback channels it can rapidly assign a modulation and coding combination per subchannel. On the other hand *distributed* subchannelization (PUSC, FUSC) employs full-channel diversity by distributing the allocated subcarriers to subchannels using a permutation mechanism. By this way, a user observes the same channel quality in all subchannels. Frequency diversity minimizes the performance degradation due to fast fading characteristics of mobile environments. It has been previously observed that adjacent subchannelization provides more capacity ($\sim 10\%$ [5],[6], [7]) than distributed methods because of the averaging effect. On the other hand especially for mobile systems distributed methods provide better channel estimation and

easiness of allocation in a fast fading environment. Hence, distributed subchannelization is the method that we used in this work.

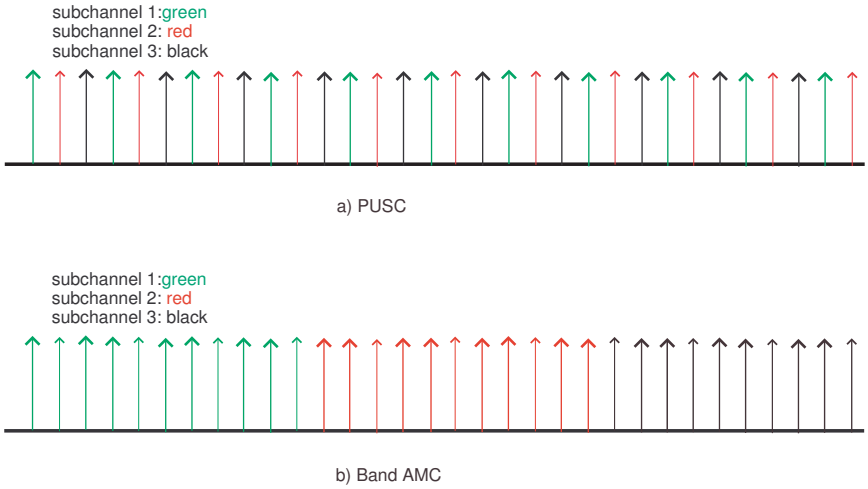

Figure 1.2: PUSC and AMC subchannelization example in a 3-subchannel OFDMA system is shown.

Advantages of OFDM with respect to its counterparts (e.g CDMA) can be summarized as follows. By using narrowband signals OFDM can combat multipath delay spread more effectively. The reason is that the wavelength of a narrowband signal is much greater than a typical multipath delay spread. This makes OFDM successful in Non-Line-of-Sight (NLOS) communication systems. Moreover, OFDM distributes the information across several subcarriers, with the use of forward error correction (FEC), if an error occurs in one subchannel, those errors are recovered by FEC. OFDM also has better spectral efficiency since intersymbol interference is eliminated by using the cyclic prefix. Therefore OFDM also doesn't require channel equalization. Besides OFDMA has the scalability advantage through using different FFT sizes without changing subcarrier

spacing (521, 1024, 2048 FFT). By this way, increasing system bandwidth doesn't affect multipath fading.

1.1.1.1 WiMax Technology

One of the reasons that we studied OFDMA based scheduling is its applications in recently developed technologies like WiMax. The system considered in this work is motivated by the recent IEEE 802.16 standard that defines the air interface and medium access control (MAC) specifications for wireless metropolitan area networks. Such networks intend to provide high speed voice, data and on demand video streaming services for end users. IEEE 802.16 standard is often referred to as WiMax and it provides substantially higher rates than cellular networks. Besides it eliminates the costly infrastructure to deploy cables, therefore it is becoming an alternative to cabled networks, such as fiber optic and DSL systems [8]. Although originally the standard [8] is for communication in 11-66 GHz range, more recent updates on this standard allows communication in 2-11 GHz frequency range, which is more suitable for non line of sight (NLOS) communications [9], [10].

WiMax networks are designed for point to multipoint communications, where a base station (BS) transmits to and receives from multiple subscriber stations (SS) in a cellular coverage area of typical size around 5miles. A SS can be either an end user itself, or be the backbone connection of a WLAN. We consider the end user scenario since it is more interesting because of user mobility and channel fading. The framework that we adopt in our work is mostly in line with the Mobile WiMax standard (IEEE 802.16e) that

is updated as of March 2006 [11], [12]. In Korea a system named as WiBro is designed according to this standard and it will be launched commercially in the middle of 2006 [13]. Initially we focus on the traffic from BS to SSs (downlink).

In this thesis our goal is to find multicarrier fair schemes that also satisfy heterogeneous stability and delay requirements. We propose resource allocation algorithms for OFDMA-based downlink and uplink communications. These two directions of communications reveal different trade-offs, which are worth investigating separately.

1.1.2 Downlink Communications

Downlink means the transmission from the base station to users in a cellular area. Main constraints in OFDM based downlink transmission is the total power and bandwidth of the base station. Fair downlink scheduling schemes with QoS considerations were proposed and studied previously for single carrier systems. Only very recently in [14], [15], [16], [17], [18], proportional fair scheduling was studied for multicarrier systems. However, in [15] it is studied without power control and no algorithm was proposed to find the optimum bandwidth allocation. The work in [14] also has a proportional rate constraint, where the rates of individual users has to be in certain proportions in order to maintain fairness. In [16] and [18] proportional fair scheduling is addressed for a single time instant, rather than the long term received rates. Besides in all of these works supporting real time traffic with QoS requirements was not addressed. The scheduling rules do not apply sufficiently to different QoS requirements and heterogeneous traffic. We have to note that the work in [19] jointly considers data, voice and video traffic,

however they do not consider power control and they don't distinguish between best effort traffic and real time traffic.

A major drawback of proportional fair scheduling is that it assumes there are infinite packets to be transmitted at time zero and no packet arrivals. For FTP sessions, it is reasonable to assume that large files are ready transmit at the beginning of a session, which is not the case for real time applications such as VoIP and Video Streaming. Different real time applications can have different arrival rates, therefore average rate in the long run should be larger than the arrival rate for each session in order to maintain stability. In [20] it was shown by some examples that Proportional Fair scheduling does not guarantee stability of the queues in some situations that can actually be stabilized. Therefore our goal is to improve Proportional Fair scheduling in order to maintain stability. This could be done by putting constraints on transmission rates. Another drawback of proportional fair scheduling is that it does not support heterogeneous QoS requirements. For example in VoIP and Video Streaming applications there is a delay requirement for each packet. If a packet can not be transmitted in a certain time interval then that packet has to be dropped, which degrades the quality of real time sessions. In proportional fair scheduling there is a long term rate requirement, while in real time sessions there is a short term rate requirement.

OFDMA based resource allocation has been studied also without the fairness and QoS objectives in [21], [22], [23], [24], [25], [26]. The work [21] and [24] propose sub-carrier and bit allocation algorithms that satisfy rate requirements of users with minimum total power. The papers [22] and [25] address maximizing total throughput subject to power and subcarrier constraints and do not address real time traffic. The authors in [23]

are interested in maximizing the worst user's capacity. Cendrillon et. al. in [26] maximize a weighted sum of users' capacities which gives a feeling of fairness however it doesn't necessarily provide proportional fairness.

Uplink means transmission from users to base station in a cellular network. This brings different trade-offs than downlink transmission. First of all unlike base stations, mobile devices carry limited-sized power sources and there is an individual power constraint unlike downlink communications. OFDMA-based resource allocation in uplink systems were studied in [27], [28] and [29]. In [27] total capacity was maximized subject to individual power constraints, while in [28] and [29] sum-power was minimized subject to individual rate constraints.

There are also some papers that study other multicarrier transmission schemes such as multicode CDMA. For example [30] study a fair queueing scheme with time varying weight assignment. Weights are proportional to the channel conditions divided by long term received rates. In [31] throughput maximizing power and spreading code allocation subject to total power and bandwidth constraints is studied. Abedi et. al. in [32], propose a QoS-based packet scheduler for HSPDA systems that are based on WCDMA technique. The proposed scheme is purely based on heuristics.

In the systems that we consider the Base Station has a large coverage area. Especially in urban areas, this may cause problems for the line-of-sight communication because tall buildings can create holes in the coverage area. In this thesis we develop resource allocation algorithms to improve the network performance by deploying fixed relay devices in order to eliminate shadowing and improve the performance. This idea has similarities with *Mesh Networks*, where each user operates also as a router and packets

are forwarded from a gateway in a multihop fashion. Unlike mesh networks we perform this relaying function by deploying *relay stations*, which act like small base stations. Base Station assigns each user either to itself or one of the relay stations. These relay stations have a single interface in order to keep them inexpensive. Hence, they can't transmit and receive simultaneously. This leads us to schedule the transmissions of base and relay stations in a TDMA manner. We develop an algorithm that allocates time , subchannel and power to each session in a frame.

1.2 System Model

Below, we briefly explain the physical, medium access control and network layer assumption that we will use throughout the thesis.

1.2.1 Adaptive Modulation and Coding

We assume a channel that experiences path loss, Rayleigh fading and Log-normal shadowing. Although the system we consider is a mobile system we do not change the distance from the BS to the MS in the analysis and simulations, but we do simulate a fast and slow fading channel for each BS-MS link, which is a reflection of mobility. Let N_0 be the noise power spectral density. We assume that this also includes the inter-cell interference. Let $g_i(t)$ be the combined channel gain for user i at time t. Then, the SINR for user i, (γ_i) is as follows:

$$\gamma_i(t) = \frac{p_i(t)g_i(t)}{N_0 w_i(t)}, \tag{1.1}$$

where $p_i(t), w_i(t)$ are the power and bandwidth allocated to user i at time t. Using pilot symbols inserted to the downlink frame the mobiles can effectively estimate the channel parameter $g_i(t)$. We assume perfect channel estimation and feedback. We assume that channel conditions are constant at each frame and therefore assume AWGN channel with SINR as in (1.1).

Adaptive Modulation and Coding and fast channel feedback are used in our system model to enhance the coverage and capacity. It has been shown in [33], [34] that adaptive modulation effectively improves the BER performance on wireless channels and relieves the effects of deep fading. In line with the IEEE 802.16 standard, in our model the base station chooses a modulation level from a set of available levels from 4-QAM to 64-QAM depending on the current signal to noise ratio (SNR) and target bit error rate (BER). We assume that at each time slot the channel gain (fading and path loss) hence SNR is constant, therefore the channel in a time slot can be considered as an AWGN channel. Performance of adaptive modulation in AWGN channels was studied in [33], [34]. There, it was shown that the BER for an M-QAM modulation can be well approximated by

$$BER \simeq 0.2 \exp\left[-1.5\gamma/(M-1)\right] \tag{1.2}$$

Let $T_i(\gamma_i)$ be the throughput, which is the number of bits that can successfully be sent in a symbol for a given SNR, γ_i for user i. Therefore for a constant BER requirement the throughput can be approximated by

$$T_i(\gamma_i) = \log_2 M(\gamma_i) = \log_2(1+\beta\gamma_i) \tag{1.3}$$

where β is equal to $-1.5/\ln(5BER)$ from (1.2). The throughput formulation has a form similar to the Shannon capacity.

In our model convolutional coding and repetition coding is applied to the uncoded bit stream before modulation to reduce the BER. Effects of using different set of modulation and coding pairs is beyond the scope of this thesis. Instead, we use predefined set of modulation/coding pairs in the IEEE 802.16 OFDMA standard [11], [35]. The table below shows the modulation levels/coding rates and corresponding throughput and optimal SNR values for a target $BER = 10^{-4}$.

Mod./Coding	**Repetition**	**Rate(bps/Hz)**	**SNR(dB)**
QPSK,1/2	6×	1/6	-2.78
QPSK,1/2	4×	1/4	-1.0
QPSK,1/2	2×	1/2	2.0
QPSK,1/2	1×	1	5
QPSK,3/4	1×	1.5	8
16QAM,1/2	1×	2	10.5
16QAM,3/4	1×	3	14
64QAM,2/3	1×	4	18
64QAM,3/4	1×	4.5	20

Table 1.1: Optimal Modulation and Coding Schemes Corresponding to SNR Values

We assume that all types of traffic traffic have same BER requirements, however, the proposed schemes can easily be extended for different BER requirements. If we plot the spectral efficiency values (in bps/Hz) in this table as a function of given SNR's, we see that using formula in (1.3) by setting $\beta = 0.25$ is a reasonable approximation. Therefore in the following chapters we will use (1.3) in the problem formulations as the rate function. Please note that the performance of the system can be improved by enlarging the set of available modulation and coding pairs. However this is beyond the

scope of this dissertation.

1.2.2 Subchannel and Power Allocation

In this work our general approach is to formulate the resource allocation problem as constrained optimization problems, where the objective function is maximized subject to some power , bandwidth and rate constraints. As for the subcarrier allocation, we consider the asymptotic case, where the available bandwidth is a continuous and infinitely divisible quantity. However, after computing the power and bandwidth for each node, we quantize the bandwidth w_i to an integer multiple of subchannel bandwidth W_{sub}. Then we update the power p_i for each node i, so that the resulting SNR values are quantized to the closest values in Table 1.1. We can always improve the performance by using more modulation/coding pairs and less subchannel bandwidth.

1.2.3 MAC Layer Scheduling

We are considering a MAC layer that supports Best Effort data traffic while simultaneously supporting Streaming Video and delay sensitive VoIP traffic over the same channel. The resource allocated to one terminal can vary from single subchannel to the entire frame. Including power control this provides a very large dynamic range of throughput to a specific user at any time. Normally the resource allocation information should be conveyed in a portion of the frame, however we neglect the number of slots and subchannels allocated for control messages.

In this work we are considering either solely the traffic from the Base Station(BS)

to mobile nodes(MS's) (*downlink*) or from MS's to BS's (downlink). Normally the two directions of traffic are separated by forming a duplex link either by dividing time or frequency.

1.2.4 Quality of Service Support

The system that we consider should be able to support a range of traffic types. Each type of traffic (flow) is associated with certain Quality of Service (QoS) parameters. The base station allocates resources according to these parameters or constraints. The traffic arriving at the Base Station is supposed to come from a high capacity wired link. The link from the Base Station to the mobile nodes (i.e. the air interface) is considered as the bottleneck. The types of services that we consider in this work are summarized as follows:

Application	**QoS Category**	**QoS Specifications**
FTP	Non-Real Time Packet Service	Minimum Reserved Rate
Web Browsing	Best Effort Service	No delay or rate Requirement
VoIP	Unsolicited Grant Service	Max. Delay Constraint
Video Streaming	Real Time Packet Service	Min. Reserved Rate & Delay Const.

Table 1.2: Supported Applications and QoS Specifications

Admission control is beyond the scope of this work therefore we assume no new session arrivals throughput the simulation time. In some problem formulations we convert the minimum reserved rate requirement to delay requirements in order to formulate delay based optimizations. We also assume that all of the sessions continue throughout the simulation time. Since we are considering simulation times on the order of seconds this

is a realistic assumption.

1.2.5 Queueing Model and Analysis

We assume that the Base Station classifies the arriving packets according to its traffic type and its intended mobile user. For simplicity we assume that a user can only demand a single type of traffic. For each user, a separate buffer is allocated .

For data applications like FTP and Web Browsing, we assume that there are always unlimited number bits waiting to be transmitted at the base station. This is a realistic assumption since the total bits in these sessions are on the order of MB's and we consider simulation times on the order of seconds.

For real time traffic sessions such as Video Streaming and VoIP we assume to have a queue with infinite capacity. We capture the performance by measuring 95^{th} percentile of packet delays. Let $q_i(t)$ be the amount of bits in the queue of user i at frame t. During frame t the queue of user i is served at rate $r_i(t)$. Let $a_i(t)$ be the number of bits that arrive at frame t. We assume that bits arrive at the beginning of a frame (before the transmission starts). Then the queue length evolution equation can be written as,

$$q_i(t+1) = q_i(t) + a_i(t) - \min(q_i(t) + a_i(t), r_i(t)T_f) \tag{1.4}$$

For VoIP sessions we assume a constant bit rate arrival process, where a constant number of bits arrive with constant time intervals. For video traffic we assume a bursty traffic model in IEEE 802.16 specifications. The details of the bit arrival process will be explained later.

Most of the previous works on multiuser wireless packet communication systems

decoupled information theory and queueing theory. The references that we sited previously either considered systems in saturation mode and proposed schemes that maximize total throughput or proportional fair capacity, and/or satisfy some rate constraints.

Joint consideration of queueing and information theory was studied for the case of single user systems in order to jointly optimize energy expenditure and delay. Energy efficient transmission has been studied previously for a single user system. For example in [36], [37], [38], the authors studied the problem of minimizing energy expenditure of transmitting randomly arriving packets subject to a transmission deadline constraint in a fading channel. The paper [39], [40] is an extension of [38] that studies joint minimization of delay and energy. In [41] Berry and Gallager obtain structural results that points out a tradeoff between delay and energy in a single user transmission. They show that the optimal power delay curve is convex. The work in [42] extends [41] and finds a closed form expression of optimal policy in terms of the optimal policy when the signal to noise ratio is one. They also find some structural results for the optimal policy and bounds for the optimum cost. However these works don't offer any formulas or expressions for delay.

For multiuser systems [43] analyze the trade-off between error probability and delay in a multiple access system. However this framework couldn't be extended because of the complex nature of wireless multiple access. In this thesis using discrete time multi-server queueing framework [44] we make a queueing analysis of a simple OFDMA based system.

1.3 Thesis Outline

The thesis is organized as follows. In Chapter 2, we consider a Base Station transmitting to a set of mobile users that demand voice, video and data sessions. We propose a power and bandwidth allocation scheme that provides long term proportional fairness to data users, while satisfying the delay requirements of voice users and rate requirements of video users. We formulate and solve a constrained optimization problem that captures these objectives. We then develop an algorithm that finds the optimal allocation. Using simulations we compare the performance of the algorithm with a well known benchmark algorithms from the literature.

In Chapter 3, we consider the complexity of the proposed algorithm in Chapter 2 and propose a simpler algorithm. In order to make the resource allocation computationally simpler, we propose *user selection metrics*, that are used by the Base Station to select Voice, Data and Video Streaming users from the total set of users. That way we decrease the number of users entering into the computation process. In addition to this we distinguished elastic and non-elastic real time traffic. We determined minimal required rates for real time sessions and formulate a constrained optimization problem to find the allocation to maximize the proportional fair capacity for elastic best effort and real time traffic subject to rate constraints for elastic real time traffic and, total power and bandwidth. We compared this algorithm with a benchmark algorithm.

In Chapter 4, we deviate from the classical downlink case and consider a system that includes fixed relay stations (RSs) located in the cellular area. These relays are useful in reducing shadowing and increasing the capacity. We develop a resource allocation

scheme in which the base station first performs a simple 2-hop routing that assigns users either to itself or one of the relays. Then the BS allocates a subframe to each BS-RS-MS microcell. Then, for each microcell the BS performs subframe allocation for BS-RS and RS-MS composite links and joint subchannel and power allocation to each link in order to provide and proportional fairness to data users subject to rate constraints of real time sessions. We did simulations in order to see the performance of the system with the performance of a system with no RS.

In Chapter 5 we consider an OFDMA based system that experiences frequency selective fading at each subchannel. We consider a simple channel-aware resource allocation scheme that allocates each subchannel to the user with maximum normalized received SNR. Using queueing theory for discrete time multiserver systems, we perform a queueing analysis for this system. First using extreme value theory we model the service process. Then we analyze the tail probability distribution of the buffer occupancy. We compare the accuracy of our analysis with the simulation results.

Chapter 2

Proportional Fair Scheduling in OFDMA-based Wireless Downlink Systems with QoS Constraints

2.1 Introduction

In this chapter we consider a base station serving users demanding heterogeneous traffic, which are best effort data, video streaming and voice. We develop a resource allocation algorithm that provides proportional fairness among data users based on their long term received data rates unlike single instant data rates as in [16] and [18]. We develop a user selection scheme that selects a number of real time sessions based on their head-of-line packet delays and received rates. We determine their rate requirements and formulate a constrained optimization problem that maximizes proportional fairness subject to those rate requirements and power and bandwidth constraints. In Section 2.2 we describe the system model. In Section 2.3 we investigate the proportional fairness and formulate the proportional fair rate allocation. In Section 2.4 we investigate and formulate the user selection and rate requirement determination process for real time sessions. In Section 2.5 we formulate and solve the joint data and real time resource allocation problem. We also look at the feasibility of a problem given the rate constraints and how to detect infeasibility. In Section 2.6, based on the solution, we describe the resource allocation algorithm and prove that it converges to the unique optimal solution. Finally in Section 2.7 we nu-

merically demonstrate the performance improvement of the proposed resource allocation algorithm.

2.2 System Model

We adopt the WirelessMAN-OFDMA profile [11], [10] at the physical layer, which is a multicarrier scheme where multiple access is provided by assigning a subset of carriers to each receiver at each time slot. Let W and P denote the total bandwidth and power, respectively. Total bandwidth W is divided into N_{sub} subchannels of length W_{sub} Hz, each consisting of a group of carriers. As we explained in the Introduction, we assume distributed subcarrier grouping as opposed to adjacent grouping. Therefore each subchannel experiences the same average fading with respect to a user.

We consider a wireless downlink system, where a base station transmits to respective stations as in Figure 2.1. The noise and interference power density is N_0, and the channel gain averaged over the entire band from the BS to user i at time t is $h_i(t)$, where $h_i(t)$ includes path loss, shadowing (log-normal fading) and fast fading. Using the averaging effect of PUSC scheme we assume flat fading, i.e. we assume that fading level is the same at each subchannel.

There are three classes of users. Users in the classes U_D, U_S and U_V demand data, video and voice traffic, respectively. The system that we consider is time slotted with time slot length T_s. The scheduler makes a resource allocation decision at each time slot. Active period in a voice conversation, streaming duration and file size are both very long with respect to the time slot size. Therefore it is realistic that during the course of

optimization the number of active sessions are fixed.

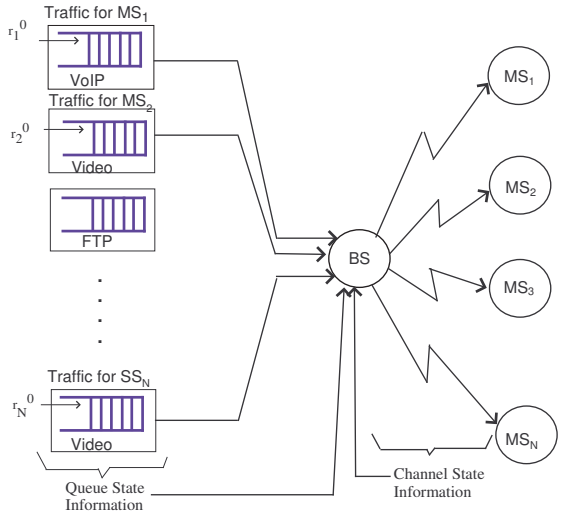

Figure 2.1: Downlink System Model

IEEE 802.16a/e standards allow several combinations of modulation and coding rates that can be used depending on the signal to noise ratio. Here assuming constant fading during a time slot, we model the channel as AWGN. Base station allocates the available power and rate among users, where $p_i(t)$ and $w_i(t)$ are the power and bandwidth allocated to user i in time slot t. For an SINR $\frac{p_i(t)h_i(t)}{N_0 w_i(t)}$, the highest order modulation and coding scheme that guarantees a BER constraint is used. We use the set of modulation coding pairs in Table 1.1.

Based on Table 1.1 it is reasonable to approximate the optimal transmission rate as an increasing and concave function of the signal to noise ratio. We will adopt the Shannon channel capacity for AWGN channel as a function for bandwidth and transmission power

assigned to user i[1]:

$$r_i(w_i(t), p_i(t)) = w_i(t) \log\left(1 + \beta \frac{p_i(t) h_i(t)}{N_0 w_i(t)}\right) \tag{2.1}$$

The reason for using Shannon capacity is its simplicity, and it also approximates rate-SINR relation in Table 1.1 with $\beta = 0.25$. The parameter $0 < \beta < 1$ compensates the rate gap between Shannon capacity and rate achieved by practical modulation and coding techniques.

2.3 Proportional Fair Resource Allocation for Data Traffic

It is proven in [1] by Tse that a proportional fair allocation for a single carrier system also maximizes the sum of the logarithms of average user rates:

$$P = \arg\max_{S} \sum_{i=1}^{N} \log R_i^{(S)} \tag{2.2}$$

where is the user set and $R_i^{(S)}$ is the average rate of user i by schedule S. In a single carrier system proportional fairness is achieved by scheduling at each time slot t, a user j according to:

$$j = \arg\max_{i} \frac{r_i(t)}{R_i(t)} \tag{2.3}$$

Here $r_i(t)$ is the instantaneous transmittable rate to user i at the current slot. $R_i(t)$ is the average data rate that user i receives over time. At each time slot the average rate is updated according to the following rule:

$$R_i(t+1) = \alpha_i R_i(t) + (1 - \alpha_i) r_i(t) \tag{2.4}$$

[1]From now on all logarithms are natural and we consider transmission in nats instead of bits

In the proportional fair scheme $T = 1/(1-\alpha_i)$ is the length of the sliding time window and average rate is computed over this time slot at each time slot. In [1] α was taken as 0.999. So this method maintains fairness in the long run, while trying to schedule the user with the best channel at each slot.

Proportional fair resource allocation problem in OFDMA systems was modeled previously in [18],[16] as follows. Maximize:

$$C(\mathbf{w},\mathbf{p}) = \sum_{i=1}^{N} \log\left(r_i(w_i,p_i)\right) \tag{2.5}$$

subject to

$$\begin{aligned} \sum_{i=1}^{N} p_i &\leq P \\ \sum_{i=1}^{N} w_i &\leq W \\ p_i, w_i &\geq 0, \forall i \end{aligned}$$

where $r_i(w_i,p_i)$ is the rate function in equation (2.1). In [18], efficient and low complexity algorithms are proposed to solve the above optimization problem. Some algorithms were also proposed for a similar model in [16]. However this formulation aims proportional fairness only in a single time slot as opposed to long term requirements.

Let us assume that there are N data users. The objective for the data users is to optimize log sum of the exponential averaged rates of the users We can model the system as a Markov decision process. The state of the system at time slot t is the vector of the averaged rates received by time $t-1$,

$$\mathbf{R}(t-1) = \left[R_1(t-1), R_2(t-1), \ldots, R_N(t-1)\right],$$

where $R(t-1) \in R^{+N}$. The control variables $\mathbf{u}(t) = (\mathbf{p}(t), \mathbf{w}(t))$ are vectors of power and bandwidth allocation at slot t denoted as $\mathbf{p}(t) = [p_1(t), p_2(t), \ldots, p_N(t)]$, $\mathbf{w}(t) = [w_1(t), w_2(t), \ldots, w_N(t)]$. The control space is denoted by $\mathcal{U}$ where $\mathcal{U} = \{\mathbf{p}, \mathbf{w} : \sum_{i=1}^{N} p_i(t) \leq P, \sum_{i=1}^{N} w_i(t) \leq W, p_i(t) \geq 0, w_i(t) \geq 0, \ \forall i\}$, where P and W are the total available power and bandwidth. The state (user rates) is updated at each time slot according to the exponential averaging formula:

$$R_i(t) = \alpha_i R_i(t-1) + (1-\alpha_i) r_i(w_i(t), p_i(t)), \forall i, t \tag{2.6}$$

where the initial state $\mathbf{R}(0)$ is a constant (possibly 0). This way we consider both current rate as well as rates given to the user in the past. Observed at time t, the highest consideration is given to the current rate $r(t)$, and the rates received at the past $t-1$, $t-1$,... carry diminishing importance. We replace the instantaneous rate $r_i(t)$ with averaged rate $R_i(t)$ in the proportional fair capacity (Equation 2.5)

$$\begin{aligned}
C(\mathbf{R}(t)) &= \sum_{i=1}^{N} \log R_i(t) \\
&= \sum_{i=1}^{N} \log \left(\alpha R_i(t-1) + (1-\alpha_i) r_i(w_i(t), p_i(t))\right) \\
&= \sum_{i=1}^{N} \log R_i(t-1) \left(\alpha_i + \frac{(1-\alpha_i) r_i(w_i(t), p_i(t))}{R_i(t-1)}\right) \\
&= C(\mathbf{R}(t-1)) + \sum_{i=1}^{N} \log \left(\alpha_i + \frac{(1-\alpha_i) r_i(w_i(t), p_i(t))}{R_i(t-1)}\right)
\end{aligned} \tag{2.7}$$

As a matter of fact, we limit ourself to *greedy* schemes in the sense that at slot t, we try to maximize the proportional fair capacity $C(\mathbf{R}(t))$ without considering the future time slots $t+1$,$t+2$, etc. Only the second term in Equation (2.7) needs consideration.

The objective for data users becomes:

$$\max_{\mathbf{p}(t),\mathbf{w}(t)} \sum_{i=1}^{N} \log\left(\alpha_i + \frac{(1-\alpha_i) r_i(w_i(t), p_i(t))}{R_i(t-1)}\right) = \max_{\mathbf{p}(t),\mathbf{w}(t)} \prod_{i}^{N} \left(\alpha_i + \frac{(1-\alpha_i) r_i(w_i(t), p_i(t))}{R_i(t-1)}\right) \quad (2.8)$$

2.4 Resource Allocation for Real Time Traffic

Our primary aim is to find a scheduling scheme that supports data traffic as well as delay sensitive traffic. Proportional fairness objective in (2.8) aims at providing fairness to data users. On the other hand, users demanding real-time traffic (voice and video) have QoS constraints on packet delay or packet drops. We assume that data traffic adjusts its transmission rate to suite its throughput (an example is TCP traffic), but it can always use any bandwidth assigned to it (its transmission queue is never empty). On the other hand, real time traffic has more strict delay and packet loss requirements. We describe below a common QoS sensitive algorithm that was commonly used in single carrier systems.

2.4.1 Benchmark Algorithm: M-LWDF-PF

In single channel systems Largest Weighted Delay First (LWDF) is shown to be throughput optimal [50]. In this scheme at each time slot the user maximizing the following quantity transmits.

$$a_i D_i^{HOL}(t) r_i(t), \quad (2.9)$$

where $D_i^{HOL}(t)$ is the head of line packet delay and $r_i(t)$ is the channel capacity of user i at time slot t. The parameter a_i is a positive constant. If QoS is defined as

$$P(D_i > D_i^{max}) < \delta_i, \tag{2.10}$$

where D_i^{max} is the delay constraint and δ_i is the probability of exceeding this constraint (typically 0.05), then the constant a_i can be defined as $a_i = -\frac{\log(\delta_i)}{D_i^{max} R_i(t)}$, which is referred to as M-LWDF-PF [50] [16]. Here, $R_i(t)$ is the average received rate which is updated as in (2.6).

Filter constant α_i should be chosen such that the average received rate is detected within the delay constraint in terms of slot durations. We will use this metric in real time session selection. M-LWDF-PF can be adapted to OFDMA systems as follows. Power is distributed equally to all subchannels. Starting from the first subchannel , the subchannel is allocated to the user maximizing (2.9). Then the received rate $R(t)$ is updated according to (2.6). All the subchannels are allocated one-by-one according to this rule. We will use this allocation scheme as a benchmark in our simulations.

2.4.2 Proposed Real Time Selection and Allocation Scheme

There are two main disadvantages of M-LWDF-PF- based resource allocation. First, the power is divided equally to over subcarriers. Performance can be increased by dynamic power adjustment. Secondly, data sessions are much different than video and voice in terms of QoS requirements. Therefore it is hard to use the same metric for data and real time sessions.

2.4.2.1 User Selection

We first choose the voice and video streaming sessions to be served in the current slot. For the data users our algorithm (which will be defined shortly) inherently selects some users and give zero rate to others. We use the following user satisfaction value for real time sessions:

$$USV_i(t) = L_i D_i^{HOL} \log\left(1 + \frac{\beta Ph_i(t)}{N_0 W}\right) \frac{r_i^0}{R_i(t)} \tag{2.11}$$

The user satisfaction metric that we use is very similar to M-LWDF-PF metric explained above except the $\frac{r_i^0}{R_i(t)}$ part at the end of the expression. If we don't use the traffic rate r_i^0 at the nominator then low-rate sessions such as VoIP gains excessively favored. Here $L_i = -\frac{\log(\delta_i)}{D_i^{max}}$, where D_i^{max} is the delay requirement of user i.

We use a simple formula to determine the fraction $F_R(t)$ of real time users scheduled in each time slot,

$$F_R(t) = \frac{1}{|U_S| + |U_V|} \sum_{i \in U_S \cup U_V} I(q_i(t) > 0.5 D_i^{max} r_i^0) \tag{2.12}$$

Here $0.5 D_i^{max} r_i^0$ denotes a queue size threshold in bits and $I(.)$ is the indicator function taking value one if the argument inside is true. As more users exceed this threshold, more fraction of real time users are scheduled. Let U_V' and U_S' be the set of voice and streaming users chosen at the current time slot and U_R' be the total set of chosen real time users . Next, we describe the joint power and bandwidth allocation that is performed on these chosen users.

2.4.2.2 Rate Allocation

The rate constraint for a chosen real time session is defined as:

$$r_i^c(t) = \frac{q_i(t)}{D_i^{max} 0.5\omega_i(t)},\ i \in U_R' \tag{2.13}$$

Here $q_i(t)$ is the queue size and $\omega_i(t)$ is the transmission frequency of user i, which is updated as follows:

$$\omega_i(t) = \alpha_i\omega_i(t-1) + (1-\alpha_i)I(r_i(t) > 0), \tag{2.14}$$

where $I(r_i(t) > 0)$ is the function that takes value one if the node receives packets in time slot t, zero otherwise. Therefore this frequency decreases if the node transmits less and less frequently. Using this frequency expression in the rate function, we compensate for the lack of transmission in the previous time slots possibly due to bad channel conditions. Choosing the rate requirement this way, we aim to empty out the current content in the queue in half duration of delay constraint.

2.5 Joint Data and Real Time Resource Allocation - FQPSA

We combine the proportional fair scheduling in (2.8) and real time user selection and rate definition in (2.11), (2.13) and propose a Fair and QoS-based Power and Subchannel Allocation (FQPSA).

We formulate a constrained optimization problem where the objective function is Equation (2.8) and the constraints are the power/bandwidth constraints and the rate requirements for chosen real time sessions defined in (2.13). Let $n_i = \frac{N_0}{\beta h_i}$. The resulting

optimization problem is as follows: [2].Find:

$$(\mathbf{p}^*, \mathbf{w}^*) = \arg\max_{\mathbf{p},\mathbf{w}} \prod_{i \in U_D} \left(\alpha_i + \frac{(1-\alpha_i) w_i \log\left(1 + \frac{p_i}{n_i w_i}\right)}{R_i} \right) \tag{2.15}$$

subject to,

$$\sum_{i \in U_D \cup U_R'} p_i^* \leq P \tag{2.16}$$

$$\sum_{i \in U_D \cup U_R'} w_i^* \leq W \tag{2.17}$$

$$w_i^* \log\left(1 + \frac{p_i^*}{n_i w_i^*}\right) \geq r_i^c, i \in U_R' \tag{2.18}$$

$$p_i^*, w_i^* \geq 0, \forall i \in U_D \cup U_R' \tag{2.19}$$

2.5.1 Solution to the Constrained Optimization Problem

The objective function (2.15) is an increasing function of (w, p), therefore the maximum is achieved only when the constraints (2.16, 2.17, 2.18) are all met with equality. For this reason we can replace these inequalities with equalities in the discussion below.

Lemma 2.1 *The problem in (2.15),(2.16),(2.17),(2.18) and (2.19) is a convex optimization problem.*

Proof 2.1 *In Appendix A.*

Before solving this optimization problem, please note that along with the rate constraint (2.18), it is required that in (2.15)

$$\alpha_i R_i + (1-\alpha_i)\left(w_i \log\left(1 + \frac{p_i}{n_i w_i}\right)\right) > 0, \ \forall i \in U_D. \tag{2.20}$$

[2]Here p_i,w_i, n_i are the values at time t. The time index is not shown for convenience.

Actually there is no guarantee that a solution can be found that satisfies both (2.18) and 2.20. The rate requirements for real time users can be too high that it may be impossible to satisfy with the given channel conditions. Below we define the feasibility of the problem:

Definition 2.1 *A feasible set of* (p,w) *is the set of power and bandwidth vectors* $(\mathbf{w},\mathbf{p})$ *such that:*

$$\alpha_i R_i + (1-\alpha_i)\left(w_i \log\left(1+\frac{\beta p_i}{n_i w_i}\right)\right) > 0,\ \forall i \in U_D. \tag{2.21}$$

$$w_i \log\left(1+\frac{p_i}{n_i w_i}\right) \geq r_i^c, \forall i \in U_R' \tag{2.22}$$

$$\sum_{i\in U_D\cup U_R'} p_i \leq P, \sum_{i\in U_D\cup U_R'} w_i \leq W, p_i, w_i \geq 0, \forall i \in U_D \cup U_R' \tag{2.23}$$

We define a feasible problem as a problem for which the feasible set is non-empty.

To start with, we assume that the problem is *feasible*. We will discuss about how to detect infeasibility of the problem and what to do in that case in the next section. We can write the Lagrangian of the problem as [47]:

$$L(\mathbf{w},\mathbf{p},\lambda_p,\lambda_w,\lambda^r) = \prod_{i\in U_D}\left(\alpha_i + \frac{(1-\alpha_i)w_i\log\left(1+\frac{p_i}{n_i w_i}\right)}{R_i}\right) + \lambda_p\left(P - \sum_{i\in U_D\cup U_R'} p_i\right)$$
$$+\lambda_w\left(W - \sum_{i\in U_D\cup U_R'} w_i\right) + \sum_{i\in U_R'}\lambda_i^r\left(w_i\log\left(1+\frac{p_i}{n_i w_i}\right) - r_i^c\right). \tag{2.24}$$

Taking the derivatives of $L(\mathbf{p},\mathbf{w},\lambda_p,\lambda_w,\lambda^{\mathbf{r}})$ w.r.t. p_i, w_i for all users, λ_p and λ_w, and λ_i^r for all chosen real-time users we get the following:

- For users $i \in U_D$:

$$\left.\frac{\partial L(\mathbf{p},\mathbf{w},\lambda_p,\lambda_w,\lambda^r)}{\partial p_i}\right|_{(\mathbf{p}^*,\mathbf{w}^*)} = 0 \quad \Rightarrow \lambda_p = \frac{1/n_i}{\left(\frac{R_i\widetilde{\alpha}_i}{w_i} + \log\left(1+\frac{p_i^*}{n_i w_i^*}\right)\right)\left(1+\frac{p_i^*}{n_i w_i^*}\right)} \tag{2.25}$$

$$\left.\frac{\partial L(\mathbf{p},\mathbf{w},\lambda_p,\lambda_w,\lambda^r)}{\partial w_i}\right|_{(\mathbf{p}^*,\mathbf{w}^*)}=0 \;\Rightarrow \lambda_w=\frac{\left(1+\frac{p_i^*}{n_i w_i^*}\right)\log\left(1+\frac{p_i^*}{n_i w_i^*}\right)-\frac{p_i}{n_i w_i^*}}{\left(1+\frac{p_i^*}{n_i w_i^*}\right)\left(R_i\widetilde{\alpha}_i+w_i^*\log\left(1+\frac{p_i^*}{n_i w_i^*}\right)\right)} \tag{2.26}$$

where $\widetilde{\alpha}_i=\frac{\alpha_i}{1-\alpha_i}$. By dividing (2.25) with (2.25) we can write for all $i\in U_D$:

$$\frac{\lambda_w}{\lambda_p}=\Lambda_x=n_i\left((1+x_i^*)\log(1+x_i^*)-x_i^*\right), \tag{2.27}$$

where $x_i^*=\frac{p_i^*}{n_i w_i^*}$ denotes the optimal *effective* SINR, which is the SINR multiplied by the SINR gap parameter β.

- For users $i\in U_R'$:

$$\left.\frac{\partial L(\mathbf{p},\mathbf{w},\lambda_p,\lambda_w,\lambda^r)}{\partial p_i}\right|_{(\mathbf{p}^*,\mathbf{w}^*)}\Rightarrow\frac{\lambda_p}{\lambda_i^r}=\frac{1}{n_i}\frac{1}{1+\frac{p_i^*}{n_i w_i^*}} \tag{2.28}$$

$$\left.\frac{\partial L(\mathbf{p},\mathbf{w},\lambda_p,\lambda_w,\lambda^r)}{\partial w_i}\right|_{(\mathbf{p}^*,\mathbf{w}^*)}\;\Rightarrow\;\frac{\lambda_w}{\lambda_i^r}\;=\;\log\left(1+\frac{p_i^*}{n_i w_i^*}\right)\;-\;\frac{\frac{p_i^*}{n_i w_i^*}}{1+\frac{p_i^*}{n_i w_i^*}} \tag{2.29}$$

Combining equation (2.28) and (2.29)(dividing $\frac{\lambda_w}{\lambda_p}$) for all $i\in U_R'$ again gives:

$$\frac{\lambda_w}{\lambda_p}=\Lambda_x=n_i\left((1+x_i^*)\log(1+x_i^*)-x_i^*\right), \tag{2.30}$$

By writing (2.30) we can eliminate λ_i^r's from the problem. It is worth noting that we get the same relation between Λ_x/n_i and x_i for all users (Eq. (2.27) and (2.30)). At this point it is convenient to define the function $f_x(x)$ as:

$$f_x(x)=(1+x)\log(1+x)-x. \tag{2.31}$$

Then we have,

$$x_i(\Lambda_x)=f_x^{-1}(\Lambda_x/n_i),\forall i\in U_D\cup U_R'. \tag{2.32}$$

Lemma 2.2 *The following properties hold:*

1. *Effective SINR ($x_i(\Lambda_x)$) is a monotonic increasing function of Λ_x for users $i \in U_D \cup U'_R$.*
2. *If $n_i < n_j$ then $x_i(\Lambda_x) > x_j(\Lambda_x)$*
3. *If $n_i > n_j$ then $x_i(\Lambda_x)n_i > x_j(\Lambda_x)n_i$*

Proof 2.2 *The proof is in Appendix B*

For real time users we also have:

$$\left.\frac{\partial L(\mathbf{p},\mathbf{w},\lambda_p,\lambda_w,\lambda^r)}{\partial \lambda_i^r}\right|_{(\mathbf{p}^*,\mathbf{w}^*)} = 0 \quad \Rightarrow r_i^c = w_i^* \log\left(1+\frac{p_i^*}{n_i w_i^*}\right), \forall i \in U'_R \quad (2.33)$$

- For all nodes $i \in U_D \cup U'_R$

$$\left.\frac{\partial L(\mathbf{p},\mathbf{w},\lambda_p,\lambda_w,\lambda^r)}{\partial \lambda_p}\right|_{(\mathbf{p}^*,\mathbf{w}^*)} = 0 \quad \Rightarrow P = \sum_{i\in U_D\cup U'_R} p_i^* \quad (2.34)$$

$$\left.\frac{\partial L(\mathbf{p},\mathbf{w},\lambda_p,\lambda_w,\lambda^r)}{\partial \lambda_w}\right|_{(\mathbf{p}^*,\mathbf{w}^*)} = 0 \quad \Rightarrow W = \sum_{i\in U_D\cup U'_R} w_i^* \quad (2.35)$$

From Equation (2.25) and (2.27) for data users we can write:

$$\frac{\left[\Lambda_p^* - n_i(1+f_x^{-1}(\frac{\Lambda_x^*}{n_i}))R_i\widetilde{\alpha}_i\right]^+}{\log(1+f_x^{-1}(\frac{\Lambda_x^*}{n_i}))(1+f_x^{-1}(\frac{\Lambda_x^*}{n_i}))n_i} = w_i^*, i \in U_D \quad (2.36)$$

$$\frac{\left[\Lambda_p^* - n_i(1+f_x^{-1}(\frac{\Lambda_x^*}{n_i}))R_i\widetilde{\alpha}_i\right]^+ f_x^{-1}(\frac{\Lambda_x^*}{n_i})}{\log(1+f_x^{-1}(\frac{\Lambda_x^*}{n_i}))(1+f_x^{-1}(\frac{\Lambda_x^*}{n_i}))} = p_i^*, i \in U_D \quad (2.37)$$

where $\Lambda_p = 1/\lambda_p$. The $[.]^+$ operator in Equations (2.36),(2.37) guarantees that $w_i, p_i \geq 0$ for all users. Given Λ_p and Λ_x we can compute the power and bandwidth for users $i \in U_D$ using (2.36) and(2.37). Given Λ_x, we can calculate the power and bandwidth for users

$i \in U_R'$ using (2.33). Please note that just from (2.33), (2.36) and (2.37), the bandwidth and power constraints (2.17) (2.16) are *not* necessarily satisfied. We need to find the right Λ_x and Λ_p so that the power and bandwidth constraints are satisfied. Let $S_p(\Lambda_x, \Lambda_p)$ and $S_w(\Lambda_x, \Lambda_p)$ be the total bandwidth and total power corresponding to Λ_x and Λ_p:

$$S_w(\Lambda_x, \Lambda_p) = \sum_{i \in U_D} \frac{\left[\Lambda_p - n_i(1 + f_x^{-1}(\frac{\Lambda_x}{n_i}))R_i\widetilde{\alpha}_i\right]^+}{\log(1 + f_x^{-1}(\frac{\Lambda_x}{n_i}))(1 + f_x^{-1}(\frac{\Lambda_x}{n_i}))n_i} + \sum_{i \in U_R'} \frac{r_i^0}{\log(1 + f_x^{-1}(\frac{\Lambda_x}{n_i}))} \quad (2.38)$$

$$S_p(\Lambda_x, \Lambda_p) = \sum_{i \in U_D} \frac{\left[\Lambda_p - n_i(1 + f_x^{-1}(\frac{\Lambda_x}{n_i}))R_i\widetilde{\alpha}_i\right]^+ f_x^{-1}(\frac{\Lambda_x}{n_i})}{\log(1 + f_x^{-1}(\frac{\Lambda_x}{n_i}))(1 + f_x^{-1}(\frac{\Lambda_x}{n_i}))} + \sum_{i \in U_R'} \frac{r_i^0 f_x^{-1}(\frac{\Lambda_x}{n_i})n_i}{\log(1 + f_x^{-1}(\frac{\Lambda_x}{n_i}))} \quad (2.39)$$

As a result, the problem is finding Λ_a^* and Λ_p^* such that

$$S_w(\Lambda_x^*, \Lambda_p^*) = W \quad (2.40)$$

$$S_p(\Lambda_x^*, \Lambda_p^*) = P \quad (2.41)$$

using Equations (2.38) and (2.39). Note that although Λ_x and Λ_p are independent variables that determine power and bandwidth for each node, they become dependent when the power and bandwidth constraints (2.40) (2.41) need to be satisfied. Using (2.38) and (2.39), and writing $\lambda_w \sum_{U_D \cup U_R'} w_i^* + \lambda_p \sum_{U_D \cup U_R'} p_i^*$, we obtain.

$$\begin{aligned} \lambda_w^* W + \lambda_p^* P &= \sum_{i \in U_D} [1 - \lambda_p^* n_i(1 + f_x^{-1}(\frac{\Lambda_x^*}{n_i}))R_i\widetilde{\alpha}_i]^+ + \sum_{i \in U_R'} r_i^c \lambda_p^* n_i(1 + f_x^{-1}(\frac{\Lambda_x^*}{n_i})) \\ \Lambda_x^* W + P &= \sum_{i \in U_D} [\Lambda_p^* - n_i(1 + f_x^{-1}(\frac{\Lambda_x^*}{n_i}))R_i\widetilde{\alpha}_i]^+ + \sum_{i \in U_R'} r_i^c n_i(1 + f_x^{-1}(\frac{\Lambda_x^*}{n_i})) \end{aligned} \quad (2.42)$$

where $\Lambda_p = 1/\lambda_p$. Let the function $\Lambda_p^*(\Lambda_x)$ be the value of Λ_p that satisfies (2.42) for Λ_x. Let U_D' be defined as $\{i \in U_D : \Lambda_p - n_i(1 + f_x^{-1}(\Lambda_x/n_i))R_i\widetilde{\alpha}_i > 0\}$.

2.5.2 Feasibility of the Solution

In the previous section we stated that there exists a solution to the problem, if the feasible set is non-empty. The feasibility of a problem means conditions (2.21), (2.22), (2.23) are all met. We now consider how to detect an infeasible problem and what to do in that case. From Eq. (2.36) and (2.37), $\Lambda_p = 0$ corresponds to the case that no bandwidth and power is allocated to data sessions. If the problem is feasible (i.e. if the available power and bandwidth is enough to satisfy rate requirements of real time sessions), then there exists a $\Lambda_x = n_i f_x(x_i), \forall i \in U_R'$ so that the following inequalities hold:

$$\sum_{i \in U_R'} \frac{r_i^0}{\log(1 + f_x^{-1}(\frac{\Lambda_x}{n_i}))} = S_w(\Lambda_x, 0) \quad \leq \quad W, \tag{2.43}$$

$$\sum_{i \in U_R'} \frac{r_i^0 f_x^{-1}(\frac{\Lambda_x}{n_i}) n_i}{\log(1 + f_x^{-1}(\frac{\Lambda_x}{n_i}))} = S_p(\Lambda_x, 0) \quad \leq \quad P. \tag{2.44}$$

Below, we prove some properties of the functions $S_w(\Lambda_x, \Lambda_p)$ and $S_p(\Lambda_x, \Lambda_p)$ that will be useful in proving the existence and uniqueness of solution to problem (2.15-2.19).

Lemma 2.3 *The following properties hold:*

i. $S_w(\Lambda_x, \Lambda_p)$ *and* $S_p(\Lambda_x, \Lambda_p)$*, are nondecreasing functions of* Λ_p *for any* $\Lambda_x \geq 0$*. Also* $\lim_{\Lambda_p \to \infty} S_w(\Lambda_x, \Lambda_p) = \infty$ *and* $\lim_{\Lambda_p \to \infty} S_p(\Lambda_x, \Lambda_p) = \infty$*.*

ii. $S_w(\Lambda_x, \Lambda_p)$ *is a decreasing function of* Λ_x *for all* $\Lambda_p \geq 0$*. Moreover,* $\lim_{\Lambda_x \to 0} S_w(\Lambda_x, \Lambda_p) = \infty$ *and* $\lim_{\Lambda_x \to \infty} S_w(\Lambda_x, \Lambda_p) = 0$ *for all* Λ_p*.*

iii. $S_p(\Lambda_x, 0)$ *is an increasing function of* Λ_x*.*

iv. Let Λ_x^0 *be the smallest* Λ_x *value that satisfies the inequality* $S_w(\Lambda_x, 0) \leq W$*. There exists such a* $\Lambda_x^0 > 0$*. The problem is feasible if and only if* $S_p(\Lambda_x^0, 0) \leq P$*.*

v. For $\Lambda_x > \Lambda_x^0$, the derivative $\frac{d\Lambda_p^*(\Lambda_x)}{d\Lambda_x}$ is positive therefore $\Lambda_p^*(\Lambda_x)$ is an increasing function of Λ_x

vi. The following inequalities hold for $\Lambda_p^*(\Lambda_x)$ and optimal Λ_x^*:

$$\Lambda_p^*(\Lambda_x) \leq \frac{\Lambda_x W + P - \sum_{i\in U_R'} r_i^c n_i (1 + f_x^{-1}(\frac{\Lambda_x}{n_i})) + \sum_{i\in U_D} n_i (1 + f_x^{-1}(\frac{\Lambda_x}{n_i})) R_i \widetilde{\alpha}_i}{|U_D|} \tag{2.45}$$

$$\min_{i\in U_D \cup U_R'} \left\{ n_i f_x \left(\frac{P}{n_i W} \right) \right\} \leq \Lambda_x^* \leq \max_{i\in U_D \cup U_R'} \left\{ n_i f_x \left(\frac{P}{n_i W} \right) \right\} \tag{2.46}$$

vii. $S_w(\Lambda_x, \Lambda_p^*(\Lambda_x))$ is a quasiconvex function of Λ_x. Specifically, it is a decreasing function of Λ_x up to a certain point Λ_x^1 and takes a value smaller than W at that point; it is an increasing function for $\Lambda_x > \Lambda_x^1$ and takes value W at limit $\Lambda_x \to \infty$. Therefore for a feasible problem $S_w(\Lambda_x, \Lambda_p^*(\Lambda_x))$ takes value W at a unique point $\infty > \Lambda_x^* \geq \Lambda_x^0$.

Proof 2.3 *Proof in Appendix I.*

Therefore before starting the optimization we can first find Λ_x^0 in order to check for feasibility (Lemma 2.3.iv). If $S_p(\Lambda_a^0, 0) > P$, the problem is not feasible and too many users have been admitted. We will then chose a user that consumes too much power and decrease its rate.

2.6 Proposed Algorithm

Using (2.38) and (2.42) we can develop an algorithms to determine the power and bandwidth allocation. The algorithm is also able to detect infeasibility if there no solution exists.

Algorithm:

1. Compute $\Lambda_x^0 = BinarySearch_x^0()$.

2. If $S_p(\Lambda_x^0, 0) < P$ then the problem is feasible from Lemma 2.3.iii. Continue with Step 3. Otherwise the problem is not feasible.

3. $(\Lambda_x^*, \Lambda_p^*) = BinarySearch_{xp}(\Lambda_x^0)$.

4. $(w_i^* p_i^*, x_i^*) = ComputePowerBandwidth(\Lambda_x^*, \Lambda_p^*)$

Subroutine: $\Lambda_x^0 = BinarySearch_x^0()$: Find Λ_x^0 s.t $S_w(\Lambda_x, 0) = W$.

i. Choose $\Delta_x > 0$. Find the smallest integer $k > 0$ s.t. $S_w(2^k\Delta_x, 0) < W$. Set $\Lambda_x^l = 2^{k-1}\Delta_x$, $\Lambda_x^h = 2^k\Delta_x$ and $\Lambda_x^m = (\Lambda_x^l + \Lambda_x^h)/2$

ii. Iteratively compute $S_w(\Lambda_x^m, 0)$ and update $(\Lambda_x^l, \Lambda_x^h)$.

- if $|\frac{\Lambda_x^h}{\Lambda_x^l} - 1| < \varepsilon$, return $\Lambda_x^0 = \Lambda_x^m$;
- else if $S_w(\Lambda_x^m, 0) < W$, $\Lambda_x^h = \Lambda_x^m$ and $\Lambda_x^m = (\Lambda_x^h + \Lambda_x^l)/2$;
- else $\Lambda_x^l = \Lambda_x^m$ and $\Lambda_x^m = (\Lambda_x^h + \Lambda_x^l)/2$.

Subroutine $\Lambda_p^* = BinarySearch_p(\Lambda_x, \Lambda_p^l, \Lambda_p^h)$: Find $\Lambda_p^*(\Lambda_x)$ that satisfies (3.21).

i. Set $\Lambda_p^m = (\Lambda_p^l + \Lambda_p^h)/2$ and run $(\mathbf{w}, \mathbf{p}, \mathbf{x}) = ComputePowerBandwidth(\Lambda_x, \Lambda_p^m)$:

ii. Binary search:

- If $|\frac{\Lambda_p^h}{\Lambda_p^l} - 1| < \varepsilon$, return $\Lambda_p^* = \Lambda_p^m$;
- else if $\Lambda_x W + P < \sum_{i \in U_D} [\Lambda_p^m - n_i(1 + f_x^{-1}(\frac{\Lambda_x}{n_i}))R_i\widetilde{\alpha}_i]^+ + \sum_{i \in U_R'} r_i^c n_i(1 + f_x^{-1}(\frac{\Lambda_x}{n_i}))$, $\Lambda_p^h = \Lambda_p^m$ and $\Lambda_p^m = (\Lambda_p^h + \Lambda_p^l)/2$;

- else $\Lambda_p^l = \Lambda_p^m$ and $\Lambda_p^m = (\Lambda_p^h + \Lambda_p^l)/2$.

Subroutine $(\Lambda_x^*, \Lambda_p^*) = BinarySearch_{xp}(\Lambda_x^0)$: If the problem is feasible finds $(\Lambda_x^*, \Lambda_p^*)$ s.t $S_w(\Lambda_x^*, \Lambda_p^*) = W$, and $S_p(\Lambda_x^*, \Lambda_p^*) = P$.

i. Determine the upper bound on Λ_x, (Λ_x^l) using (2.46). Based on this bound, find the upper bound on Λ_p, (Λ_p^h) using (2.45) (To do this, we have to find the SINR values corresponding to Λ_x^h by the subroutine $(\mathbf{x}, \mathbf{w}, \mathbf{p}) = ComputePowerBandwidth(\Lambda_x^h, 0)$). Set $\Lambda_x^l = \Lambda_p^l = 0$ and $\Lambda_x^m = (\Lambda_x^h + \Lambda_x^l)/2$.

ii. Iteratively compute $\Lambda_p^m = BinarySearch_p(\Lambda_x^m, \Lambda_p^l, \Lambda_p^h)$, and update $(\Lambda_x^l, \Lambda_x^h, \Lambda_p^l, \Lambda_p^h)$ based on $S_w(\Lambda_x^m, \Lambda_p^m)$.

- if $|\frac{\Lambda_x^h}{\Lambda_x^l} - 1| < \varepsilon$, return $(\Lambda_x^*, \Lambda_p^*) = (\Lambda_x^m, \Lambda_p^m)$;
- else if $S_w(\Lambda_x^m, \Lambda_p^*(\Lambda_x^m)) < W'$, $\Lambda_x^h = \Lambda_x^m$, $\Lambda_x^m = (\Lambda_x^h + \Lambda_x^l)/2$ and $\Lambda_p^h = \Lambda_p^m$;
- else $\Lambda_x^l = \Lambda_x^m$, $\Lambda_x^m = (\Lambda_x^h + \Lambda_x^l)/2$ and and $\Lambda_p^l = \Lambda_p^m$.

After we find Λ_x^* and Λ_p^*, we compute the optimal SNR, bandwidth and power values for all nodes with the following subroutine:

Subroutine $(\mathbf{x}, \mathbf{w}, \mathbf{p}) = ComputePowerBandwidth(\Lambda_x, \Lambda_p)$:

i. Optimal SNR values (scaled by β) for all chosen users, x_i^*:

$$x_i^* = f_x^{-1}(\Lambda_x^*/n_i),\ \forall i \in U_D \cup U_R' \tag{2.47}$$

where $f_x(x) = (1+x)\log(1+x) - x$.

ii. Optimal bandwidth values, w_i^*:

- For $i \in U_D$:

$$w_i^* = \frac{[\Lambda_p^* - n_i(1+x_i^*)R_i\widetilde{\alpha}_i]^+}{\log(1+x_i^*)(1+x_i^*)n_i} \tag{2.48}$$

- For $i \in U_R'$:

$$w_i^* = \frac{r_i^c}{\log(1+x_i^*)} \tag{2.49}$$

iii. Optimal power values for all nodes, $p_i^*,\ i \in U_D \cup U_R'$:

$$p_i^* = n_i w_i^* x_i^* / \beta,\ \forall i \in U_D \cup U_R' \tag{2.50}$$

Proposition 2.1 *If the problem is infeasible, the Algorithm always detects it.*

Proof 2.4 *From Lemma 2.3.iv we know that if the problem is infeasible than* $S_p(\Lambda_x^0, 0) > P$, *where* Λ_x^0 *is the smallest* Λ_x *such that* $S_w(\Lambda_x, 0) \leq W$. *As a corollary of Lemma 2.3.ii we also know that* $S_w(\Lambda_x, 0)$ *is a monotonic decreasing function of* Λ_x. *Therefore we can use the subroutine BinarySearch*$_a^0()$ *in order to find* Λ_x^0 *and compute* $S_p(\Lambda_x^0, 0)$ *in order to check for feasibility of the problem.*

Proposition 2.2 *Existence of a unique solution: If the problem is feasible there exists a unique point* $(\Lambda_x^*, \Lambda_p^*(\Lambda_x^*))$ *that satisfies (2.40) and (2.41).*

Proof 2.5 *From Lemma 2.3.v* $S_w(\Lambda_x, \Lambda_p^*(\Lambda_x)) \geq W$ *for* $\Lambda_x = \Lambda_x^0$ *and from Lemma 2.3.vii* $S_w(\Lambda_x, \Lambda_p^*(\Lambda_x))$ *is a monotonically decreasing function of* Λ_x. *Hence the problem has a unique solution.*

Figure 2.2 illustrates the characteristics of the sum-powers $S_p(\Lambda_x, \Lambda_p^*(\Lambda_x))$ and sum-bandwidth $S_w(\Lambda_x, \Lambda_p^*(\Lambda_x))$ versus Λ_x for 30 data users for the case of $R_i = 0,\ \forall i \in U_D$ and $R_i > 0,\ \forall i \in U_D$ at one point in time. From the graph we see that indeed sum-power

and sum-bandwidth crosses the power and bandwidth constraints at one point, which is the unique optimal solution. For the case of non-zero received rates, we observe discontinuities in sum-power and bandwidth functions. This is because at each point of discontinuity, the expression $\Lambda_p^*(\Lambda_x) - n_i(1 + f_x^{-1}(\frac{\Lambda_x}{n_i}))R_i\widetilde{\alpha}_i$ changes sign for one of the nodes $i \in U_D$.

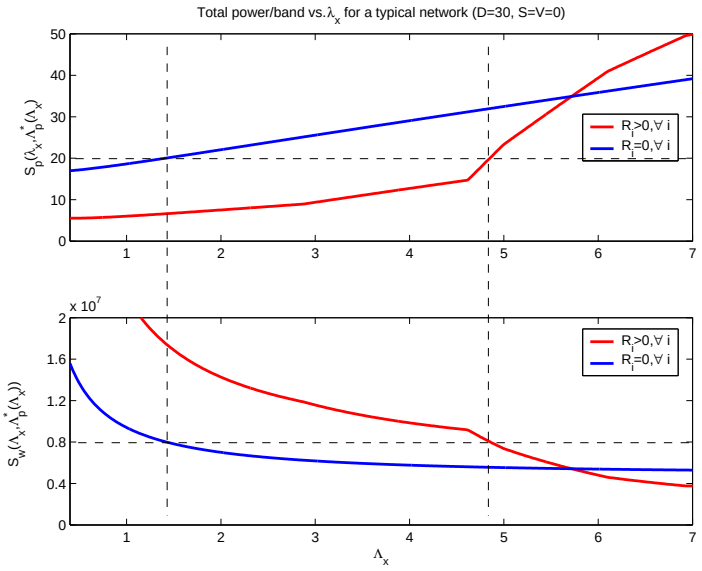

Figure 2.2: Existence of a Solution

Proposition 2.3 *Convergence of the algorithm to the unique solution: The Algorithm converges to the globally optimum solution to the set of Equations (2.40) and (2.41).*

Proof 2.6 *In Appendix II, we prove that the objective function in (2.15) is a strictly concave function of both power and bandwidth for all users. We also prove in Appendix II that the constraint set defined in (2.17), (2.16), (2.18) defines a convex set. Together this means there exists a unique local maximum of the optimization problem which is also the*

global maximum. As Equations (2.38) and (2.39) define the local maximum of the problem, its solution is the sole maximum of the problem which is also the global maximum.

2.6.1 SINR/Bandwidth Quantization and Reshuffling

In practice, bandwidth allocation is in terms of integer number of subchannels. As mentioned before there also exists a set of modulation/coding pairs and corresponding SINR thresholds, which also requires power reshuffling to quantize the SINR. Hence, we have to apply the following resource shuffling procedure

1. Quantize the bandwidth values to the nearest number of subchannel. Quantize to 1 subchannel if it is less than that. For the real time sessions recompute the power value that satisfies rate constraint.

2. Quantize the SINR values to the nearest one in Table 1.1. Quantize to the lowest SINR if it is less than that.

3. If the total bandwidth is greater than W, then find the node that has the largest bandwidth and decrease one subchannel. Adjust powers so that SINR values remain the same. Repeat this until constraint is satisfied

4. If total power is greater than P than find the node that consumes largest amount of power and take one subchannel remaining the SINR same. Repeat this until power constraint is satisfied.

5. If total bandwidth is smaller than W, then find the node with the best channel condition and give it one more bandwidth, adjusting the power so that SINR remains

the same. This is because users with good channel condition are more dependent on bandwidth.

6. If total power is smaller than P then find the node with worst channel condition. Boost its SINR to the next level (if possible). This is because users with worse channel conditions are more dependent on power. Repeat this procedure until it is impossible to do so.

A similar resource shuffling procedure can be found in [18].

2.7 Performance Evaluation

For the numerical evalutations we divide the users to 5 classes according to the distances, 0.3,0.6,0.9,1.2,1.5 km. For instance if there are 5 voice users in the system, at each distance class a single Voice user is located. For $k \times 5$ user there are k users for each session of the same type is located at each distance point. We use the parameters in Table 2.1.

2.7.1 OFDMA-Related Parameters

Table 2.2 summarizes the OFDMA-related parameters used in this simulation and their derivations. Here FFT Size means the number of samples in the Fast Fourier Transformation. Number of used subcarriers N_{used} is smaller than N_{FFT} because the outer carriers in a subchannel does not carry modulation data.

Parameter	Value
Cell radius	1.5km
User Distances	0.3,0.6,0.9,1.2,1.5 km
Total power (P)	20 W
Total bandwidth (W)	10 MHz
Frame Length	1 msec
Voice Traffic	CBR 32kbps
Video Traffic	802.16 - 128kbps
FTP File	5 MB
AWGN p.s.d.(N_0)	-169dBm/Hz
Pathloss exponent (γ)	3.5
$\psi_{DB} \sim N(\mu_{\psi_{dB}}, \sigma_{\psi_{dB}})$	N(0dB,8dB)
Coherent Time (Fast/Slow)	(5msec/400msec.)
Pathloss(dB, d in meters)	$-31.5 - 35\log_{10} d + \psi_{dB}$

Table 2.1: Simulation Parameters

2.7.2 Performance Criteria

We will compare our algorithm with the benchmark M-WLDF algorithm with proportional fairness. Delay exceeding probability is taken as $\delta_i = 0.05$ for all users. Delay constraint for voice and video users are 0.1 and 0.4 second, respectively. For M-LWDF algorithm we assume that the delay constraint is 2 seconds and buffer length is infinite. We assume a constant HOL delay of 1 second for the data sessions. For the FPSQA algorithm resource allocation for data traffic does not depend on delay. Filter values are $\alpha_i = 0.998, 0.995, 0.98$ for data, streaming and voice sessions.

Performance criteria are as follows. We will observe the total throughput for all

Parameter	Value
Nominal Channel Bandwidth	$W = 10MHz$
FFT size	$N_{FFT} = 1024.$
Number of used Subcarriers	$N_{used} = 840.$
Sampling Factor	$n_s = F_s/W = 8/7$
Sampling Frequency	$F_s = \lfloor n \times W/8000 \rfloor \times 8000 = 11.424MHz$
Subcarrier spacing	$\Delta f = F_s/N_{FFT} = 1.1156 \times 10^4 Hz$
Used Bandwidth	$N_{used} \times \Delta f = 9.37125MHz$
Useful symbol Time	$T_b = 1/\Delta f = 89.638\mu s$
Guards Period ratio	1/8
OFDM Symbol time	$T_s = (1 + 1/8) \times T_b = 0.1008msec$
Subchannelization mode	DL-PUSC
Tones per subchannel	24
Subchannel bandwidth	$W_{sub} = 24 \times \Delta f = 267.744KHz$
Number of subchannels	$N_{sub} = 30$

Table 2.2: OFDMA-Related Parameters

data users and also the total throughput for the users at the edge (users at 1500m). For data users we will also observe total log-sum rate $C(t) = \sum_{i \in U_D} \log R_i(t)$. For real time users we will measure the 95^{th} Percentile Delay for Voice and Streaming Sessions.

2.7.3 Increasing Number of Voice Users

Figures 2.3, 2.4 show the effects of increasing the number of voiceusers. In Figure 2.3 we observe that FPQS Algorithm is better than M-LWDF algorithm in terms delay performance. With FQPSA delay for voice and video sessions stay within acceptable

bounds, while with M-LWDF, it exceeds the bounds for the user at the cell edge when $V \geq 30$. Besides, according to Figure 2.4 FQPSA provides at least 10 percent increase in total throughput. Total throughput decreases linearly with increasing number of voice users. Although 10 voice users adds up to 0.32 Mbps, adding 10 users decreases the total throughput approximately by 1.2-1.4 Mbps. This is because voice has a very strict delay requirement and a voice session may have to be transmitted despite bad channel conditions. Throughput for LWDF decreases with a little bit slower rate but that reflects to the voice and video performance negatively. Log-sum performance of FQPSA is also better than that of M-LWDF, which shows that our algorithm provides fairness. We also observed that a voice session almost always uses one subchannel, when scheduled. This is approximately the case for video sessions. Users at the edge sometimes users 2 subchannels in a slot.

2.7.4 Increasing Number of Streaming Users

In Figure 2.5 we see the effects of increasing number of video streaming users on delay. We see that our algorithm is better than M-LWDF in all criteria. 95^{th} percent delay for edge users demanding voice and video exceeds the acceptable region for $S > 40$, while for our algorithm it stays within the threshold. There is still more than 10 percent increase in total throughput. Log-sum of long term received rates is also greater. Adding 10 video users (which means 1.28Mbps) decreases the throughput by 2 Mbps on the average. The inefficiency is less compared to increasing voice users because video has a looser delay requirement.

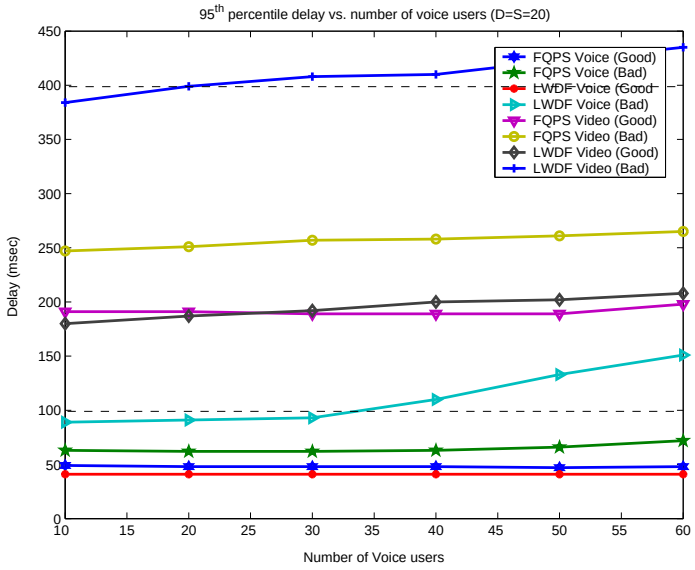

Figure 2.3: 95 percentile queue size(bits) vs. number of voice users

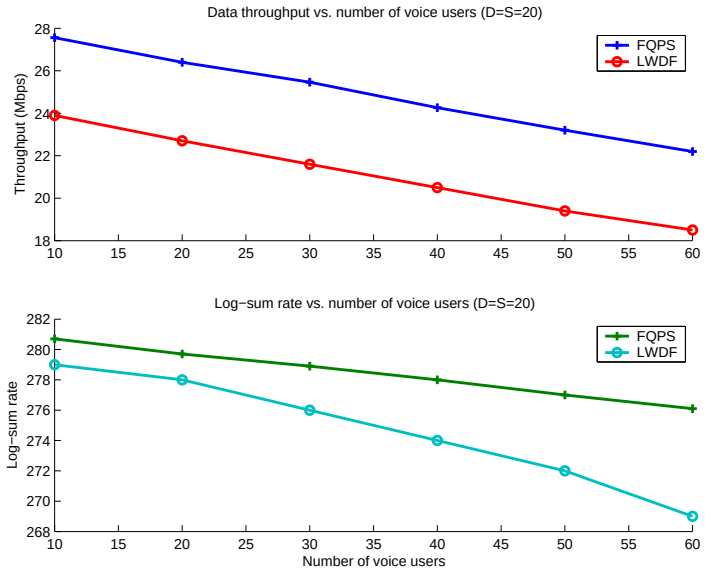

Figure 2.4: 95 percentile delay vs. number of voice users

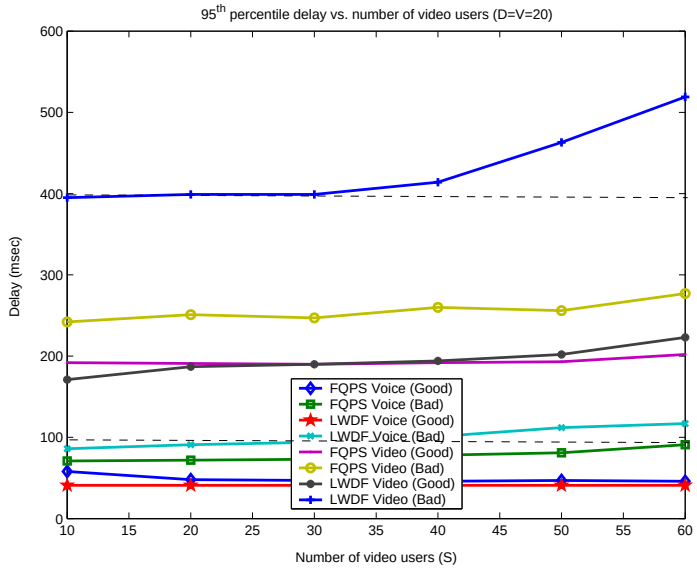

Figure 2.5: 95 percentile queue size(bits) vs. number of video users

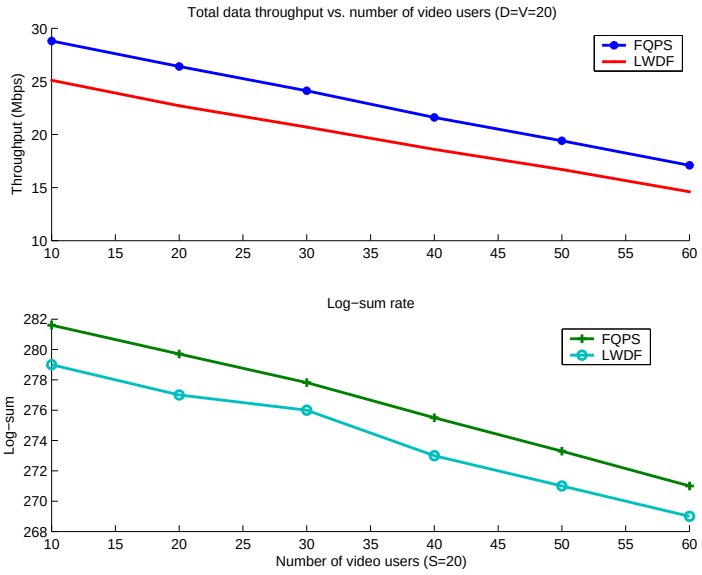

Figure 2.6: 95 percentile delay vs. number of video users

2.7.5 Increasing Number of Data Users

In Figure 2.7 and 2.8 we can observe the effects of increasing the number of data users.We observe that delay for both voice and video streaming sessions stay approximately constant. Delay performance is much better than that of M-LWDF algorithm. Data performance is 10 percent better than M-LWDF. Total throughput increases with number of data users, but the increase diminishes as D increases.

2.8 Summary

In this chapter we formulated and a resource allocation problem for OFDMA-based downlink transmission. We proposed an algorithm that converges to the unique optimal solution of the problem. Finally we numerically showed that when compared with the M-LWDF scheme, our scheme both provides better proportional fairness for data sessions and provides better QoS for real time sessions.

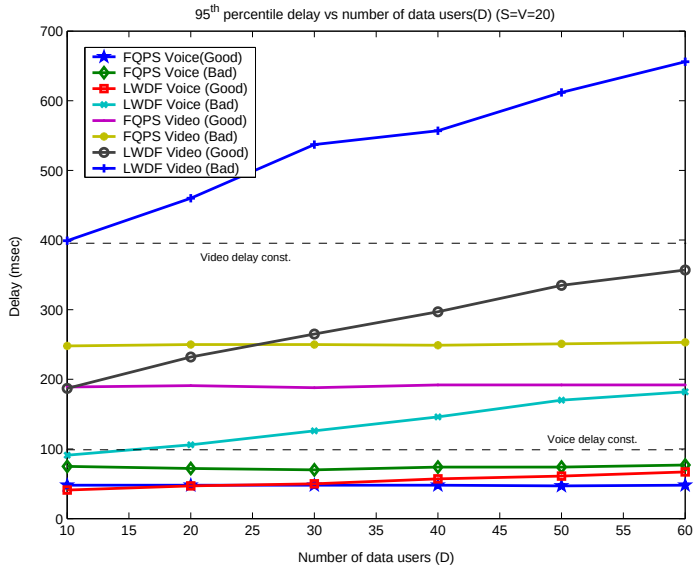

Figure 2.7: 95 percentile queue size(bits) vs. number of data users

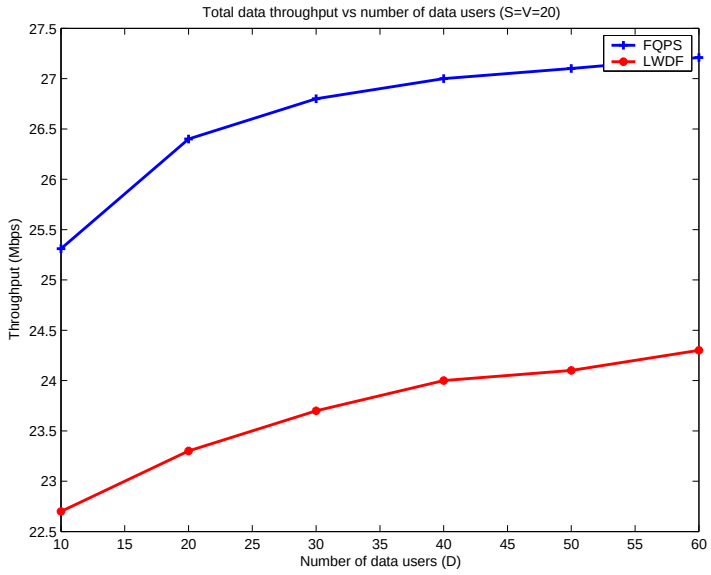

Figure 2.8: 95 percentile delay vs. number of data users

Chapter 3

Practical Scheduling of Heterogeneous Traffic in OFDMA-based Wireless Downlink Systems

3.1 Introduction

In Chapter 2 we considered the problem of resource allocation for long term proportional fairness of data sessions and satisfying QoS requirements for real time traffic. The base station allocates available power and bandwidth to individual users based on long term average received rates, QoS constraints and channel conditions. Although the proposed scheme in Chapter 2 is theoretically sound, the complexity of the algorithm motivates us in finding a simpler version of it.

In the proposed algorithm in Chapter 2, although few of the data sessions transmitted most of the time, the algorithm had to involve all data nodes in the computation and perform the look-up table operation for all data nodes at every step of the binary search. In this Chapter we add two new steps such as data user selection and minimal resource allocation. We select only a fraction of data users. We formulate and solve a proportional fair resource allocation problem for the selected data and video users subject to minimum rate requirements for video users. For selected voice users we calculate a minimal resource and exclude them from the optimization. At this point we distinguish video sessions from voice sessions in terms of elasticity and give them chance to get more rates depending

on their channel conditions. The rest of the chapter is organized as follows: In Section 3.2 we explain the system model. In Section 3.3 we describe the user selection for data and real time sessions and rate requirement determination process for the real time sessions. In Section 3.4 we formulate the problem of joint power and bandwidth allocation. Section 3.5 consists of algorithm description. Finally, we evaluate the performance of the proposed algorithm numerically in Section 3.6 and conclude the chapter.

3.2 System Model

We consider a cellular system consisting of a base station transmitting to N mobile users. Time is slotted and at each time slot base station allocates the total bandwidth W and total power P among the users. In the simulations we keep the users fixed, however we simulate mobility by fast and slow fading. Fast fading is Rayleigh distributed and slow fading is log-normal distributed. Total channel gain is the product of distance attenuation, fast and slow fading. Let $h_i(t)$ be the channel gain of user i at time t. For an AWGN channel with noise p.s.d. N_0, signal to interference plus noise ratio (SINR) is,

$$SINR_i = \frac{p_i(t)h_i(t)}{N_0 w_i(t)}, \tag{3.1}$$

where $p_i(t)$ and $w_i(t)$ are the power and bandwidth allocated to user i at time t. The BS uses a set of modulation and coding (convolutional coding and repetitions) corresponding to certain SINR thresholds defined in Table 1.1.

In order to allocate resources in a fair manner we will solve a constrained optimization problem. In that formulation, we will use the following rate function.

$$r_i(t) = w_i(t)\log\left(1+\beta\frac{p_i(t)h_i(t)}{N_0 w_i(t)}\right), \tag{3.2}$$

The network can support different traffic types such as real time (VoIP), video streaming, data applications with some rate requirements (FTP) and best effort traffic. We assume that each user demands a single type of traffic. We will consider the following traffic types:

1. FTP: FTP traffic consists of a sequence of file transmissions separated by random reading times. File sizes are on the order of megabytes. In the simulations we will consider transmission of a single file and will make a full buffer assumption, that is, there will always be unlimited number of packets to transmit throughout the simulation. FTP traffic is typically non real time, which has a minimum rate requirement.

2. Video Streaming: A video session consists of video frames arriving at regular intervals There are a fixed number of packets (slices) at each frame. Each packet in a frame consists of a random number of bytes. Video traffic has a minimum rate requirement. As long as this minimum rate requirement is satisfied, the excess traffic can be treated equally as FTP and Web traffic.

3. VoIP: A VoIP session consists of a stream of packet arrivals with deterministic interarrival time and fixed packet lengths. Therefore satisfying the minimum rate requirement is enough for such traffic types.

We classify the traffic into two groups as elastic and non-elastic traffic. BE traffic is elastic, that is, a BE user can use any available traffic. Fairness and throughput are the performance objectives for BE traffic. Proportional fairness provides a good balances between the two. Voice traffic is non-elastic; it is a CBR traffic with strict delay

requirements. If a voice user can receive its short term required rate level, it doesn't need excessive resources. On the other hand Video streaming traffic is in between the two types. It has a basic rate requirement with certain delay constraints, however it is possible to achieve higher quality video transmission if the user experiences good channel conditions. In this work we aim to satisfy the basic rate requirement for voice and video users, while treating excessive rate allocation for video users similarly as BE users.Typical rates for these traffic types are listed in Table 3.2.

3.3 User Selection

Our proposed scheduling algorithm consists of user selection and rate allocation. After selecting the users, the subchannels and power is allocated. We use the same user satisfaction value as in Chapter 2.

$$USV_i(t) = L_i D_i^{HOL} \log\left(1 + \frac{\beta p_i(t) h_i(t)}{N_0 w_i(t)}\right) \frac{r_i^0}{R_i(t)} \tag{3.3}$$

Here $L_i = -\frac{\log(\delta_i)}{D_i^{max}}$ and r_i^0 is the data rate requirement for user i. Let U_D, U_S and U_V be the BE, Video and Voice users. Let $U_R = U_S \cup U_V$ be the set of real time users. Let U_E and $\overline{U_E}$ be the set of users demanding elastic traffic and the rest, respectively.

In this setting the quantity or fraction of users chosen from data and real time users is also an important parameter. Choosing too much real time users gives excessive rate to those users and is bad for the data users. Choosing too much data is users both bad for real time users and it may also decrease the achievable rate. Our scheme puts the real time (streaming, voice) users and data users in separate pools. Let D, S and V be the number of data , streaming and voice users. We use a simple formula to determine the fraction

$F_R(t)$ of real time users scheduled in each time slot,

$$F_R(t) = \frac{1}{|U_S| + |U_V|} \sum_{i \in U_S \cup U_V} I(q_i(t) > 0.5 D_i^{max} r_i^0) \tag{3.4}$$

Here $0.5 D_i^{max} r_i^0$ denotes a queue size threshold in bits and $I(.)$ is the indicator function taking value one if the argument inside is true. As more users exceed this threshold, more fraction of real time users are scheduled. For data users, we simply choose a fraction of 0.2 of users. Next, we describe the joint power and bandwidth allocation that is performed on these chosen users.

3.4 Joint Power and Bandwidth Allocation

After the users are chosen, joint power and bandwidth allocation is performed. Let U_D', U_S' and U_V' be the chosen users that belong to all three traffic classes. The algorithm is as follows:

3.4.1 Basic Rate Allocation for Real Time Users

For the real time (voice, streaming) users, first the nominal SNR γ_i^0 is determined according to the uniform power per bandwidth allocation as $\gamma_i^0 = \frac{Ph_i(t)}{N_0 W}$. Then γ_i^0 is quantized by decreasing $\frac{Ph_i(t)}{N_0 W}$ to the closest SNR level in Table 1.1. If γ_i^0 is smaller than the smallest SNR level, then the ceiling is taken. Based on this nominal SINR, nominal bandwidth efficiency $S_i^0(t)$ is determined using Table 1.1. Then the rate is determined. The basic rate for real time sessions is,

$$r_i^c(q_i(t), \omega_i(t)) = \left(\frac{q_i(t)}{T_s}, \frac{r_i^0}{\omega_i(t)}, \right), \; i \in U_R' \tag{3.5}$$

Here $q_i(t)$ is the queue size and $\omega_i(t)$ is the transmission frequency of user i, which is updated as follows:

$$\omega_i(t) = \alpha_i \omega_i(t-1) + (1-\alpha_i) I(r_i(t) > 0), \tag{3.6}$$

where $I(r_i(t) > 0)$ is the function that takes value one if the node receives packets in time slot t, zero otherwise. Therefore this frequency decreases if the node transmits less and less frequently. Using this frequency expression in the basic rate function, we compensate for the lack of transmission in the previous time slots possibly due to bad channel conditions.

For the chosen real time users with non-elastic traffic ($i \in \overline{U_E} \cap U_R'$) basic resource allocation is enough to support the session. For these users we allocate the basic resource as follows, and don't include them in the rate allocation which will be defined later. First, the nominal SNR γ_i^0 is determined according to the uniform power per bandwidth allocation as $\gamma_i^0 = \frac{Ph_i(t)}{N_0 W}$. Then γ_i^0 is quantized by decreasing $\frac{Ph_i(t)}{N_0 W}$ to the closest SNR level in Section 3.2. If γ_i^0 is smaller than the smallest SNR level, then the ceiling is taken. Based on this nominal SINR, nominal bandwidth efficiency $S_i^0(t)$ (in bps/Hz) is determined again using the values above. Using this basic rate and the nominal bandwidth efficiency, basic bandwidth for non-elastic traffic is determined as $w_i^{min} = \frac{r_i^{min}(t)}{S_i^0(t)}, i \in \overline{U_E} \cap U_R'$. Then this bandwidth is quantized to a multiple of subchannel bandwidth by $w_i^{min} = \max(1, \lfloor w_i^{min} \rfloor) W_{sub}$. Minimal power for this user is then $p_i^{min} = \gamma_i^0 w_i^{min} N_0 / h_i(t)$, $\forall i \in \overline{U_E} \cap U_R'$. Hence $p_i = p_i^{min}$ and $w_i = w_i^{min}$ for these users. [1]

Let the residual power and bandwidth after non-elastic real time traffic allocations

[1] After the basic allocation, if the total bandwidth or power is greater then the available resource, the user with the largest power is chosen, bandwidth is decreased by one subchannel and the power is also decreased

be $P' = \sum_{i \in \overline{U_E} \cap U'_R} p_i^{min}$ and $W' = \sum_{i \in \overline{U_E} \cap U'_R} w_i^{min}$. For real time users with elastic traffic ($i \in U'_R \cap U_E$) we include the basic rate as a constraint in joint residual bandwidth-power allocation, which will be explained next.

3.4.2 Proportional Fair Resource Allocation for Data and Video Streaming

At this stage the residual power (P') and bandwidth (W') is allocated among the chosen users demanding elastic traffic in a proportional fair manner. The PF resource allocation problem in (3.7) is solved among the chosen streaming and data users.

Find $(\mathbf{p}^*, \mathbf{w}^*)$ such that:

$$\max_{\mathbf{p},\mathbf{w}} \prod_{i \in U_E \cap (U'_R \cup U'_D)} \left(w_i \log \left(1 + \frac{p_i}{n_i w_i} \right) \right)^{\phi_i} \tag{3.7}$$

subject to,

$$w_i \log \left(1 + \frac{p_i}{n_i w_i} \right) \geq r_i^{min}, \ \forall i \in U_E \cap U'_R \tag{3.8}$$

$$\sum_{i \in U_E \cap (U'_R \cup U'_D)} p_i \leq P' \tag{3.9}$$

$$\sum_{i \in U_E \cap (U'_R \cup U'_D)} w_i \leq W' \tag{3.10}$$

$$p_i, w_i \geq 0, \forall i \in U_E \cap (U'_R \cup U'_D) \tag{3.11}$$

Here log-sum is written as a product. The above problem is a convex optimization problem with a concave objective function and convex set [47]. In this optimization we also included the parameter ϕ_i, which depends on the traffic type. Since data users

in order to keep the SINR fixed. This process is continued until the total bandwidth and power for voice and video users becomes smaller than the available resources.

typically can tolerate more rate and video users are already allocated basic bandwidth, we can give higher ϕ_i for data users. We can solve this problem using the Lagrange multipliers.

$$L(\mathbf{w},\mathbf{p},\lambda_p,\lambda_w) = \sum_{i\in U_E\cap(U'_R\cup U'_D)} \log\left(w_i \log\left(1+\frac{p_i}{n_i w_i}\right)\right)^{\phi_i} + \sum_{i\in U_E\cap(U'_R\cup U'_D)} \lambda_r^i \left(w_i \log\left(1+\frac{p_i}{n_i w_i}\right) - r_i^{min}\right) + \lambda_p \left(P' - \sum_{i\in U_E\cap(U'_R\cup U'_D)} p_i\right) + \lambda_w (W' - \sum_{i\in U_E\cap(U'_R\cup U'_D)} w_i) \quad (3.12)$$

$$\left.\frac{\partial L(\mathbf{p},\mathbf{w},\lambda_p,\lambda_w)}{\partial p_i^*}\right|_{(\mathbf{p}^*,\mathbf{w}^*)} = 0 \quad \Rightarrow \lambda_p = \frac{1/n_i\left(\phi_i + \lambda_r^i w_i^* \log\left(1+\frac{p_i^*}{n_i w_i^*}\right)\right)}{w_i^* \log\left(1+\frac{p_i^*}{n_i w_i^*}\right)\left(1+\frac{p_i^*}{n_i w_i^*}\right)} \quad (3.13)$$

$$\left.\frac{\partial L(\mathbf{p},\mathbf{w},\lambda_p,\lambda_w)}{\partial w_i}\right|_{(\mathbf{p}^*,\mathbf{w}^*)} = 0 \quad \Rightarrow \lambda_w = \left(\frac{1}{w_i^*} - \frac{\frac{p_i^*}{n_i w_i^*}}{w_i^*\left(1+\frac{p_i^*}{n_i w_i^*}\right)\left(1+\frac{p_i^*}{n_i w_i^*}\right)}\right) \times \left(\phi_i + \lambda_r^i w_i^* \log\left(1+\frac{p_i^*}{n_i w_i^*}\right)\right) \quad (3.14)$$

By dividing (3.14) to (3.13) we can write for all $i \in U_E \cap (U'_R \cup U'_D)$:

$$\frac{\lambda_w}{\lambda_p} = \Lambda_x = n_i\left((1+x_i^*)\log(1+x_i^*) - x_i^*\right), \quad (3.15)$$

where $x_i^* = \frac{p_i^*}{n_i w_i^*}$ denotes the optimal *effective* SINR, which is the SINR multiplied by the SINR gap parameter β. Let's define function $f_x(x) = (1+x)\log(1+x) - x$. This is an increasing convex function as proved in Lemma 2.2. For given λ_w, λ_p, we can find the SINR from $f_x^{-1}(\frac{\lambda_w}{\lambda_p n_i})$.

Combining Equations (3.13) and (3.14), and denoting $\Lambda_p = 1/\lambda_p$ we also write,

$$\frac{\phi_i + \lambda_r^{i*} w_i^* \log\left(1+\frac{p_i^*}{n_i w_i^*}\right) - \Lambda_p^* p_i^*}{w_i^*} = \lambda_w^* \quad (3.16)$$

Please note that from Kuhn-Tucker conditions $\lambda_r^{i*}\left(w_i^*\log\left(1+\frac{p_i^*}{n_i w_i^*}\right)-r_i^{min}\right)=0$, therefore if $\lambda_r^i>0$, then $w_i^*\log\left(1+\frac{p_i^*}{n_i w_i^*}\right)=r_i^{min}$. Carrying w_i^* to the right hand side and adding (3.16) for all $i\in U_E\cap(U_R'\cup U_D')$ and using the power and bandwidth constraints we get,

$$\sum_{i\in U_E\cap(U_R'\cup U_D')}\phi_i+\sum_{i\in U_E\cap(U_R'\cup U_D')}\lambda_r^{i*}r_i^{min}=\lambda_w^*W'+\lambda_p^*P' \tag{3.17}$$

Using (3.16) we can write,

$$\lambda_r^{i*}=\left[\lambda_p\left(1+f_x^{-1}(\frac{\Lambda_x}{n_i})\right)n_i-\frac{\phi_i}{r_i^{min}}\right]^+ \tag{3.18}$$

Finally, using (3.13), (3.15) and (3.18) we can write the sum-bandwidth and using the relation $p_i=w_i n_i f_x^{-1}(\frac{\Lambda_x}{n_i})$ we can write the sum-power in terms of Λ_x,

$$S_w(\Lambda_x,\Lambda_p)=\sum_{i\in U_E\cap(U_R'\cup U_D')}w_i=\sum_{i\in U_E\cap(U_R'\cup U_D')}\frac{\max\left(\phi_i\Lambda_p,\left(1+f_x^{-1}(\frac{\Lambda_x}{n_i})\right)n_i r_i^{min}\right)}{\Lambda_x+n_i f_x^{-1}(\frac{\Lambda_x}{n_i})} \tag{3.19}$$

$$S_p(\Lambda_x,\Lambda_p)=\sum_{i\in U_E\cap(U_R'\cup U_D')}p_i=\sum_{i\in U_E\cap(U_R'\cup U_D')}\frac{\max\left(\phi_i\Lambda_p,\left(1+f_x^{-1}(\frac{\Lambda_x}{n_i})\right)n_i r_i^{min}\right)f_x^{-1}(\frac{\Lambda_x}{n_i})}{\frac{\Lambda_x}{n_i}+f_x^{-1}(\frac{\Lambda_x}{n_i})} \tag{3.20}$$

Let us define the function $\Lambda_p^*(\Lambda_x)$ as the relation between Λ_p and Λ_x that satisfies both sum-power and sum-bandwidth conditions. Using (3.19) and (3.20) we can write,

$$\sum_{i\in U_E\cap(U_R'\cup U_D')}\max\left(r_i^{min}\left(1+f_x^{-1}(\frac{\Lambda_x}{n_i})\right)n_i,\Lambda_p^*(\Lambda_x)\phi_i\right)=\Lambda_x W'+P' \tag{3.21}$$

Using (3.19) and (3.21) we can find the optimal values of Lagrange multipliers. Please note that for the special case of $r_i^{min}=0,\forall i\in U_E\cap(U_R'\cup U_D')$ we can write $\frac{\Lambda_x^*}{n_i}=\frac{P}{Wn_i}\left(\Lambda_p^*\frac{\Phi}{P}-1\right)$, where $\Phi=\sum_{i\in U_D'\cup U_S'}\phi_i$. Let us define function $\Lambda_p^*(\Lambda_x)$ that gives the relationship between Λ_p and Λ_x based on Equation (3.21).

$$\Lambda_p(\Lambda_x) \;=\; \frac{\Lambda_x W'+P'-\sum_{r_i=r_i^{min}}r_i^{min}n_i(1+f_x^{-1}(\frac{\Lambda_x}{n_i}))}{\sum_{r_i>r_i^{min}}\phi_i} \tag{3.22}$$

Lemma 3.1 *The following properties hold:*

1. $\Lambda_p(\Lambda_x) < \frac{\Lambda_x W' + P'}{\Phi}, \forall \Lambda_x$, *where* $\Phi = \sum_{i \in U_E \cap (U'_R \cup U'_D)} \phi_i$.

2. $\Lambda_p(\Lambda_x)$ *is an increasing function of* Λ_x *for* $\Lambda_x > \Lambda_x^0$. $\Lambda'_p(0) = \frac{W'}{\sum_{i \in U_E \cap (U'_R \cup U'_D)} \phi_i}$.

3. *Let* Λ_x^0 *satisfy the equality* $W' = S_w(\Lambda_x, 0)$. *If* $S_w(\Lambda_x, 0) > P'$ *then the problem is infeasible, that there is no* (Λ_x, Λ_p) *that solves both (3.19) and (3.20).*

4. *As* Λ_x *goes to zero,* $\Lambda_p(\Lambda_x)$ *goes to* $\frac{P'}{\Phi}$. *In this case* $S_w(\Lambda_x, \Lambda_p^*(\Lambda_x))$ *goes to infinity and* $S_p(\Lambda_x, \Lambda_p^*(\Lambda_x))$ *goes to* P'. *On the other hand as* Λ_x *goes to infinity,* $S_p(\Lambda_x, \Lambda_p^*(\Lambda_x))$ *goes to infinity and* $S_w(\Lambda_x, \Lambda_p^*(\Lambda_x))$ *goes to* W'.

Proof 3.1 1. *From equation (3.21) we can write that*

$$\max\left(r_i^{min}\left(1 + f_x^{-1}(\frac{\Lambda_x}{n_i})\right) n_i, \Lambda_p^*(\Lambda_x)\phi_i\right) \geq \Lambda_p^*(\Lambda_x)\phi_i$$

for all i. Therefore we can write

$$\begin{aligned} \sum_{i \in U'_D \cup U'_S} \Lambda_p^*(\Lambda_x)\phi_i &\leq \Lambda_x^* W' + P' \\ \Lambda_p^*(\Lambda_x) &\leq \frac{\Lambda_x^* W' + P'}{\sum_{i \in U'_D \cup U'_S} \phi_i} \end{aligned}$$

, hence inequality is satisfied.

2. *Using equation (3.21) we can take the derivative of* $\Lambda_p^*(\Lambda_x)$ *w.r.t.* Λ_x *and obtain the following:*

$$\frac{d\Lambda_p(\lambda_w)}{d\Lambda_x} = \frac{W' - \sum_{r_i = r_i^{min}} \frac{r_i^{min}}{\log(1 + f_x^{-1}(\Lambda_x / n_i))}}{\sum_{r_i > r_i^{min}} \phi_i} \tag{3.23}$$

As we can see the denominator of the derivative is $W' - S_w(0, \Lambda_x)$. *From the definition of feasibility for* $\Lambda_x \geq \Lambda_x^0$, $W' - S_w(0, \Lambda_x) > 0$, *which implies that the derivative is positive (function is increasing) for all* $\Lambda_x \geq \Lambda_x^0$.

3. *$S_p(\Lambda_x, \Lambda_p)$ is an nondecreasing function of Λ_x for all Λ_p. $S_w(\Lambda_x, \Lambda_p)$ is an non-increasing function of Λ_x for all Λ_p. Both $S_p(\Lambda_x, \Lambda_p)$ and $S_w(\Lambda_x, \Lambda_p)$ are nondecreasing functions of Λ_p for all Λ_x. Therefore, if $S_w(\Lambda_x^0, 0) = W'$ and $S_p(\Lambda_x^0, 0) > P'$ then from the above monotonicity properties $S_p(\Lambda_x, \Lambda_p) > P'$ for all $\Lambda_x > \Lambda_x^0$. We also know from monotonicity properties that $S_w(\Lambda_x, \lambda_p) > W'$ for all $\Lambda_x < \Lambda_x^0$ and $\Lambda_p > 0$. Therefore the problem has no solution for this case and the problem is infeasible.*

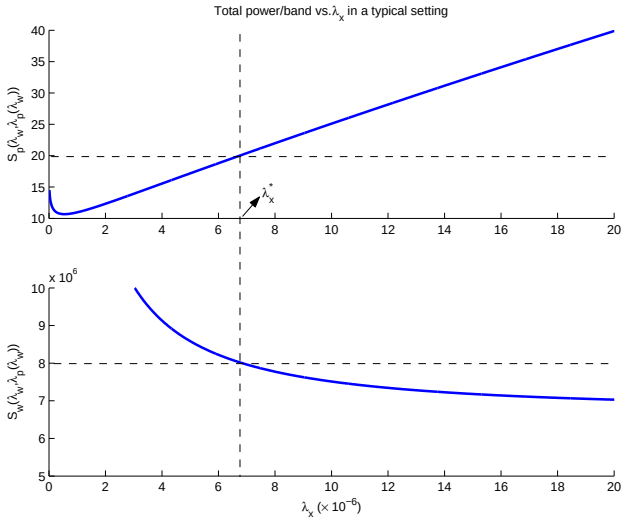

Figure 3.1: Convergence of Algorithm

3.5 Proposed Algorithm

In this section we present the algorithm that determines the power and bandwidth allocation. The algorithm is also able to detect infeasibility if there no solution exists.

Algorithm:

1. Compute $\Lambda_x^0 = BinarySearch_x^0()$.

2. If $S_p(\Lambda_x^0, 0) < P$ then the problem is feasible from Lemma 3.1.iii. Continue with Step 3. Otherwise the problem is not feasible.

3. $(\Lambda_x^*, \Lambda_p^*) = BinarySearch_{xp}(\Lambda_x^0)$.

4. $(w_i^* p_i^*, x_i^*) = ComputePowerBandwidth(\Lambda_x^*, \Lambda_p^*)$

Subroutine: $\Lambda_x^0 = BinarySearch_x^0()$: Find Λ_x^0 s.t $S_w(\Lambda_x, 0) = W$.

i. Choose $\Delta_x > 0$. Find the smallest integer $k > 0$ s.t. $S_w(2^k \Delta_x, 0) < W$. Set $\Lambda_x^l = 2^{k-1}\Delta_x$, $\Lambda_x^h = 2^k \Delta_x$ and $\Lambda_x^m = (\Lambda_x^l + \Lambda_x^h)/2$

ii. Iteratively compute $S_w(\Lambda_x^m, 0)$ and update $(\Lambda_x^l, \Lambda_x^h)$.

- if $|\frac{\Lambda_x^h}{\Lambda_x^l} - 1| < \varepsilon$, return $\Lambda_x^0 = \Lambda_x^m$;
- else if $S_w(\Lambda_x^m, 0) < W$, $\Lambda_x^h = \Lambda_x^m$ and $\Lambda_x^m = (\Lambda_x^h + \Lambda_x^l)/2$;
- else $\Lambda_x^l = \Lambda_x^m$ and $\Lambda_x^m = (\Lambda_x^h + \Lambda_x^l)/2$.

Subroutine $\Lambda_p^* = BinarySearch_p(\Lambda_x, \Lambda_p^l, \Lambda_p^h)$: Find $\Lambda_p^*(\Lambda_x)$ that satisfies (3.21).

i. Set $\Lambda_p^m = (\Lambda_p^l + \Lambda_p^h)/2$ and run $(\mathbf{w}, \mathbf{p}, \mathbf{x}) = ComputePowerBandwidth(\Lambda_x, \Lambda_p^m)$:

ii. Binary search:

- If $|\frac{\Lambda_p^h}{\Lambda_p^l} - 1| < \varepsilon$, return $\Lambda_p^* = \Lambda_p^m$;
- else if $\sum_{i \in U_D' \cup U_S'} \max\left(r_i^{min}(1 + x_i) n_i, \Lambda_p^m \phi_i\right) > \Lambda_x W' + P'$, $\Lambda_p^h = \Lambda_p^m$ and $\Lambda_p^m = (\Lambda_p^h + \Lambda_p^l)/2$;
- else $\Lambda_p^l = \Lambda_p^m$ and $\Lambda_p^m = (\Lambda_p^h + \Lambda_p^l)/2$.

Subroutine $(\Lambda_x^*, \Lambda_p^*) = BinarySearch_{xp}(\Lambda_x^0)$: If the problem is feasible finds $(\Lambda_x^*, \Lambda_p^*)$ s.t $S_w(\Lambda_x^*, \Lambda_p^*) = W$, and $S_p(\Lambda_x^*, \Lambda_p^*) = P$.

i. Choose $\Delta_x > 0$. Find the smallest integer $k > 0$ s.t. $S_w(2^k\Delta_x, \Lambda_p^*(2^k\Delta_x)) > P$. Set $\Lambda_x^h = 2^k\Delta_x$. If $k = 0$ then set $\Lambda_x^l = \Lambda_x^0$ else set $\Lambda_x^l = 2^{k-1}\Delta_x$. Set $\Lambda_x^m = (\Lambda_x^l + \Lambda_x^h)/2$. $\Lambda_p^l = BinarySearch_p(\Lambda_x^l, 0, \frac{\Lambda_x^l W' + P'}{\Phi})$ and $\Lambda_p^h = BinarySearch_p(\Lambda_x^l, 0, \frac{\Lambda_x^h W' + P'}{\Phi})$

ii. Iteratively compute $\Lambda_p^m = BinarySearch_p(\Lambda_x^m, \Lambda_p^l, \Lambda_p^h)$, and update $(\Lambda_x^l, \Lambda_x^h, \Lambda_p^l, \Lambda_p^h)$ based on $S_w(\Lambda_x^m, \Lambda_p^m)$.

- if $|\frac{\Lambda_x^h}{\Lambda_x^l} - 1| < \varepsilon$, return $(\Lambda_x^*, \Lambda_p^*) = (\Lambda_x^m, \Lambda_p^m)$;
- else if $S_w(\Lambda_x^m, \Lambda_p^*(\Lambda_x^m)) < W'$, $\Lambda_x^h = \Lambda_x^m$, $\Lambda_x^m = (\Lambda_x^h + \Lambda_x^l)/2$ and $\Lambda_p^h = \Lambda_p^m$;
- else $\Lambda_x^l = \Lambda_x^m$, $\Lambda_x^m = (\Lambda_x^h + \Lambda_x^l)/2$ and and $\Lambda_p^l = \Lambda_p^m$.

After we find Λ_x^* and Λ_p^*, we compute the optimal SNR, bandwidth and power values for all nodes with the following subroutine:

Subroutine $(\mathbf{w}, \mathbf{p}, \mathbf{x}) = ComputePowerBandwidth(\Lambda_x, \Lambda_p)$:

i. Optimal SNR values for all chosen users, x_i:

$$x_i = f_x^{-1}(\Lambda_x / n_i),\ \forall i \in U_E \cap (U_R' \cup U_D') \tag{3.24}$$

where $f_x(x) = (1+x)log(1+x) - x$. We use a look-up table to perform this operation.

ii. For $i \in U_E \cap (U_R' \cup U_D')$, bandwidth values, w_i:

$$w_i = \frac{\max\left(\phi_i \Lambda_p, (1+x_i)\, n_i r_i^{min}\right)}{\Lambda_x + n_i x_i} \tag{3.25}$$

iii. For $i \in U_E \cap (U_R' \cup U_D')$, power values, p_i:

$$p_i = n_i w_i x_i,\ \forall i \in U_D' \cup U_S' \tag{3.26}$$

3.5.1 Bandwidth and SINR quantization and Reshuffling

After the resources are allocated, first the bandwidth for data and video streaming users is quantized as $w_i = \max(1, \lfloor w_i \rfloor) W_{sub}$. Then the SINR is quantized and transmit power is determined. Unlike FTP transmission, queue size plays an important role in real time transmissions. As a result of the above optimization some streaming time sessions may get more rates than that is enough to transmit all bits in the queue. Some of the bandwidth is taken from video users in order to obey this queue constraint. After these modifications, if the total bandwidth is grater than the available, then the user with the highest power is found and its bandwidth decreased. Power is recalculated in order to keep the SINR fixed. This process is continued until bandwidth constraint is satisfied. If total power is still greater than the available then again choosing the user with highest power and decreasing bandwidth, power constraint is satisfied. If after these processes there is a leftover bandwidth, then choosing the user that has the highest channel a subchannel is added and power is increased accordingly (if there is enough power to do so). If there is some leftover power, then starting from the user with lower channel gains, SINR is boosted to the next power level (if there is enough power to do so). For the real time sessions we don't increase bandwidth or power if there isn't enough buffer content.

3.6 Numerical Evaluation

For the numerical evaluations we divide the users to 5 classes according to the distances, 0.3,0.6,0.9,1.2,1.5 km. For instance if there are 5 voice users in the system, at each distance class a single Voice user is located. For $k \times 5$ user there are k users for each

session of the same type is located at each distance point. We use the parameters in Table 3.1.

Parameter	Value
Cell radius	1.5km
User Distances	0.3,0.6,0.9,1.2,1.5 km
Total power (P)	20 W
Total bandwidth (W)	10 MHz
Frame Length	1 msec
Voice Traffic	CBR 32kbps
Video Traffic	802.16 - 128kbps
FTP File	5 MB
AWGN p.s.d.(N_0)	-169dBm/Hz
Pathloss exponent (γ)	3.5
$\psi_{DB} \sim N(\mu_{\psi_{dB}}, \sigma_{\psi_{dB}})$	N(0dB,8dB)
Coherent Time (Fast/Slow)	(5msec/300msec.)
Pathloss(dB, d in meters)	$-31.5 - 35\log_{10} d + \psi_{dB}$

Table 3.1: Simulation Parameters

We performed the simulations using MATLAB. We compared our algorithm with the benchmark M-LWDF algorithm with proportional fairness. Delay exceeding probability is taken as $\delta_i = 0.05$ for all users. The traffic and resource allocation parameters are listed in Table 3.2. Since we choose data users separately from others, the parameters L_i and head of line delay D_i^{HOL} are not used for data users.

Traffic	$r^0(kbps)$	$r^{max}(kbps)$	$D^{max}(s)$	L_i	ϕ_i	α_i
VoIP	32	32	0.1	13	-	0.98
Streaming	128	1024	0.4	3.25	1	0.995
BE	0	∞	2	0.65	-	0.998

Table 3.2: Minimum required and maximum sustained rates for different types of traffic.

The measured performance metrics are 95^{th} percentile delay for real time sessions and total throughput for data sessions. We will observe these parameters with respect to number of users for each type of sessions. For the delay, we observe the users in the range 0.3-1.2 separately as *good* users and the ones at 1.5km as *bad* users.

3.6.1 Fixed Rate Video Traffic

3.6.1.1 Increasing Number of Voice Users

In Figure 3.2 we plotted the 95 percentile delay of real time sessions vs increasing number of voice users. For this simulation we kept the number of data and Video users fixed at 20 each. We see that there is a slight increase in delay with increasing voice sessions. Delay for bad users exceeds the threshold with the M-LWDF algorithm, while for DRA they are in the acceptable range. In Figure 3.3 we see that DRA algorithm is also better in terms of total throughput. We also observe that total throughput decreases linearly with increasing real time sessions.

3.6.1.2 Increasing Number of Video Users

In Figure 3.4, we plotted the 95 percentile delays of real time sessions vs increasing number of video users. For this simulation we kept the number of data and Voice users fixed at 20. Again we observe that 95 percentile delay for video sessions increases exponentially with number video users, while delays for the users at the edge is within the acceptable range for DRA unlike M-LWDF.

In figure 3.5 we see that total data rate decreases linearly with increasing video

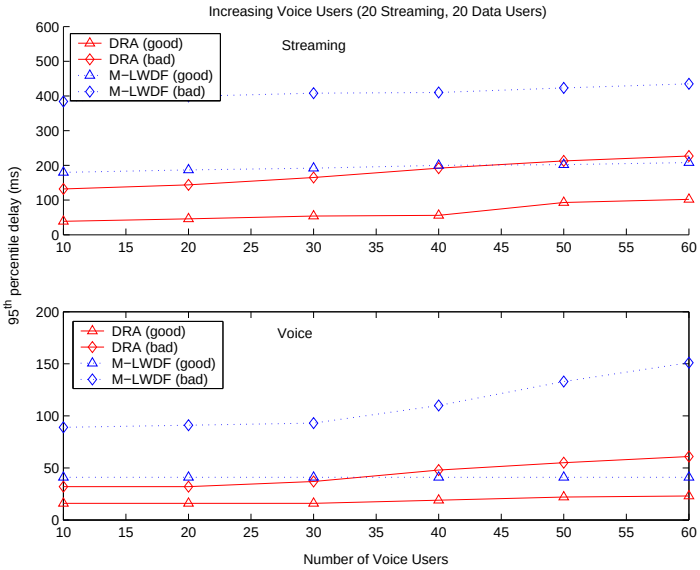

Figure 3.2: 95 percentile queue size(bits) vs. number of voice users

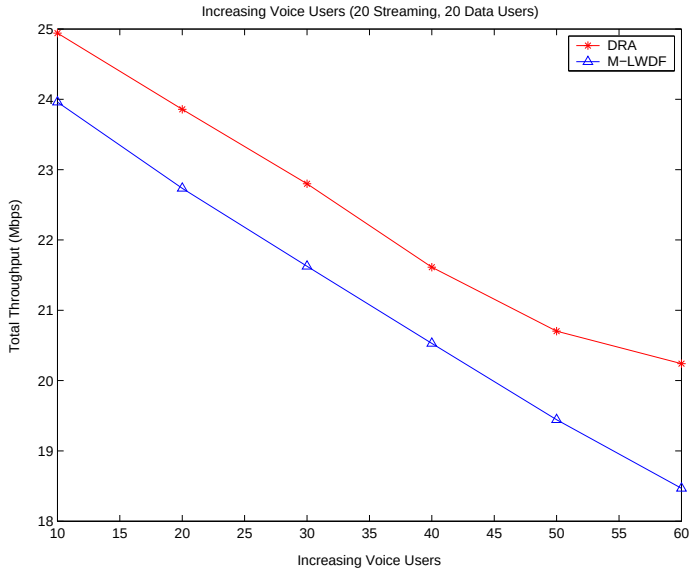

Figure 3.3: 95 percentile queue size(bits) vs. number of voice and video users

users. Data performance of DRA is again better than M-LWDF.

3.6.1.3 Increasing Number of FTP Users

In Figure 3.6, 95^{th} percentile delay for video and voice sessions are plotted for increasing number of data sessions. The number of Streaming and Voice sessions are kept fixed at 20. We observe a linear increase in the delay w.r.t. number of data sessions with M-LWDF. The delay increase is negligible for DRA.

In Figure 3.7 we see that total throughput increases as the number of FTP users increases for both algorithms. This is because of multiuser diversity. After some increase, the total throughput reaches a capacity. Capacity corresponding to DRA is approximately 10 percent higher than that of M-LWDF.

3.6.2 Elastic Video Traffic

In the second part of the simulations we considered video traffic rate that varies with packet delays. We implemented a simple rate control scheme that looks at the average head of line packet delay and increases or decreases according to a threshold policy. We defined rate levels $r_i^0\lambda_i$, ($\lambda_i \in \{1,2,\ldots,8\}$) that are integer multiples of 128kbps. Interarrival times are the same for level 1 and k, however for level k packet size is k times larger for each packet. For each user $i \in U_E \cap U_R$ and at each update instant.

- if $\overline{D_i^{HOL}}(t) < 0.125 D_i^{max}$ then $\lambda_i = \min\{\lambda_i + 1, \lambda^{max}\}$
- if $\overline{D_i^{HOL}}(t) > 0.25 D_i^{max}$ then $\lambda_i = \max\{\lambda_i - 1, 1\}$

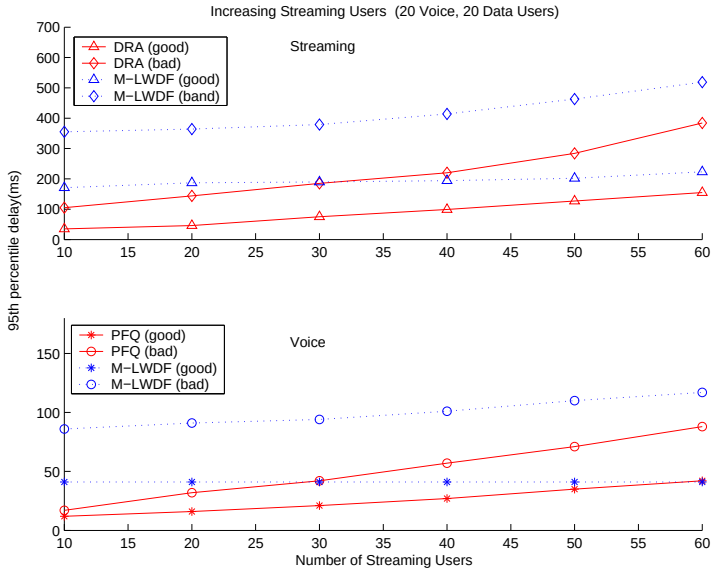

Figure 3.4: 95 percentile queue size(bits) vs. number of video users

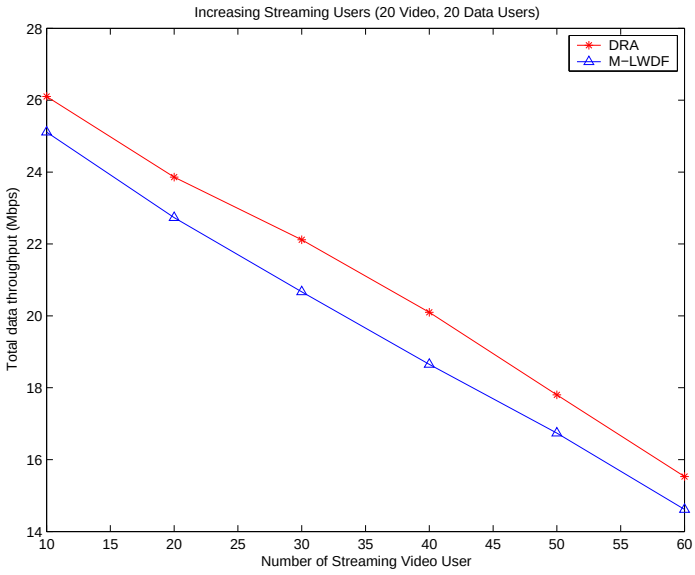

Figure 3.5: 95 percentile queue size(bits) vs. number of voice and video users

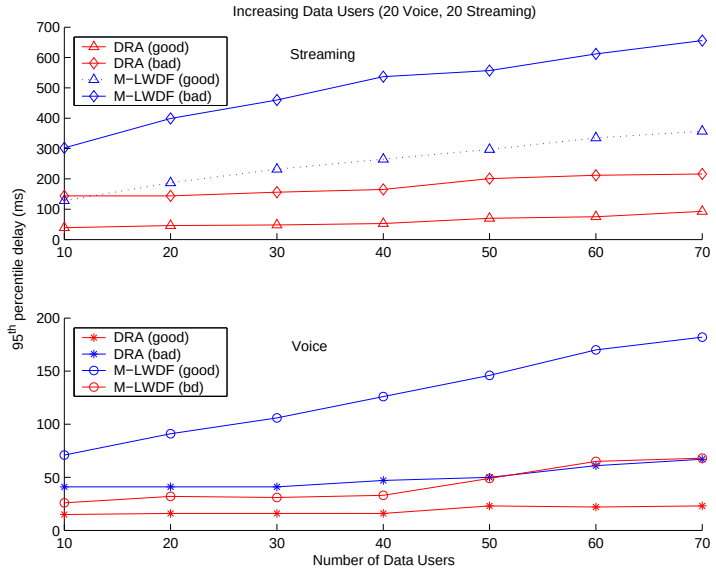

Figure 3.6: 95 percentile queue size(bits) vs. number of FTP users

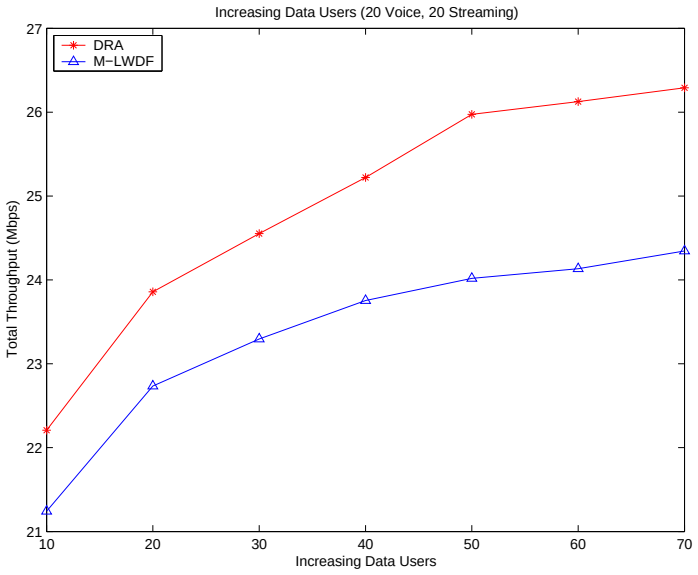

Figure 3.7: Total throughput(bps) vs. number of FTP users

Here $\overline{D_i^{HOL}}(t)$ denotes mean HOL packet delay in the last 400 frames. The updates are made at each 200 frames.

Figure 3.8 shows the evolution of rate levels along with queue sizes for video users at distances 300, 900 and 1500 meters. We observe that users closer to the BS can achieve higher rates.

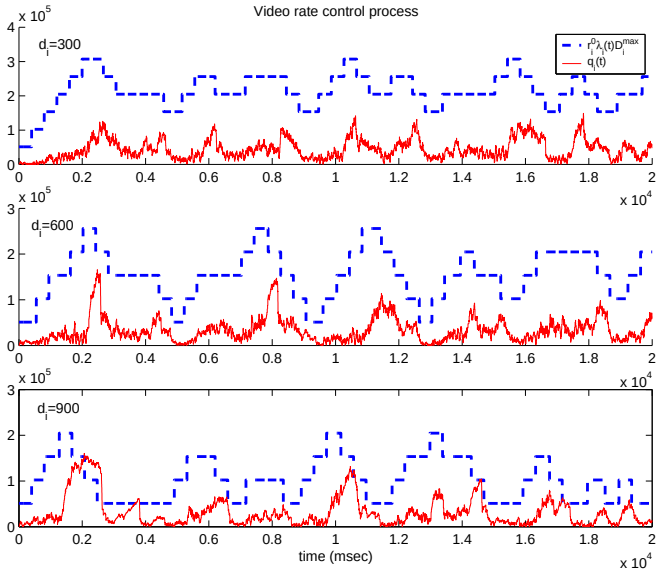

Figure 3.8: Evolution of Video rate along with queue sizes for users at 300, 600 and 900meters

In Figure 3.9 we observe the comparison of delay and throughput for the DRA and LWDF schemes.We see that DRA system satisfies delay constraints for voice users unlike LWDF. As for throughput, we see that DRA can provide significantly better throughput for video users at all distances. Total data/video throughput and log-sum throughput (proportional fairness) is also better for DRA scheme.

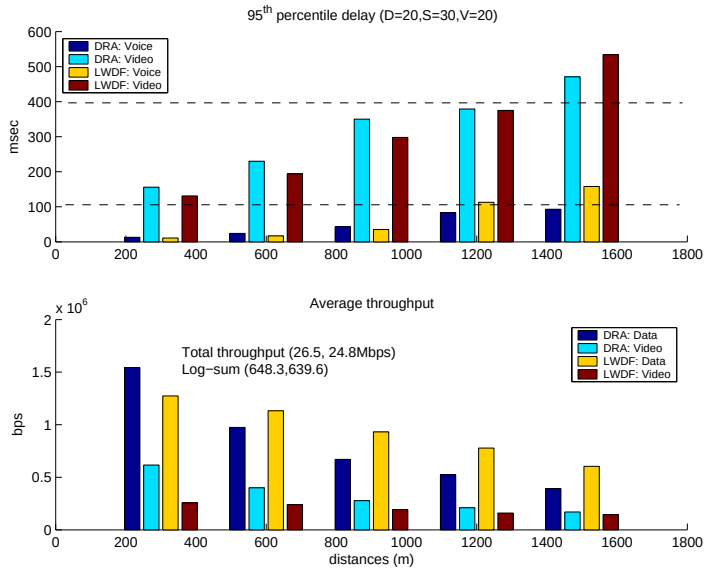

Figure 3.9: 95^{th} percentile delay and average throughput for users at different distances.

3.7 Summary

In this chapter we proposed a simpler resource allocation algorithm as an alternative to the algorithm proposed in Chapter 2. The simulation results show that the algorithm has a better performance than the benchmark algorithm and it is comparable to the one proposed in 2.

Chapter 4

Resource Allocation for Wireless Downlink System with Relays

4.1 Introduction

In Chapters 2 and 3 we considered a cellular system consisting of a single base station and mobile users. We observed that users at the cell edge often suffer from bad channel conditions and observe lower SINR. In an urban environment, big buildings pose a serious blockage to users behind and sometimes generate coverage holes. Signal penetration and attenuation inside buildings or tunnels also degrade the signal quality significantly. Often it is not possible to improve the signal qualities to these under-serviced areas by increasing the transmission power or changing the antenna configurations. Reducing the cell size and deploying more base stations will improve the situation, but this is often not possible due to limited access to traditional cell sites and wired backhaul links and the associated high operating cost. Using radio relay stations is an effective way to increase the signal quality of the users by replacing a long, low quality link between a Base Station(BS) and a Mobile Station(MS) with multiple shorter, high quality links through one or multiple Relay Stations (RS). As relay stations do not require their own wired backhauls, and are often less sophisticated than a full functional BS, relay stations are less expensive to deploy and operate than a traditional base station. The standard for relay in WiMax networks is being developed by the 802.16j Relay working group [51].

In this chapter we address the problem of OFDM based resource allocation in a

cellular network with fixed relay stations. In a realistic multihop relay network the traffic between the BS and MSs can be forwarded via multiple hops through RSs. However in this work we assume that there is at most one RS between the BS and a MS. A RS communicates to the BS like a MS, and communicates with the MS in its coverage area (called RS-*microcell*) like a BS. We describe the system model in Section 4.2. In line with recent IEEE 802.16j standard we schedule microcell transmissions in a TDMA manner in a MR frame. We first allocate a time interval of the frame to each microcell. As in the previous chapters we apply a user selection and rate requirement determination for each real time session link. We study real time rate assignment and time allocation problem in Section 4.3. In Section 4.4 we formulate a constrained optimization problem that allocates the available bandwidth, power and time to sessions in the BS-RS and RS-MS *composite links*. Our objective is to maintain proportional fairness among the data sessions in a microcell while guaranteeing required rates for real time (voice and video) sessions. We propose an algorithm that solves this problem in Section 4.5 and numerically evaluate the proposed algorithm in Section 4.6. We compare the performance of the relay network with our proposed algorithm in Chapter 2.

The use of relays in broadband cellular networks have not been studied sufficiently in the past. The existing studies involve TDMA based schemes [52],[53],[54]. In these models, transmission from the BS to RS and from RS to MS happen in consecutive equal-length time intervals. The work in [52] concentrates on a single tandem link. At each time slot either one of the queues are served and the authors formulate the problem in the context of dynamic programming and propose a link scheduling and power control scheme to jointly optimize energy expenditure and delay. In [53], the authors consider

high speed network with multiple CDMA codes and constant power. They allocate two consecutive slots to a tandem queue (fixed), therefore each link in the tandem queue transmits every other time slot. The paper only considers data communication. In [54], the authors propose a power control scheme that minimizes the interference in a relay network. Unlike all these works we propose a frame-by-frame scheduler, where in each time slot, the time slots and subcarriers in a frame are allocated to each transmission in order to optimize a QoS-based objective.

4.2 System Model and Notation

Figure 4.1 shows a typical multihop relay (MR) network. The base station is at the center, and there are a number of RSs located in the cell area. We assume that the MSs are located randomly in the cell are and they are fixed. Relay stations are also fixed and each MS is assigned to the BS or one of the RSs, based on the distance. [1]

In this work we consider frame by frame downlink resource allocation. Total frame duration is T_f seconds and it is divided into time slots of duration T_s. Total bandwidth is W Hz, which is divided into N_{sub} subchannels of W_{sub}Hz bandwidth. We assume PUSC as the subchannelization method [9], where a subchannel is formed by randomly sampling subcarriers from the entire frequency range. Because of sampling, all subchannels are of equal channel quality with respect to a user. While modeling the allocation problem we will consider time and bandwidth as a continuously divisible quantity. After finding the optimal values, we will quantize them to the integer multiples of subchannel bandwidth

[1] We consider a fixed system but simulate mobility of MSs through fast Rayleigh fading and slow Log-normal shadowing.

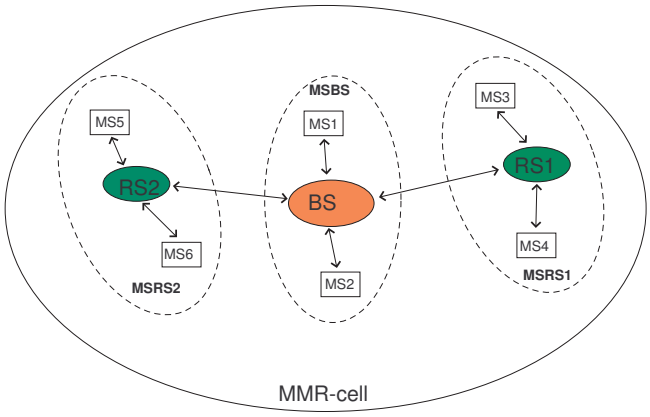

Figure 4.1: Topology of a MR cell with a BS and two relay stations (RS_1 and RS_2). The BS is serving the MSs in the set MS_{BS} directly (MS_1 and MS_2). Two relay stations (RS_1, RS_2) are used to extend the coverage of BS and serve MSs in the set MS_{RS1} (MS_3, MS_4) and MS_{RS2} (MS_5, MS_6). The MR cell includes the coverage area of the BS and all the RSs.

and time slot duration. We assume for simplicity that each user demands only one type of traffic, data, video streaming or voice. Let U_D and U_R be the set of data and real time sessions. Set of nodes assigned to RS_i is denoted as MS_{RSi} and set of nodes directly connected to BS is called MS_{BS}. This assignment is based on path loss. A node is assigned directly to the BS or one of the RSs that maximizes the received signal strength. We assume that this assignment is fixed. The BS keeps separate queues for each user, while each RS also keeps separate queues for the set of nodes MS_{RS}. We make the following definitions:

- *Microcell*: A microcell is formed by a group of MSs directly connected to a station (BS or RS). Let M-1 be the number of RSs. Including the MSs directly connected

to the BS, there are M microcells. Let MC_i be the i^{th} microcell, where MC_M denotes the microcell that contains the MSs directly connected to the BS. In the example in Figure 4.1 there are 3 microcells.

- *Composite Link*: There are three types of composite links. The set of transmissions through $BS \rightarrow RS_i$, $RS_i \rightarrow MS_{RSi}$ for all $i = 1,\ldots,M-1$ and $BS \rightarrow MS_{BS}$ are all composite links. Figure 4.2 illustrates a typical downlink frame. As seen in the figure, transmissions belonging to different composite links are scheduled in a TDMA fashion in a downlink frame. As an example in Figure 4.1, there are 5 composite links and hence the downlink frame is divided into 5 TDMA subframes.

- *Tandem queue*: A tandem queue l_j is the two cascading queues $BS \rightarrow RS_i \rightarrow MS_j$, where $j \in MS_{RSi}$. Let h_j^{BS} and h_j^{RS} be the channel gains for the links $BS \rightarrow RS_i$ and $RS_i \rightarrow MS_j$, respectively. Obviously $h_j^{BS} = h_k^{BS}$ for all $j,k \in MS_{RS_i}$, because those queues follows the same $BS \rightarrow RS_i$ link. Let q_j^{BS} and q_j^{RS} be the number of bits waiting in those queues to be transmitted.

In an MR network, bandwidth is often limited and has to be shared by the base stations and multiple relay stations to serve all the MSs in the MR-cell. We assume that a relay station has a single radio interface in order to reduce the cost, which also mandates the RS to use the same channel to communicate with the BS and with its MSs (and potentially with other RSs). Because of the single interface constraint of relay stations, transmissions $BS \rightarrow RS_i$ and $RS_i \rightarrow MS_{RSi}$ should also be scheduled in a TDMA fashion. Considering this and for simplicity we follow a TDMA approach in scheduling transmissions of each composite link.

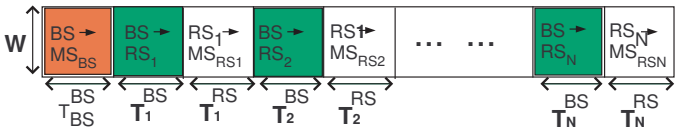

Figure 4.2: Downlink subframe for the TDD frame structure of a MR cell. BS and N RSs share the DL subframes on a TDMA basis. The order of the medium access in a DL or UL subframe is arbitrary and can be interchanged without affecting the proposed scheme. On the downlink, T_i^{BS} includes all the time slots assigned to the traffic destined from BS and RSi, while T_i^{RS} is for the traffic destined from RS_i to MS_{RSi}. Uplink subframe is just the symmetric of DL subframe.

Let P^{BS} and P^{RS} be the available power budget for the BS and each RS, respectively. We consider a channel with Rayleigh fading and Log-normal shadowing. At each time frame the channel gain is assumed constant and we consider an equivalent AWGN channel. In order to determine the bandwidth efficiency as a function of SNR, we use the values in Table 1.1.

In the problem formulation we use the following function for the number of bits that is transmitted through a link

$$r_j^{\phi} = T_i^{\phi} w_j^{\phi} \log\left(1 + \frac{\beta p_j^{\phi} h_j^{\phi}}{w_j^{\phi} N_0}\right), \phi = BS, RS, j \in MS_{RSi} \tag{4.1}$$

Here T_i^{ϕ} is the part of the frame (in seconds) that is allocated to the composite link ϕ (BS or RS) of microcell i. Let p_j^{ϕ} w_j^{ϕ} be the power and bandwidth user j in microcell i gets ($\phi = BS(RS)$ for the BS-RS (RS-MS) link). We perform resource allocation in two main steps:

1. Cellular Time Allocation: In this step we consider a TDMA scheme among composite links, where all sessions in a composite link transmit simultaneously in a T_i^{ϕ} second subframe and share the available bandwidth and power. Before performing TDMA allocation we also determine rate requirements for each real time session.

2. Microcell Resource allocation: In this step we separately perform joint power/bandwidth allocation for each $BS \rightarrow RS \rightarrow MS_{RSi}, \forall i \in MC$.

4.3 Cellular Time Allocation

In this section we consider resource allocation in a single microcell, which includes the transmissions through composite links $BS \rightarrow RS_i$ and $RS_i \rightarrow MS_{RSi}$. For the data sessions let R_j be the average transmitted rate through the tandem queue of data user $j \in MS_{RSi} \cap U_D$. For the real time sessions, let $r_j^{c,BS}$ and $r_j^{c,RS}$ be the required rates for the tandem queues of session j.

4.3.1 Real Time Session Rates

First a number of real time session links are chosen in BS-RS and RS-MS composite links to be transmitted in the current frame. We use the following user satisfaction value for real time sessions:

$$USV_j^{\phi}(t) = -\frac{\log(\delta_j)}{D_j^{max}} D_j^{\phi} \log\left(1 + \frac{\beta P^{\phi} h_j^{\phi}(t)}{N_0 W}\right) \frac{\lambda_j}{R_j^{\phi}(t)} \tag{4.2}$$

This metric resembles the Largest Weighted Delay First (LWDF) metric except the $\lambda_j / R_j^{\phi}(t)$ term at the end. Here λ_j is the bit arrival rate and $R_j^{\phi}(t)$ is the service rate for

user j. Service rate is updated as $R_j^{\phi}(t+1) = \alpha R_j^{\phi}(t) + (1-\alpha) r_j^{\alpha}(t)$, where $\phi = BS$ for the BS-RS transmission and $\phi = RS$ for RS-MS transmission. D_j^{max} and D_j^{ϕ} are the maximum allowable delay and current head of line delay for the link. δ_j is typically chosen as 0.05 and reflects the probability of exceeding the delay constraint. The BS chooses a number of real time sessions according to this metric, where U_R' denotes the set of chosen real time users. The rate constraint for a chosen real time session is defined as:

$$r_j^{c,\phi}(q_j^{\phi}(t), \omega_j^{\phi}(t)) = \max\left(\lambda_j, \frac{q_j^{\phi}(t)}{D_j^{max} 0.5 \omega_j^{\phi}(t)}\right), \ j \in U_R' \tag{4.3}$$

Here ω_j^{ϕ} is the transmission frequency of the corresponding link of user j, which is calculated as follows:

$$\omega_i(t) = \alpha \omega_j^{\phi}(t-1) + (1-\alpha) I(r_j^{\phi}(t) > 0), \tag{4.4}$$

where $I(r_j^{\phi}(t) > 0)$ is the function that takes value one if the node receives packets in time slot t, zero otherwise. Therefore this frequency decreases if the link transmits less and less frequently. Using this frequency expression in the rate function, we compensate for the lack of transmission in the previous time slots possibly due to bad channel conditions.

4.3.2 Time Allocation for each Microcell

In this section we will propose a method for allocating time intervals for each microcell. We assume uniform power allocation. By this assumption , we will be able to allocate times for each microcell in a simple manner, then with these time values we will determine the times for each composite link along with the power and bandwidth of each individual link in these composite links. Let the spectral efficiency be defined as

$S_j^{\phi} = \log\left(1+\frac{P^{\phi}}{n_j^{\phi}W}\right),\ \forall j \in U, \phi = BS, RS.$ Then the number of nats transmitted is equal to $r_j^{\phi} = T_i^{\phi} w_j^{\phi} S_j^{\phi}$ nats. We can define time-bandwidth product as $b_j^{\phi} = T_i^{\phi} w_j^{\phi}$ for $j \in MS_{RSi}$ and allocate resources subject to a time bandwidth constraint $\sum_{j\in U_D \cup U_R} b_j^{BS} + b_j^{RS} \leq WT_f$. For a real time session link j required time bandwidth product can be directly computed as $b_j^{\phi} = \frac{r_j^{c,\phi}}{S_j^{\phi}}, \phi = BS, RS.$ So we can do a resource allocation only for data sessions and subject to the constraint $\sum_{j\in U_D} b_j^{BS} + b_j^{RS} \leq (WT_f)' = WT_f - \sum_{j\in U_R} \frac{r_j^{c,BS}}{S_j^{BS}} + \frac{r_j^{c,RS}}{S_j^{RS}}$. Since we assume uniform power allocation , it becomes a much simpler task to allocate times.

$$\max_{\mathbf{b},\mathbf{r}} \sum_{j\in U_D} \log\left(\alpha_j R_j + (1-\alpha_j) r_j\right) \tag{4.5}$$

$$\sum_{j\in U_D} b_j^{BS} + b_j^{RS} \leq (WT_f)' \tag{4.6}$$

$$b_j^{\phi} S_j^{\phi} \geq r_j,\ \phi = BS, RS, \forall j U_D \tag{4.7}$$

The problem above has a concave objective function increasing in each data session rate. The constraint set is convex, hence we can solve the problem by using Lagrange multipliers.

$$L(\mathbf{b},\mathbf{r},\lambda_b,\overline{\lambda_r}) = \sum_{j\in U_D} \log\left(\alpha_j R_j + (1-\alpha_j) r_j\right) + \lambda_b((WT_f)' - \sum_{j\in U_D} b_j^{BS} + b_j^{RS}) + \sum_{\phi=BS,RS} \sum_{j\in U_D} \lambda_r^{j,\phi}\left(b_j^{\phi} S_j^{\phi} - r_j\right) \tag{4.8}$$

We won't go into details of the solution. Using similar methods as in the microcell problem solution of this problem requires a simple binary search on λ_b that solves the following equations:

$$r_j(\lambda_b) = \left[\frac{1}{\lambda_b(\frac{1}{S_j^{BS}} + \frac{1}{S_j^{RS}})} - \widetilde{\alpha} R_j\right]^+ \forall j \in U_D \tag{4.9}$$

$$WT_f = \sum_{j\in U_D} r_j(\lambda_b)\left(\frac{1}{S_j^{BS}} + \frac{1}{S_j^{RS}}\right) + \sum_{j\in U_R}\left(\frac{r_j^{c,BS}}{S_j^{BS}} + \frac{r_j^{c,RS}}{S_j^{RS}}\right) \tag{4.10}$$

As we see sum of time-bandwidth resources is a monotonic decreasing function of λ_b. Based on these result of this optimization in order to share the frame in a TDMA manner time allocated to composite links in microcell i can be computed as $T_i^{\phi}(\lambda_b) = \frac{1}{W}\sum_{j\in MS_{RSi}} r_j(\Lambda_b)b_j^{\phi}, \phi = BS, RS, \forall i \in MC$.

4.3.3 Feasibility of the Problem

The analysis above is made with a feasibility assumption. By feasibility we mean that the available resources are enough to support at least the required rates for real time sessions. Let us define $\underline{T_i^{BS}}$ and $\underline{T_i^{RS}}$ be the minimum required time to support the real time sessions in BS-RS and RS-MS links. We can find them by taking the limit $\underline{T_i^{\phi}} = \lim_{\lambda_b\to\infty} T_i^{\phi}(\lambda_b)$. Looking at the rate equation in (4.28) we see that limit $\lambda_b \to \infty$ makes the data session rates equal to zero and real time sessions are unaffected. If $\sum_{i\in MC}\underline{T_i^{BS}} + \underline{T_i^{RS}} > T_f$, then we find the non-zero-rate link with worst channel condition and change its required rate equal to zero.

4.4 Composite Link Resource Allocation

Let $\mathbf{p} = \{p_j^{BS}, p_j^{BS} | j \in MS_{RSi}\}$, $\mathbf{p} = \{w_j^{BS}, w_j^{BS} | j \in MS_{RSi}\}$ be the set of powers and rates allocated for links in this microcell and let $\mathbf{T}_i = \{T_i^{BS}, T_i^{RS}\}$ be the allocated time for $BS \to RS$ and $RS \to MS_{RS}$ transmissions. The objective is to maximize the log sum of data rates.

$$C_i(\mathbf{w}, \mathbf{p}, \mathbf{T}_i) = \sum_{j\in MS_{RSi}\cap U_D} \log\left(\alpha R_j + (1-\alpha) r_j\right) \tag{4.11}$$

The constraints are the real time sessions rate requirements defined in the previous part, and total power, rate, time constraints. The problem is formulated as follow:

$$\max_{\mathbf{w},\mathbf{p},\mathbf{T}_i} C(\mathbf{w},\mathbf{p},\mathbf{T}_i) \tag{4.12}$$

$$\begin{aligned}
s.t.\ T_i^{BS} + T_i^{RS} &\leq T_i && (4.13)\\
\sum_{j\in MS_{RSi}} p_j^{\phi} &\leq P^{\phi}\ \phi = BS, RS && (4.14)\\
\sum_{j\in MS_{RSi}} w_j^{\phi} &\leq W,\ \phi = BS, RS && (4.15)\\
T_i^{\phi} w_j^{\phi} \log\left(1 + \frac{p_j^{\phi}}{n_j^{\phi} w_j^{\phi}}\right) &\geq r_j,\ \phi = BS, RS, \forall j \in MS_{RSi} \cap U_D && (4.16)\\
T_i^{\phi} w_j^{\phi} \log\left(1 + \frac{p_j^{\phi}}{n_j^{\phi} w_j^{\phi}}\right) &\geq r_j^{0,\phi},\ \phi = BS, RS, \forall j \in MS_{RSi} \cap U_R && (4.17)
\end{aligned}$$

The problem above has a concave objective function increasing in each data session rate. The constraint set is convex, hence we can solve the problem by using Lagrange multipliers.

$$\begin{aligned}
L(\mathbf{w},\mathbf{p},\mathbf{T}_i,\overline{\lambda_p},\overline{\lambda_w},\overline{\lambda_r},\lambda_T) &= C_i(\mathbf{w},\mathbf{p},\mathbf{T}_i) + \lambda_T(T_i - T_i^{BS} - T_i^{RS})\\
&+ \sum_{\phi=BS,RS} \lambda_p^{\phi}\left(P^{\phi} - \sum_{j\in MS_{RSi}} p_j^{\phi}\right) + \sum_{\phi=BS,RS} \lambda_w^{\phi}\left(W - \sum_{j\in MS_{RSi}} w_j^{\phi}\right)\\
&+ \sum_{\phi=BS,RS}\ \sum_{j\in MS_{RSi}\cap U_D} \lambda_r^{j,\phi}\left(T_i^{\phi} w_j^{\phi} \log\left(1 + \frac{p_j^{\phi}}{n_j^{\phi} w_j^{\phi}}\right) - r_j\right)\\
&+ \sum_{\phi=BS,RS}\ \sum_{j\in MS_{RSi}\cap U_R} \lambda_j^{r,\phi}\left(T_i^{\phi} w_j^{\phi} \log\left(1 + \frac{p_j^{\phi}}{n_j^{\phi} w_j^{\phi}}\right) - r_j^{0,\phi}\right)
\end{aligned} \tag{4.18}$$

The problem can be solved by taking derivative with respect to resources and Lagrange multipliers. Since the rate is an increasing function of resources the optimal can be achieved only when all the time, power and bandwidth is used. Therefore all Lagrange multipliers are positive. Derivatives with respect to resources are as follows:

4.4.0.1 Derivative w.r.t. r_j for users $j \in MS_{RSi} \cap U_D$, $\phi = BS, RS$

$$\frac{\partial L(\mathbf{w}, \mathbf{p}, \mathbf{T}_i, \overline{\lambda_p}, \overline{\lambda_w}, \overline{\lambda_r}, \lambda_T)}{\partial r_j} = 0 \Rightarrow \left[\frac{1}{\lambda_j^{r,BS} + \lambda_j^{r,RS}} - \frac{\alpha R_j}{1-\alpha} \right]^+ = r_j \tag{4.19}$$

4.4.0.2 Derivative w.r.t. w_j^ϕ and p_j^ϕ for users $j \in MS_{RSi}$, $\phi = BS, RS$

$$\frac{\partial L(\mathbf{w}, \mathbf{p}, \mathbf{T}_i, \overline{\lambda_p}, \overline{\lambda_w}, \overline{\lambda_r}, \lambda_T)}{\partial w_j^\phi} = 0 \Rightarrow T_i^\phi \left(\log\left(1 + \frac{p_j^\phi}{n_j^\phi w_j^\phi}\right) - \frac{\frac{p_j^\phi}{n_j^\phi w_j^\phi}}{1 + \frac{p_j^\phi}{n_j^\phi w_j^\phi}} \right) = \frac{\lambda_w^\phi}{\lambda_j^{r,\phi}} \tag{4.20}$$

$$\frac{\partial L(\mathbf{w}, \mathbf{p}, \mathbf{T}_i, \overline{\lambda_p}, \overline{\lambda_w}, \overline{\lambda_r}, \lambda_T)}{\partial p_j^\phi} = 0 \Rightarrow \frac{T_i^\phi}{n_j^\phi \left(1 + \frac{p_j^\phi}{n_j^\phi w_j^\phi}\right)} = \frac{\lambda_p^\phi}{\lambda_j^{r,\phi}} \tag{4.21}$$

Dividing Eq. (4.20) to (4.21) we get the following relation:

$$\frac{\lambda_x^\phi}{n_j^\phi} = \frac{\lambda_w^\phi}{n_j^\phi \lambda_p^\phi} = \left(1 + \frac{p_j^\phi}{n_j^\phi w_j^\phi}\right) \log\left(1 + \frac{p_j^\phi}{n_j^\phi w_j^\phi}\right) - \frac{p_j^\phi}{n_j^\phi w_j^\phi} \tag{4.22}$$

Let's define $\lambda_x^\phi = \lambda_w^\phi / \lambda_p^\phi$ and the function $f_x(x) = (1+x)\log(1+x) - x$. This is a monotonic increasing and convex function. Using $\lambda_x^\phi(i)$ we can find the SINR $x_j^\phi = \frac{p_j^\phi}{n_i^\phi w_j^\phi}$ as $f_x^{-1}(\lambda_x^\phi / n_j^\phi)$.

4.4.0.3 Derivative w.r.t. T_i^ϕ, for $\phi = BS, RS$

$$\frac{\partial L(\mathbf{w}, \mathbf{p}, \mathbf{T}_i, \overline{\lambda_p}, \overline{\lambda_w}, \overline{\lambda_r}, \lambda_T)}{\partial T_i^\phi} = 0 \Rightarrow \sum_{j \in MS_{RSi}} \lambda_j^{r,\phi} w_j^\phi \log\left(1 + \frac{p_j^\phi}{n_j^\phi w_j^\phi}\right) = \lambda_T \tag{4.23}$$

Using Equations (4.23), (4.21) and (4.22) we can write,

$$\begin{aligned}
\lambda_T &= \sum_{j\in MS_{RSi}} \frac{\lambda_p^{\phi}}{T_i^{\phi}} n_j^{\phi} w_j^{\phi} \left(1+\frac{p_j^{\phi}}{n_j^{\phi} w_j^{\phi}}\right) \log\left(1+\frac{p_j^{\phi}}{n_j^{\phi} w_j^{\phi}}\right) \\
\lambda_T &= \sum_{j\in MS_{RSi}} \frac{\lambda_p^{\phi}}{T_i^{\phi}} n_j^{\phi} w_j^{\phi} \left(\frac{\lambda_w^{\phi}}{\lambda_p^{\phi} n_j^{\phi}} + \frac{p_j^{\phi}}{n_j^{\phi} w_j^{\phi}}\right) \\
\lambda_T &= \sum_{j\in MS_{RSi}} \left(\frac{w_j^{\phi}\lambda_w^{\phi}}{T_i^{\phi}} + \frac{\lambda_p^{\phi} p_j^{\phi}}{T_i^{\phi}}\right) \\
T_i^{\phi} &= \frac{W\lambda_w^{\phi} + P^{\phi}\lambda_p^{\phi}}{\lambda_T}
\end{aligned} \tag{4.24}$$

Using Equation (4.21) and (4.24) we get:

$$\begin{aligned}
\lambda_j^{r,\phi} &= \lambda_T \lambda_p^{\phi} \frac{n_j^{\phi}\left(1+f_x^{-1}(\frac{\lambda_x^{\phi}}{n_j^{\phi}})\right)}{W\lambda_w^{\phi} + P^{\phi}\lambda_p^{\phi}} \\
&= \lambda_T \frac{n_j^{\phi}\left(1+f_x^{-1}(\frac{\lambda_x^{\phi}}{n_j^{\phi}}\right)}{W\lambda_x^{\phi} + P^{\phi}}
\end{aligned} \tag{4.25}$$

Combining (4.19) and (4.25) we obtain the rate function for data users in terms of λ_x^{BS}, λ_x^{RS} and λ_T as,

$$r_j(\lambda_x^{BS},\lambda_x^{RS},\lambda_T) = \left(\frac{1}{\frac{\lambda_T n_j^{BS}\left(1+f_x^{-1}(\frac{\lambda_x^{BS}}{n_j^{BS}})\right)}{W\lambda_x^{BS}+P^{BS}} + \frac{\lambda_T n_j^{RS}\left(1+f_x^{-1}(\frac{\lambda_x^{RS}}{n_j^{RS}})\right)}{W\lambda_x^{RS}+P^{RS}}} - \widetilde{\alpha}R_j\right)^{+} \tag{4.26}$$

Taking the derivative of (4.18) w.r.t. λ_p^{ϕ} we obtain the power constraints and combining it with (4.22) we obtain,

$$P^{\phi} = \sum_{j\in MS_{RSi}} f_x^{-1}\left(\frac{\lambda_x^{\phi}}{n_j^{\phi}}\right) n_j^{\phi} w_j^{\phi},\ \phi = BS, RS \tag{4.27}$$

Since $n_j^{BS} = n^{BS}$ for $j \in MS_{RSi}$, SINR's in those links are the same using (4.22). Therefore optimal SINRs of all users in the BS-RS links are equal to $x^{BS*} = f_x^{-1}(\frac{\lambda_x^{BS*}}{n^{BS}}) =$

$\frac{P^{BS}}{n^{BS}W}$. After some arrangements in (4.26) we can write the rates of all data sessions as a function of only λ_x^{RS} and λ_T as follows:

$$r_j(\lambda_x^{RS},\lambda_T) = \left(\frac{W/\lambda_T}{\frac{1}{\log\left(1+\frac{P^{BS}}{n^{BS}W}\right)} + \frac{\left(1+f_x^{-1}(\frac{\lambda_x^{RS}}{n_j^{RS}})\right)}{\frac{\lambda_x^{RS}}{n_j^{RS}}+\frac{P^{RS}}{n_j^{RS}W}}} - \widetilde{\alpha}R_j \right)^+ \tag{4.28}$$

Rate $r_j(\lambda_x^{RS},\lambda_T)$ is a nonincreasing function of λ_x^{RS} for $0 \leq \lambda_x^{RS} \leq n_j^{RS} f_x(\frac{P^{RS}}{n_j^{RS}W})$ and it is a nondecreasing function of λ_x^{RS} for $n_j^{RS} f_x(\frac{P^{RS}}{n_j^{RS}W}) \leq \lambda_x^{RS}$. For finite λ_T it always takes finite values.

4.4.0.4 Calculation of times

Sum of the time-bandwidth products in BS-RS and RS-MS composite links is as follows,

$$\sum_{j\in MS_{RSi}} T_i^{\phi}(\lambda_x^{RS},\lambda_T) w_j^{\phi} = \sum_{j\in MS_{RSi}\cap U_D} \frac{r_j(\lambda_x^{RS},\lambda_T)}{\log\left(1+f_x^{-1}(\frac{\lambda_x^{\phi}}{n_j^{\phi}})\right)} + \sum_{j\in MS_{RSi}\cap U_R} \frac{r_j^{c,\phi}}{\log\left(1+f_x^{-1}(\frac{\lambda_x^{\phi}}{n_j^{\phi}})\right)} \tag{4.29}$$

Each node in a composite link transmits using the same time interval, but different frequency bands, where the sum of the bandwidths is equal to W. Dividing the total time-bandwidth product to W we can find the time intervals allocated to BS-RS and RS-MS

composite links:

$$T_i^{BS}(\lambda_x^{RS},\lambda_T) = \sum_{j\in MS_{RSi}\cap U_D} \frac{r_j(\lambda_x^{RS},\lambda_T)}{W\log\left(1+\frac{P^{BS}}{n^{BS}W}\right)} + \sum_{j\in MS_{RSi}\cap U_R} \frac{r_j^{c,BS}}{W\log\left(1+\frac{P^{BS}}{n^{BS}W}\right)} \quad (4.30)$$

$$T_i^{RS}(\lambda_x^{RS},\lambda_T) = \sum_{j\in MS_{RSi}\cap U_D} \frac{r_j(\lambda_x^{RS},\lambda_T)}{W\log\left(1+f_x^{-1}(\frac{\lambda_x^{RS}}{n_j^{RS}})\right)} + \sum_{j\in MS_{RSi}\cap U_R} \frac{r_j^{c,RS}}{W\log\left(1+f_x^{-1}(\frac{\lambda_x^{RS}}{n_j^{RS}})\right)} \quad (4.31)$$

Total time equation is,

$$S_T(\lambda_x^{RS},\lambda_T) = T_i^{BS}(\lambda_x^{RS},\lambda_T) + T_i^{RS}(\lambda_x^{RS},\lambda_T) \quad (4.32)$$

4.4.0.5 Calculation of total power

Sum of the powers in BS-RS and RS-MS transmissions can be found as follows:

$$S_p^{BS}(\lambda_x^{RS},\lambda_T) = \sum_{j\in MS_{RSi}\cap U_D} \frac{r_j(\lambda_x^{RS},\lambda_T)n^{BS}\frac{P^{BS}}{n^{BS}W}}{T_i^{BS}(\lambda_x^{RS},\lambda_T)\log\left(1+\frac{P^{BS}}{n^{BS}W}\right)} + \sum_{j\in MS_{RSi}\cap U_R} \frac{r_j^{c,BS}n^{BS}\frac{P^{BS}}{n^{BS}W}}{T_i^{BS}(\lambda_x^{RS},\lambda_T)\log\left(1+\frac{P^{BS}}{n^{BS}W}\right)} \quad (4.33)$$

$$S_p^{RS}(\lambda_x^{RS},\lambda_T) = \sum_{j\in MS_{RSi}\cap U_D} \frac{r_j(\lambda_x^{RS},\lambda_T)n_j^{RS}f_x^{-1}(\frac{\lambda_x^{RS}}{n_j^{RS}})}{T_i^{RS}(\lambda_x^{RS},\lambda_T)\log\left(1+f_x^{-1}(\frac{\lambda_x^{RS}}{n_j^{RS}})\right)} + \sum_{j\in MS_{RSi}\cap U_R} \frac{r_j^{c,RS}n_j^{RS}f_x^{-1}(\frac{\lambda_x^{RS}}{n_j^{RS}})}{T_i^{RS}(\lambda_x^{RS},\lambda_T)\log\left(1+f_x^{-1}(\frac{\lambda_x^{RS}}{n_j^{RS}})\right)} \quad (4.34)$$

Combining Equations (4.30),(4.31),(4.33) and (4.34) we obtain a second equation for total time,

$$T_i = \sum_{j\in U_D}\left[\frac{1}{\lambda_T^*} - \widetilde{\alpha}R_j\left(\frac{n_j^{BS}(1+f_x^{-1}(\frac{\lambda_x^{BS}}{n_j^{BS}}))}{W\Lambda_x^{BS}+P^{BS}} + \frac{n_j^{RS}(1+f_x^{-1}(\frac{\lambda_x^{RS}}{n_j^{RS}}))}{W\Lambda_x^{RS}+P^{RS}}\right)\right]^+ + \sum_{j\in U_R'}\left(\frac{r_j^{c,BS}n_j^{BS}(1+f_x^{-1}(\frac{\lambda_x^{BS}}{n_j^{BS}}))}{W\Lambda_x^{BS}+P^{BS}} + \frac{r_j^{c,RS}n_j^{RS}(1+f_x^{-1}(\frac{\lambda_x^{RS}}{n_j^{RS}}))}{W\Lambda_x^{RS}+P^{RS}}\right) \quad (4.35)$$

$$T_i = \sum_{j\in U_D}\left[\frac{1}{\lambda_T^*} - \widetilde{\alpha}R_j\left(\frac{1}{W\log(1+\frac{P^{BS}}{n^{BS}W})} + \frac{n_j^{RS}(1+f_x^{-1}(\frac{\lambda_x^{RS}}{n_j^{RS}}))}{W\Lambda_x^{RS}+P^{RS}}\right)\right]^+ + \sum_{j\in U_R'}\left(\frac{r_j^{c,BS}}{W\log(1+\frac{P^{BS}}{n^{BS}W})} + \frac{r_j^{c,RS}n_j^{RS}(1+f_x^{-1}(\frac{\lambda_x^{RS}}{n_j^{RS}}))}{W\Lambda_x^{RS}+P^{RS}}\right) \quad (4.36)$$

$$\sum_{j\in MS_{RSi}\cap U_D} r_j(\lambda_x^{RS},\lambda_T)(A_j(\lambda_x^{RS}) - B_j(\lambda_x^{RS})) + \sum_{j\in MS_{RSi}\cap U_R} r_j^{c,RS}(A_j(\lambda_x^{RS}) - B_j(\lambda_x^{RS})) = 0 \quad (4.37)$$

where $A_j(\lambda_x^{RS}) = \frac{1}{W\log\left(1+f_x^{-1}(\frac{\lambda_x^{RS}}{n_j^{RS}})\right)}$ and $B_j(\lambda_x^{RS}) = \frac{1}{W\log\left(1+f_x^{-1}(\frac{\lambda_x^{RS}}{n_j^{RS}})\right)}\frac{f_x^{-1}(\frac{\lambda_x^{RS}}{n_j^{RS}})}{\frac{P^{RS}}{n_j^{RS}W}}$ $A_j(\lambda_x^{RS})$ is a decreasing and $B_j(\lambda_x^{RS})$ is an increasing function of λ_x^{RS}.

Lemma 4.1 *Left hand side of (4.37) is a monotonic nonincreasing function of λ_x^{RS} that decreases from $+\infty$ to $-\infty$ and crosses zero at a single point.*

Proof 4.1 *We will start the analysis from a single user. For a data user j and for $\lambda_T > 0$, the function $r_j(\lambda_x^{RS}, \lambda_T)$ takes finite values for all $0 < \lambda_x^{RS}$. It is a either zero or a decreasing function of λ_x^{RS} for $\lambda_x^{RS} < n_j^{RS} f_x(\frac{P^{RS}}{n_j^{RS} W})$ and either zero or increasing function of λ_x^{RS} for $\lambda_x^{RS} > n_j^{RS} f_x(\frac{P^{RS}}{n_j^{RS} W})$. For real time users rate function is constant.*

It can also easily be shown that $A_j(\lambda_x^{RS}) - B_j(\lambda_x^{RS})$ is a decreasing function of λ_x^{RS} which takes positive values for $\lambda_x^{RS} < n_j^{RS} f_x(\frac{P^{RS}}{n_j^{RS} W})$ and negative values for $\lambda_x^{RS} > n_j^{RS} f_x(\frac{P^{RS}}{n_j^{RS} W})$. For $\lambda_x^{RS} < n_j^{RS} f_x(\frac{P^{RS}}{n_j^{RS} W})$ the LHS of (4.37) is a product of two positive decreasing functions and it is decreasing for user j. For $\lambda_x^{RS} > n_j^{RS} f_x(\frac{P^{RS}}{n_j^{RS} W})$ it is the product of a positive increasing and negative decreasing function hence it is also decreasing in this region for user j. Hence, LHS of (4.37), summation of such functions for all users is a monotonic decreasing function

Let $\lambda_x^{RS*}(\lambda_T)$ be the Lagrangian multiplier that satisfies Equation (4.37) (Please note that the power constraint is automatically satisfied for BS-RS by setting $x_j^{BS} = \frac{P^{BS}}{W n^{BS}}$ for $j \in MS_{RSi}$). Since the total power is an increasing function of λ_x^{RS}, this value can be found by a simple binary search. Then $T_i^{BS}(\lambda_x^{RS*}(\lambda_T), \lambda_T)$ and $T_i^{RS}(\lambda_x^{RS*}(\lambda_T), \lambda_T)$ become the corresponding time allocated to BS-RS and RS-MS transmissions. We are looking for the Lagrange multiplier values $(\lambda_x^{RS*}(\lambda_T^*), \lambda_T^*)$ that satisfies both $S_p^{RS}(\lambda_x^{RS*}(\lambda_T^*), \lambda_T^*) = P$ and $T_i^{BS}(\lambda_x^{RS*}(\lambda_T^*), \lambda_T^*) + T_i^{RS}(\lambda_x^{RS*}(\lambda_T^*), \lambda_T^*) = T_i$. We need another binary search on λ_T. Hence we can find the optimal power,bandwidth and time by two nested binary searches. The algorithm will be described later in more detail.

4.5 Algorithm

Main Algorithm:

1. Determine required rates for real time sessions

2. Test feasibility: If $\underline{T}_i^{BS} + \underline{T}_i^{RS} > T_i$ then find the real time transmitting link with non-zero rate and worst channel condition and drop it.

3. **Run** $(\lambda_x^{RS*}, \lambda_T^*) = BinarySearchTime()$

4. **Run** $(\mathbf{p}, \mathbf{w}, \mathbf{T}_i) = ComputePowerBandTime(\lambda_x^{RS*}, \lambda_T^*)$

Procedure : $(\lambda_x^{RS*}, \lambda_T^*) = BinarySearchTime()$:

1. **Run** $BinarySearch_x^{RS}(2^k \Delta\lambda_T)$ and find the smallest k such that $T_i^{BS}(\lambda_x^{RS*}(2^k\Delta\lambda_T), 2^k\Delta\lambda_T) + T_i^{RS}(\lambda_x^{RS*}(2^k\Delta\lambda_T), 2^k\Delta\lambda_T) > T_i$. Set $\lambda_T^h = 2^k\Delta\lambda_T$, $\lambda_T^l = 2^{k-1}\Delta\lambda_T$.

   Repeat Step 2 until $\left| \frac{T_i}{T_i^{BS*} + T_i^{BS*}} - 1 \right| < \varepsilon$

2. Set $\lambda_T^m = (\lambda_T^h + \lambda_T^l)/2$ and **run** $\lambda_x^{RS*}(\lambda_T^m) = BinarySearch_x^{RS}(\lambda_T^m)$.

   - If $T_i^{BS}(\lambda_x^{RS*}(\lambda_T^m), \lambda_T^m) + T_i^{RS}(\lambda_x^{RS*}(\lambda_T^m), \lambda_T^m) > T_i$ then $\lambda_T^l = \lambda_T^m$.
   - else $\lambda_T^h = \lambda_T^m$.

Procedure : $\lambda_x^{RS*}(\lambda_T) = BinarySearch_x^{RS}(\lambda_T)$: Finds the $\lambda_x^{RS*}(\lambda_T)$ so that $S_p^{RS}(\lambda_x^{RS*}(\lambda_T), \lambda_T) = P^{RS}$.

1. Find the smallest k such that $S_p^{RS}(2^k\Delta\lambda_x^{RS}, \lambda_T) > P^{RS}$. Set $\lambda_x^{RS,h} = 2^k\Delta\lambda_x^{RS}$, $\lambda_x^{RS,l} = 2^{k-1}\Delta\lambda_x^{RS}$.

   Repeat Step 2 until $\left| \frac{S_p(\lambda_x^{RS,m}, \lambda_T)}{P^{RS}} - 1 \right| < \varepsilon$

2. Set $\lambda_T^m = (\lambda_T^l + \lambda_T^h)/2$. If $S_p^{RS}(\lambda_x^{RS,m}, \lambda_T) < P^{RS}$ then $\lambda_T^l = \lambda_T^m$ else $\lambda_T^h = \lambda_T^m$

Procedure:$(\mathbf{p}, \mathbf{w}, \mathbf{T}_i) = ComputePowerBandTime(\lambda_x^{RS}, \lambda_T)$

1. Calculate $r_j, j \in MS_{RSi}$ using (4.28).

2. Calculate T_i^{RS} and T_i^{BS} using (4.31) and (4.30).

Proposition 4.1 *The problem presented in (4.12)-(4.17) (for a feasible case) has a concave objective function and a convex constraint set. Therefore it has a solution.*

Proof 4.2 *The proof is very similar to the proof for Lemma 2.1 and it is omitted.*

Figure 4.3 shows a typical binary search process for a microcell. At each step a $(\lambda_T, \lambda_x^{RS}(\lambda_T))$ pair is found such that the sum of powers is equal to P^{RS}. Since for such pairs time $T_i^{RS} + T_i^{BS}$ is monotonic decreasing in λ_T (as seen in the figure), we are able to find the optimal λ_T by a binary search. Since the channel condition in the access (BS-RS) link is usually much better, usually $T_i^{BS} < T_i^{RS}$. In this example time slot length is $0.1msec$, and after the optimization, all times will be rounded to this value. Therefore we can stop the search when we come less that 0.05msec close to the time constraint (which is 2msec in this example)

4.6 Numerical Evaluation

Figure 4.4 shows a sample MR system. We consider a tandem network of 2km radius, where the BS is at the (0,0) coordinate. The RSs are located at 1400m to the end of the MR-cell. MSs are located at 400,800,1200,1600 and 2000 meters. In order to make

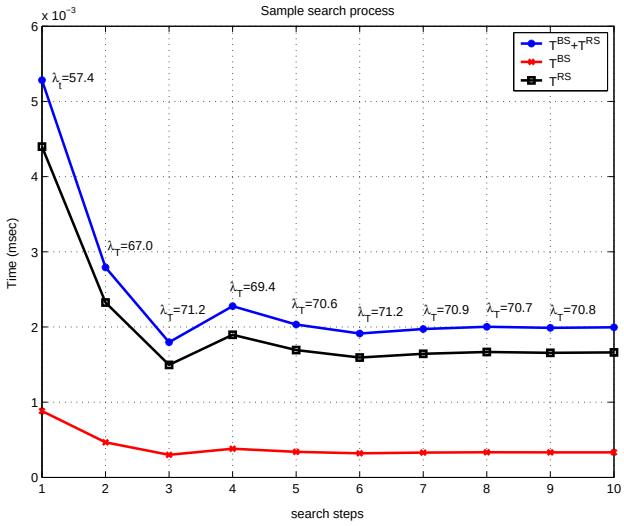

Figure 4.3: A sample binary search process

the station assignment, all the stations (BS and RSs) send broadcast signals (Transmission power for the BS and RSs is P^{BS} and P^{RS}, respectively). Each MS is assigned to the station (either the BS or one of the RSs) that maximizes the received power. [2]

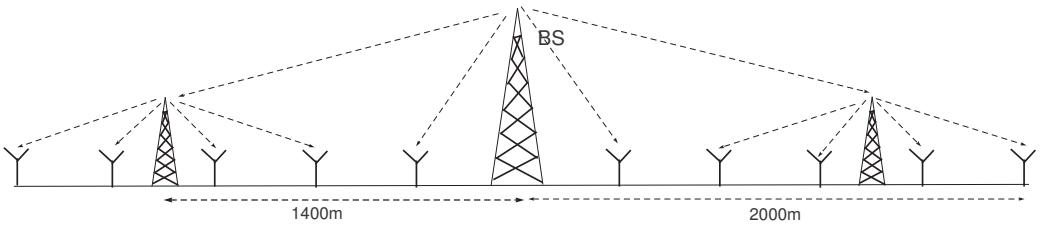

Figure 4.4: A sample MR model for numerical evaluation

As for the path loss, we use the IEEE 802.16j channel model proposed in [55]. For $BS \rightarrow MS$ and $RS \rightarrow MS$ we use the Non-line-of-sight (NLOS) and for $BS \rightarrow RS$ and we use the LOS model. We assume log-normal fading with variance equal to 8Db

[2] In a real system RSs can be located according to the user density in the MR cell area.

for BS transmissions and 3.1dB for RS transmissions, and Rayleigh fading with mean equal to 0.6. We assume that rayleigh fading stays constant at each frame and log-normal fading stays constant during 5 frames. Frame length is equal to 20 slots and each slot is $T_s = 1msec$. Base station and each relay has $P^{BS} = 20W$ and $P^{RS} = 5W$s of power, respectively. Bandwidth is equal to $B = 10MHz$.

Our traffic model is based on [56], and it is as follows: For each data (FTP) session we assume a single 5MB file arriving at the queue at time zero. We assume 32kbps VoIP sessions, where a 320-bit packet arrives at every 10 time slots. Finally we assume 128kbps video streaming sessions, with a fixed video frame duration of 100msec. During each frame there are 8 packets (slices). Packet size is Truncated Pareto distributed with certain min, max and shape parameters. Interarrival time between packets is also Truncated Pareto distributed with certain min, max and shape parameters such that all packets arrive during a 100ms frame. We assume that bits arrive at the end of a time slot and they are ready to transmit at the beginning of the next time slot.

Performance Criteria are

1. 95 percentile delay for voice sessions
2. 95 percentile delay for video sessions
3. average throughput for data sessions

Keeping the number of data and voice users at 20 each, we vary the number of video user from 20 to 50. Figure 4.5 shows the 95^{th} percentile delay for voice sessions. We can observe that in the 2-RS system users at all distances delay stays under the required 100msec level, while for the system with no RSs, users at 1.6km and 2.0km experience

severe delays. Since the coherence time for the log-normal fading is much longer than the voice delay constraint, delays for edge users by far exceed the required levels.

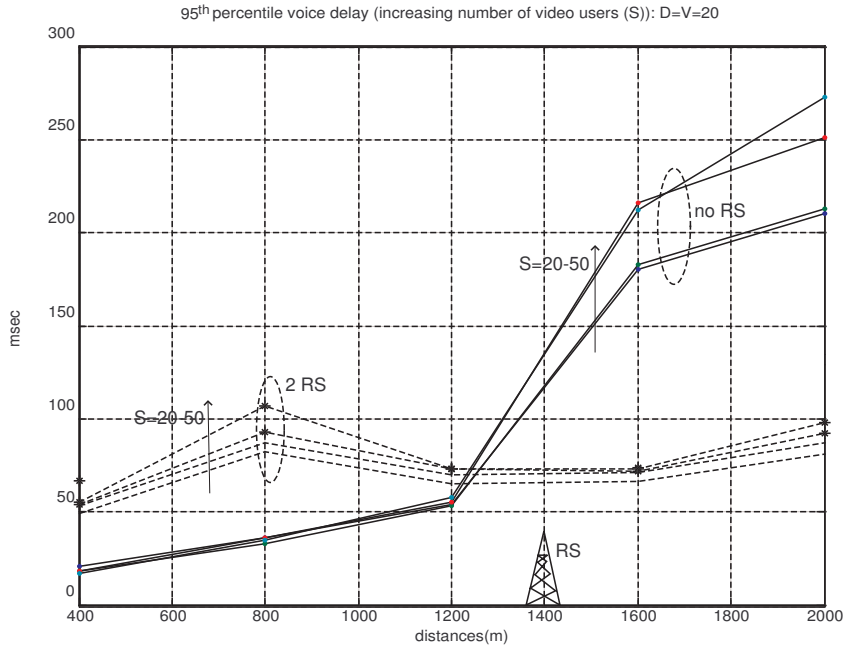

Figure 4.5: 95^{th} percentile voice delay vs. distance to the BS for increasing number of video sessions.

Figure 4.6 shows the 95^{th} percentile delay for video sessions. We again observe that using relays we can prevent QoS violation for users at all distance levels in the cell. Without RSs, users at the cell edge experience high delays.

Figure 4.7 shows the total throughput for users at different distance levels. Here we observe the negative effect of using relays on throughput. Sessions in the RS-microcells have to travel two links. These two links are both very likely to experience a better channel condition than a single BS-MS link, however transmission of a packet requires two frames. Because of this trade-off be observe from Figure 4.7 that users at 0.4, 0.8km

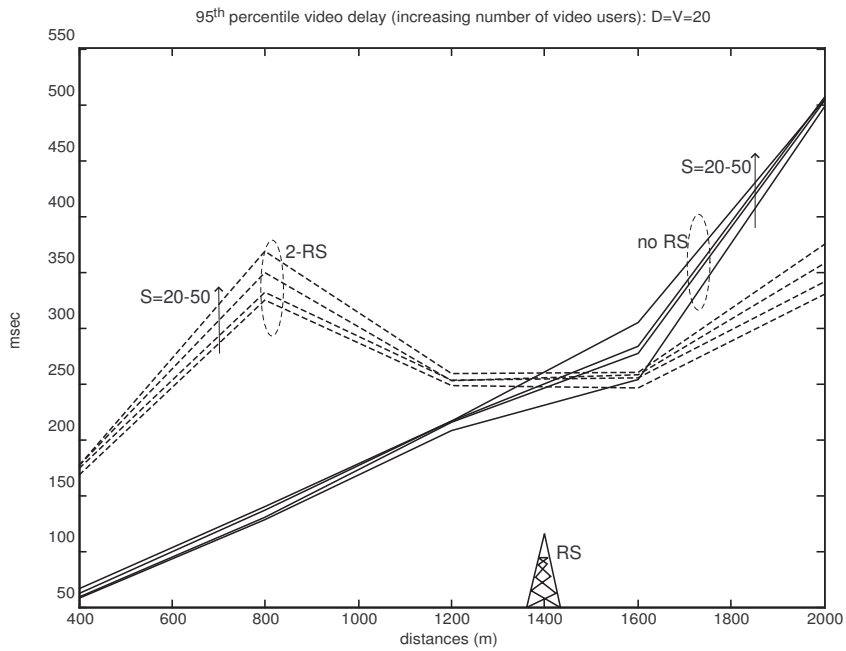

Figure 4.6: 95^{th} percentile video delay vs. distance to the BS for increasing number of video sessions.

receive more throughput in the 0-RS case. On the other hand users at 1.2, 1.6 and 2.0km receive more throughput in the 2-RS system.

We also observe that total throughput decreases more with increasing number of video users in the 0-RS case. In the 2-RS case a video user takes less throughput. Therefore in the case of large number of video users, a system with relays is expected to provide more throughput to data users. We can better observe this in Figure 4.8. We see that throughput for the 2-RS case is better for $S \geq 40$, and log-sum of throughput s better for the 2-RS case for $S \geq 30$.

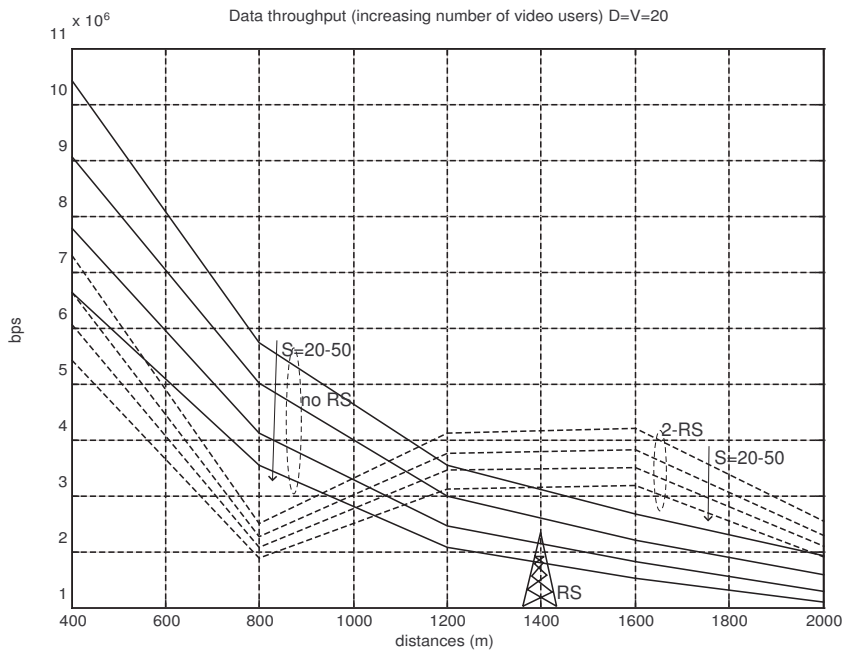

Figure 4.7: Total throughput of data users vs. distance to the BS for increasing number of video sessions.

4.7 Summary

In this Chapter we proposed a joint time, power and bandwidth allocation scheme for downlink transmission in the presence of single-interface relay stations. The proposed scheme consists of two steps, namely subframe allocation for each microcell and joint time , power,bandwidth allocation for links in each microcell. Numerical results show that it is possible to increase the cell size and decrease the number of base stations by adding low-cost relay stations. Multihop relay systems satisfy the QoS requirements of all real time sessions, for the cases, in which regular cellular systems are not sufficient.

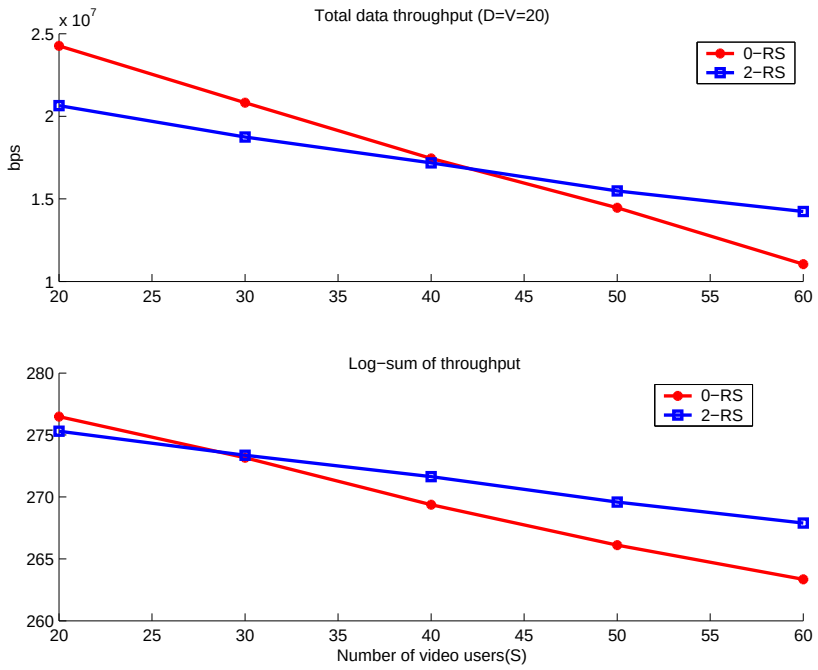

Figure 4.8: Total throughput and log-sum of throughput of data users vs. number of video sessions.

Parameter	**Value**
Cell radius	2km
User Distances	0.4,0.8,1.2,1.6,2.0km
RS Distance	1.4km
# microcells (M)	3
BS,RS Power (P^{BS}, P^{RS})	(20,5) W
W_sub, N_{sub}	267KHz, 30
Frame Length T_f	2 msec
Slot Length T_s	0.1 msec
Voice Traffic	CBR 32kbps
Video Traffic	802.16 - 128kbps
FTP File	5 MB
AWGN p.s.d.(N_0)	-174dBm/Hz
Coherent Time (Fast/Slow)	(4msec/400msec.)
BS-RS PL(d)(in dB)	$36.5 + 23.5 \log_{10} d + \psi_{dB}^{BS-RS}$
RS-MS PL(d)(in dB)	$31.5 + 35 \log_{10} d + \psi_{dB}^{RS-MS}$
BS-MS PL(d)(in dB)	$31.5 + 35 \log_{10} d + \psi_{dB}^{BS-MS}$
ψ_{dB}^{BS-MS}, ψ_{dB}^{RS-MS}	$\sim N(0dB, 8dB)$
ψ_{dB}^{BS-RS}	$\sim N(0dB, 3.1dB)$

Table 4.1: Simulation Parameters

Chapter 5

Queueing Analysis of an OFDMA-based Resource Allocation Scheme

5.1 Introduction

In the previous chapters we have studied power-bandwidth allocation for downlink communication. We considered systems supporting heterogeneous traffic. For real time sessions we determined rate constraints and developed resource allocation algorithm that maximizes proportional fair capacity for data users, while satisfying rate requirements for real time sessions. A crucial assumption in previous chapters was that frequency selective fading among subcarriers was eliminated with the help of distributed subcarrier grouping. This leads to simpler resource allocation algorithms as proposed in previous chapters.

If we use adjacent grouping instead, each subchannel experiences different fading, as we mentioned before. Pursuing our previous objectives in this setting requires more complex algorithms, however in this setting we can propose simple schemes that take advantage of multiuser diversity. In this chapter we will consider such a scheme. We will consider an OFDMA based system, where each user experiences independent and identically distributed fading (i.i.d.) at each subchannel and time slot. A fixed power level is used at each subchannel and each subchannel is allocated to the user that maximizes the signal to noise ratio (SNR). Such a system was analyzed in [57], [58], where the author studied the asymptotic throughput analysis using extreme value theory [59]. Moreover, for users with different distances to the BS (hence different average SINRs)

the author considered allocation of the subchannel to the best *normalized* SINR. Extreme order statistics can be used to approximate the distribution of maximizing random variable in a large set of random variables. Using this method the author in [57], [58] carried out a throughput analysis of the system and proved that asymptotic analysis is quite accurate. In [57] an analysis of delay was also attempted, however apparently it is not realistic. The author models the system as a continuous time M/G/1 system, however the system is inherently discrete-time, since the channel condition changes and new allocations are made at every time slot. In this chapter, modeling as a discrete time multiserver queueing system [44] and using generating function approach we estimate the tail probability of buffer occupancy at a node. Probability of exceeding a certain buffer occupancy threshold is determined as the QoS metric. We look at the trade-off between transmission power and QoS.

If the nodes have different average SNRs (due to differences in distance or log-normal fading) we can revise the scheme to schedule user with best normalized SNR. The rest of the chapter is organized as follows. In Section 5.2 we describe our system model. In this section we also describe the extreme value methodology. In Section 5.3 we make an analysis for the tail probability of queue size. In Section 5.4, we evaluate accuracy of tail probability analysis by simulations. We also look at the trade-off between transmission power and supported traffic rate. In Section 5.5 we look at the case of heterogeneous average SNRs. We numerically compare tail probability estimates with simulations results. This scheme is especially suitable for uplink transmission, since the user can adjust its traffic rate depending on the tail probability estimates.

5.2 System Model

We consider a system, where total bandwidth of W Hz is divided into K subchannels of bandwidth W_{sub}. A fixed power P per subchannel is used by all nodes. We assume that each subchannel is subject to i.i.d. fading which is constant each slot and varies from slot to slot. In a realistic OFDMA system this can be achieved by forming the subchannels using Adaptive Modulation and Coding (AMC) method where each subchannel is a superposition of a number of adjacent subcarriers. Since fading level is fixed at each slot, we assume an AWGN channel and use the tight SNR-BER relations derived in [45]. Let $\gamma_{i,k}$ be the instantaneous SNR of user i at subchannel k. For a target BER the number of packets transmitted in a subchannel as a function of SNR is,

$$r_{i,k} = \frac{W_{sub}T_s}{L}\log_2(1+\beta\gamma_{i,k}) \tag{5.1}$$

where $\beta = -1.5/\ln(5\times BER)$. This formulation was proposed for M-QAM modulation however, it also effectively models continuous rate adaptation [34]. The scheduling mechanism is as follows, each subchannel is allocated to the user with maximum SNR on that subchannel. We assume that each user has identical average SNR and identical fading distribution.

We will start from a simple case, the channel condition of each user at each subchannel is i.i.d Rayleigh distributed with mean γ_0 for all i and k, that is $F_\gamma(\gamma_{i,k}) = 1-e^{-\frac{\gamma_{ik}}{\gamma_0}}$.

5.2.1 Extreme Value Theory

In order to analyze such a system we need to derive the probability distribution of the maximizing SNR at each subchannel. We can use extreme value theory in finding the

asymptotic distributions of extreme values in a set of i.i.d. variables.

Let $\Gamma_k = max_{i\in\mathcal{N}}\gamma_{i,k}$ as the maximizing SNR in subchannel k. For large N, we can approximate the distribution of Γ_k as an extreme value distribution, if some conditions are satisfied [59]. Let $\gamma_{1,k},\gamma_{2,k},\ldots,\gamma_{N,k}$ be independent and identically distributed random variables with distribution function $F_\gamma(x)$. If there exists constants $a_N \in R, b_N > 0$, and some nondegenerate distribution function H such that the distribution of $(\Gamma_k - a_N)/b_N$ converges to H, then H belongs to one of the three standard extreme value distributions:Frechet, Weibull and Gumbel distributions. Since channel conditions are i.i.d. and average SNR's are same for all users we can drop the subchannel subscript. The distribution function of $\gamma_{i,k}$, $F(x)$, determines the exact limiting distribution. If a distribution function $F(x)$ results in one limiting distribution, then $F(x)$ belongs to the domain of attraction of this function.

Lemma 5.1 *[57], [59] Let $\gamma_{1,k},\gamma_{2,k},\ldots,\gamma_{N,k}$ be i.i.d. random variables distribution function $F(x)$. Define $\omega(F) = \sup\{x : F(x) < 1\}$. Assume that there is a real number x_1 such that, for all $x_1 < x < \omega(F)$, $f(x) = F'(x)$ and $F''(x)$ exist and $f(x) \neq 0$. If*

$$\lim_{x\to\omega(F)} \frac{d}{dx}\left(\frac{1-F(x)}{f(x)}\right) = 0$$

then there exists constants a_N and $b_N > 0$ such that $(\Gamma - a_N)/b_N$ uniformly converges in distribution to a normalized Gumbel random variable as $N \to \infty$. The normalized constants are

$$a_N = F^{-1}\left(1-\frac{1}{N}\right) \tag{5.2}$$

$$b_N = F^{-1}\left(1-\frac{1}{Ne}\right) - F^{-1}\left(1-\frac{1}{N}\right) \tag{5.3}$$

where $F^{-1} = \inf\{y : F(y) \geq x\}$

Rayleigh distributed random i.i.d random variables ($f_\gamma(\gamma) = \frac{1}{\gamma_0} e^{-\frac{\gamma}{\gamma_0}}$ *and* $F_\gamma(\gamma) = 1 - e^{-\frac{\gamma}{\gamma_0}}$*) satisfy the above Lemma.*

For $\gamma_{i,k}$ Rayleigh distributed with mean γ_0 , the parameters are: $a_N = \gamma_0 \ln N$ and $b_N = \gamma_0$. Therefore the random variable $\frac{\Gamma - \gamma_0 \ln N}{\gamma_0}$ can be approximated as a normalized Gumbel random variable. A normalized Gumbel distributed random variable, Γ with distribution function $e^{-e^{-\Gamma}}, -\infty < z < \infty$ has expectation $\mathrm{E}(\Gamma) = E_0 = 0.5772..$ and variance $\mathrm{Var}(\Gamma) = \frac{\pi^2}{6}$.

Let $r(\gamma_{i,k}) = \frac{W_{sub} T_s}{L} \log_2(1 + \beta\gamma_{i,k})$ be the number of packets that can be transmitted by user i in subchannel k. Let's define the rate of the SNR-maximizing user in subchannel k as $R^k_{max,N} = \max_{i \in \mathcal{N}}(r(\gamma_{i,k}))$. Since the SNR's are i.i.d, the distribution of $R^k_{max,N}$ is invariant of subchannels, therefore we can drop the subchannel index k. In [57], it was proven that if the SNR distribution satisfies Lemma 5.1, then rate of the maximum-SNR user also converges to Gumbel distribution. More specifically $\frac{R_{max,N} - a_N}{b_N}$ converges to normalized Gumbel distribution, where,

$$a_N = \frac{W_{sub} T_s}{L} \log_2(1 + \beta\gamma_0 \ln N) \tag{5.4}$$

$$b_N = \frac{W_{sub} T_s}{L} \log_2\left(\frac{1 + \beta\gamma_0(1 + \ln N)}{1 + \beta\gamma_0 \ln N}\right) \tag{5.5}$$

Mean and standard deviation of rate of maximum-SNR user in any subchannel is the following,

$$E\{R_{max,N}\} = b_N E_0 + a_N \tag{5.6}$$

$$Std\{R_{max,N}\} = b_N \frac{\pi}{\sqrt{6}} \tag{5.7}$$

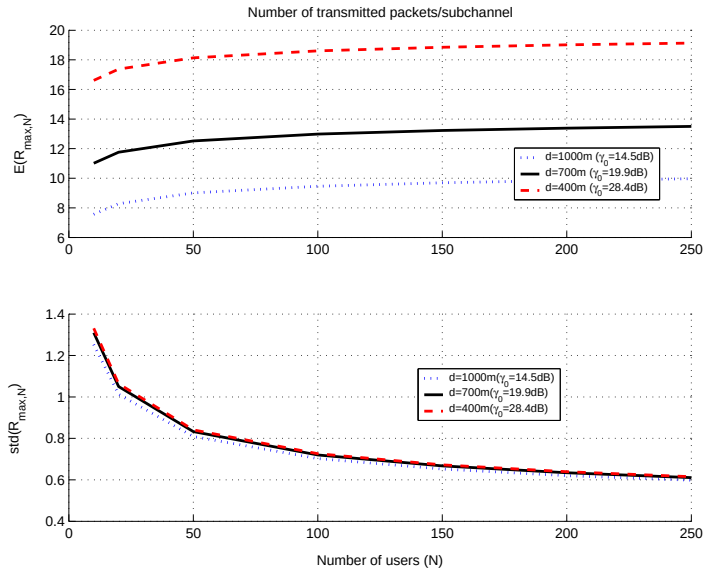

Figure 5.1: Mean and standard deviation

Looking at (5.5), we see that as $M \to \infty$, $a_N \to \infty$ and $b_N \to 0$, and R_N converges to its mean value, $R_{max,N} \approx b_N E_0 + a_N$.

$$E[R_{max,N}] = \frac{W_{sub}T_s}{L}\left(\log_2\left(\frac{1+\beta\gamma_0(1+\ln N)}{1+\beta\gamma_0 \ln N}\right)E_0 + \log_2(1+\beta\gamma_0 \ln N)\right) \tag{5.8}$$

Figure 5.1 shows the mean and standard deviation of R_{max_N}. These results numerically verify that standard deviation decreases and mean increases as $N \to \infty$. Standard deviation is smaller than 1 packet even for moderate number of users, therefore we can assume that a user can transmit $\lfloor b_N E_0 + a_N \rfloor - 1$, $\lfloor b_N E_0 + a_N \rfloor$ or $\lceil b_N E_0 + a_N \rceil$ packets, if allocated. Lets define $R(z) = P(R_{max,N} < \lfloor b_N E_0 + a_N \rfloor) z^{\lfloor b_N E_0 + a_N \rfloor - 1} + P(\lfloor b_N E_0 + a_N \rfloor < R_{max,N} < \lceil b_N E_0 + a_N \rceil) z^{\lfloor b_N E_0 + a_N \rfloor} + P(R_{max,N} > \lceil b_N E_0 + a_N \rceil) z^{\lceil b_N E_0 + a_N \rceil}$ Each user has equal chance of allocating a subchannel, therefore probability of allocation of channel k by a user is $\frac{1}{N}$ for all users and subchannels. Therefore number of allocated subchannels is Binomial distributed. Let $\sigma(s)$ be the probability of total number of pack-

ets that can be transmitted in a time slot being equal to s. Let $\Sigma(z)$ be the probability generating function of $\sigma(z)$.

$$\Sigma(z) = C(K,0)\left(1-\frac{1}{N}\right)^{K} + \sum_{k=1}^{K} C(K,k)\left(\frac{1}{N}\right)^{k}\left(1-\frac{1}{N}\right)^{K-k}(R(z))^{k} \qquad (5.9)$$

5.3 Queueing Analysis

Since the channel conditions for each user and at every subchannel is i.i.d. and channel allocation is performed purely based on normalized channel condition we can decouple the queues of each user and avoid the problem of interacting queues. In queueing theory this system can be modeled as a multiserver system, where the number of active servers is random according to probability vector σ and an active server can transmit a packet in one time slot. We use the generating function approach that was used in [44] for different system. Queueing model for our system can be summarized as follows.

1. Arrivals: A random number of L-bit packets arrive at each time slot. The arrivals occur at the end of the time slot, which means that the data unit that arrives in the current slot can be transmitted in the future time slots. Let a_t denote the number of data units arriving at time slot t. Let $A(z) = E[z^a]$ be the probability generating function function (p.g.f.) of a_t, where E[.] denotes the expected value. For poisson distributed arrivals $A(z) = e^{\lambda(z-1)}$, where $E[a] = \lambda$ packets. For geometric distribution it is $A(z) = \frac{1}{1+\lambda-\lambda z}$.

2. Service: We assume that services start at the beginning of a time slot and end before the new arrivals come.

Let's define $c = K \times \lceil R_{max,N} \rceil$ as the number of servers and let s_t be the number of packets served at time slot t.

$$s_t = s,\ w.p.\ \sigma(s), s = 0, 1, \ldots, \min(q_t, c) \tag{5.10}$$

We define the conditional probability generating function $S_i(z)$ (given that there are i packets in the buffer o a node) as,

$$\begin{aligned} S_i(z) &= E[z^{s_t} | \min(q_t, c) = i],\ i = 0, 1, \ldots, c & (5.11) \\ &= \sum_{s=0}^{i-1} \sigma(s) z^s + \sum_{s=i}^{c} \sigma(s) z^i & (5.12) \end{aligned}$$

Channel allocation is purely based on SNR values and sometimes a user may be allocated more resources than that is enough to empty out the queue. For the simplicity of analysis, in this case we assume that dummy packets are transmitted on the excess subchannels. We also assume that services are independent of arrivals.

3. Overflows: Let D_{max} be the delay constraint in slots. We convert this to a queue size constraint $Q_{max} = \lambda \times D_{max}$ packets using Little's result. Normally, if an arriving packet finds the system full, then it is considered dropped. However, for the simplicity of analysis we are considering an infinite capacity buffer and define the QoS metric as the overflow probability, which is the tail probability of buffer content distribution ($Prob[q_t > Q_{max}]$).

The system equation of the buffer content with respect to time can be written as follows,

$$q_{t+1} = q_t - s_t + a_t \tag{5.13}$$

Let $Q_t(z)$ denote the pgf of q_t. Considering the independence of arrival and service processes and using standard z-transform techniques, we can convert the system equation into the z-domain as follows,

$$Q_{t+1}(z) = A(z)E[z^{q_t - s_t}] = A(z)\left(Q_t(z)S_c(\frac{1}{z}) + \sum_{i=0}^{c-1} q(i)z^i\left(S_i(\frac{1}{z}) - S_c(\frac{1}{z})\right)\right), \quad (5.14)$$

where $q(i)$ denotes the probability that there are i packets in the queue. We are interested in stable systems, where the buffer content distribution reaches a steady state. When the steady state is reached, $Q_t(z)$ and $Q_{t+1}(z)$ converge to a steady state p.g.f. $Q(z)$. Solving the above equation for equilibrium, we get the expression for Q(z).

$$\begin{aligned} Q(z) &= \frac{z^c A(z)\sum_{i=0}^{c-1}\left(S_i(\frac{1}{z}) - S_c(\frac{1}{z})\right)q(i)z^i}{z^c - z^c S_c(\frac{1}{z})A(z)} && (5.15)\\ &= \frac{z^c A(z)\sum_{i=0}^{c-1}\left(\sum_{s=i}^{c}\sigma(s)(z^{-i} - z^{-s})\right)q(i)z^i}{z^c - z^c\sum_{s=0}^{c}\sigma(s)z^{-s}A(z)} && (5.16)\\ &= \frac{A(z)\sum_{i=0}^{c-1}\left(\sum_{s=i}^{c}\sigma(s)(z^c - z^{c-s+i})\right)q(i)}{z^c - \sum_{s=0}^{c}\sigma(s)z^{c-s}A(z)} && (5.17) \end{aligned}$$

where $q(i) = Prob[q_n = i], i = 0,1,\ldots,c-1$ are the buffer occupancy probabilities.

In order to derive $Q(z)$ completely, we need to find the c unknown probabilities $q(i)$ for $i = 0,2,\ldots,c-1$. Here we need the analyticity property of $Q(z)$ inside the unit disk ($z : |z| < 1$). A complex function is said to be analytic in a region if it is defined and differentiable at every point in the region. In order to have the analyticity property, poles of Q(z) inside the unit disk must also be the zeros of Q(z). At this point Rouche's theorem [61] stated below can be utilized to show the number of roots of the denominator inside the unit disk.

Theorem 5.1 *Rouche's Theorem[61] says that: If $f(z)$ and $g(z)$ are analytic functions of s inside and on a closed contour C, and also if $|g(z)| < |f(z)|$ on C, then $f(z)$ and*

$f(z)+g(z)$ have the same number of zeroes inside C. Assuming geometric distributed arrivals $\frac{1}{1+\lambda-\lambda z}$ the denominator of $Q(z)$, $z^c(1+\lambda-\lambda z)-\sum_{s=0}^{c}\sigma(s)z^{c-s}$, has c roots inside and including ($z:|z|<1$).

Proof 5.1 *Let's define $f(z)=z^c(1+\lambda)$ and $g(z)=-\lambda z^{c+1}-\sum_{s=0}^{c}\sigma(s)z^{c-s}$. For the value $|z|=1+\varepsilon$:*

$$\begin{aligned}
|f(z)|-|g(z)| &= |z^c(1+\lambda)|-|\lambda z^{c+1}+\sum_{s=0}^{c}\sigma(s)z^{c-s}| \\
&\geq |z|^c(1+\lambda)-(\lambda|z|^{c+1}+\sum_{s=0}^{c}\sigma(s)|z|^{c-s}) \\
&\geq (1+\varepsilon)^c(1+\lambda)-(\lambda(1+\varepsilon)^{c+1}+\sum_{s=0}^{c}\sigma(s)(1+\varepsilon)^{c-s}) \\
&= (1+c\varepsilon)(1+\lambda)-(\lambda(1+(c+1)\varepsilon)+\sum_{s=0}^{c}\sigma(s)(1+(c-s)\varepsilon))+o(\varepsilon) \\
&= \varepsilon(-\lambda+\sum_{s=0}^{c}\sigma(s)s)+o(\varepsilon)>0 \qquad (5.18)
\end{aligned}$$

We see that under the condition $\sum_{s=0}^{c}\sigma(s)s=cp>\lambda$ (which is also the stability condition) $|f(z)|>|g(z)|$. Since $f(z)$ has c roots, then the denominator has also c zeros. One of them is at $z=1$, and the others are inside the unit disk. Denominator polynomial has order $c+1$, therefore there is a single zero outside unit disk.

Let's denote these roots by z_j, $j=1,2,\ldots,c-1$. Because of the analyticity of $Q(z)$ for $|z|<1$, the numerator must also be zero at these points.

$$\sum_{i=0}^{c-1}\left(\sum_{s=i}^{c}\sigma(s)(1-z_j^{-s+i})\right)q(i)=0,\ j=1,2,\ldots,c-1 \qquad (5.19)$$

We obtain the c^{th} equation from the equality $Q(1)=1$.

$$\sum_{i=0}^{c-1}\left(\sum_{s=i}^{c}\sigma(s)(s-i)\right)q(i)=\sum_{s=0}^{c}\sigma(s)s-A'(1) \qquad (5.20)$$

From the stability assumption, the right hand side of (5.20) has to be greater than zero. From these K equations, the probabilities $q(i),\ i = 0, 1, \ldots, K-1$ can be calculated[1].

5.3.1 Tail Probabilities of the Queue Size

Let $P(q > Q_{max})$ denote the tail probability of the queue size. Tail probability can be used to approximate the overflow probability of a limited buffer. It has been previously found in [44],[62],[63],[64] that for sufficiently large values of Q_{max}, the tail distribution of queue size can be approximated as,

$$Prob[q > Q_{max}] \approx -R_q \frac{z_q^{-Q_{max}-1}}{z_q - 1}, \tag{5.21}$$

where z_q is the real positive pole of $Q(z)$ with the smallest modulus outside the unit disk, i.e. it is the dominant pole of $Q(z)$. R_q is the residue of $Q(z)$ at $z = z_q$. Assuming geometric distributed arrivals the p.g.f of queue size $Q(z)$ has only one pole outside unit circle (therefore it is real), one pole at z=1 and the rest inside the unit circle. It can be derived by evaluating $(z - z_q)Q(z)$ at $z = z_q$.

[1]Since we consider a large number of users, allocation probability of a subchannel to a user is very low. Probability of allocation of k subchannels to a user diminishes very quickly as k increases. When solving equations (5.19), (5.20) in MATLAB, errors occur because of the precision of the software. To prevent this, we can crop the probability vector σ without losing accuracy. This also speeds up the computation

$$
\begin{aligned}
R_q &= (z-z_q)Q(z)\big|_{z=z_q} && (5.22)\\
&= \left.\frac{(z-z_q)A(z)\sum_{i=0}^{c-1}\left(\sum_{s=i}^{c}\sigma(s)(z^c-z^{c-s+i})\right)q(i)}{z^c-\sum_{s=0}^{c}\sigma(s)z^{c-s}A(z)}\right|_{z=z_q} && (5.23)\\
&= \left.\frac{A(z)\sum_{i=0}^{c-1}\left(\sum_{s=i}^{c}\sigma(s)(z^c-z^{c-s+i})\right)q(i)}{cz^{c-1}-\sum_{s=0}^{c}\sigma(s)(c-s)z^{c-s-1}A(z)-\sum_{s=0}^{c}\sigma(s)z^{c-s}A'(z)}\right|_{z=z_q} && (5.24)\\
&= \left.\frac{A(z)\sum_{i=0}^{c-1}\left(\sum_{s=i}^{c}\sigma(s)(z^c-z^{c-s+i})\right)q(i)}{\frac{1}{z}\sum_{s=0}^{c}\sigma(s)sz^{c-s}A(z)-\frac{z^cA'(z)}{A(z)}}\right|_{z=z_q} && (5.25)\\
&= \frac{A(z_q)\sum_{i=0}^{c-1}\left(\sum_{s=i}^{c}\sigma(s)(1-z_q^{-s+i})\right)q(i)}{\frac{1}{z_q}\sum_{s=0}^{c}\sigma(s)sz_q^{-s}A(z_q)-\frac{A'(z_q)}{A(z_q)}} && (5.26)
\end{aligned}
$$

Equation (5.24) come from the L'Hospital rule and (5.25) is written using the fact that denominator of $Q(z)$ is zero at $z=z_q$. As the system load increases, z_q approaches to 1, the probability of exceeding a buffer occupancy threshold increases. For geometric arrival process (i.e. $A(z)=\frac{1}{1+\lambda-\lambda z}$) the residue is written as follows:

$$
R_q=\frac{\sum_{i=0}^{c-1}\left(\sum_{s=i}^{c}\sigma(s)(1-z_q^{-s+i})\right)q(i)}{\frac{1}{z_q}\sum_{s=0}^{c}\sigma(s)sz_q^{-s}-\lambda} \qquad (5.27)
$$

5.4 Numerical Evaluations

We performed a numerical study to evaluate the accuracy of tail probability estimates and see the energy-QoS trade-off by varying the transmission power. We assume a system of K=30 subchannels, where each subchannel is of $W_sub = 200KHz$. System is slotted with slot length $T_s = 0.001$sec. Pathloss in (dB's) is $31.5+35*\log_{10}(d)$, where d is the distance of the node to the base station. We assume Rayleigh fading with mean equal to one that is constant at each time slot and is i.i.d. from slot to slot. In Figure 5.2, we considered 100 users and two packets sizes $L=100$ and 50 bits,

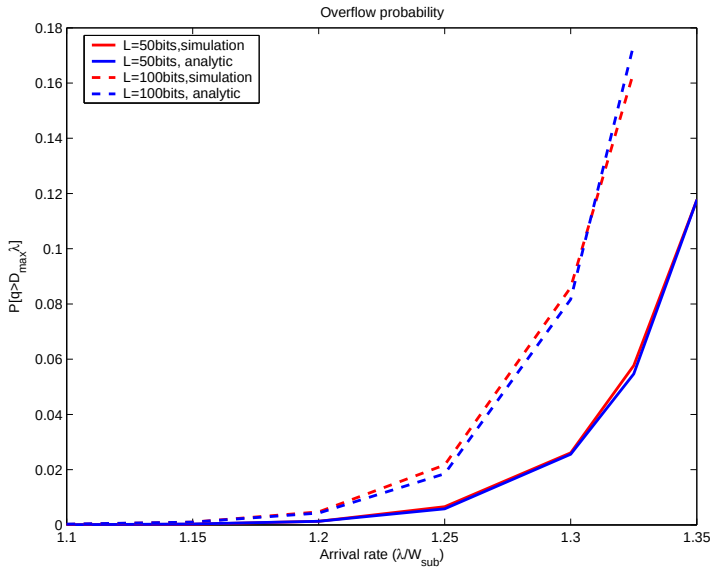

Figure 5.2: Tail probability vs. traffic rate

Distances of users are $d = 1000m$ for each user, therefore their average SNRs are the same. Arrival process for each user is geometric distributed with mean varying from 220Kbps to 260Kbps. Delay constraint is 0.1msec, which is converted to $Q_{max} = \lambda \times 0.1$ bits for each arrival rate. Figure 5.2 shows the analytical and simulation results for overflow probability versus power per subchannel for this system. We observe that analytical results are very close to the simulation results and overflow probability is increasing and convex as a function of arrival rate.

5.5 Normalized SNR-based scheduling

In reality average SNRs of users are different due to differences in distances to the base stations and effects of shadowing. In this case scheduling the best user causes unfairness in the network. However, when we schedule users based on their normalized

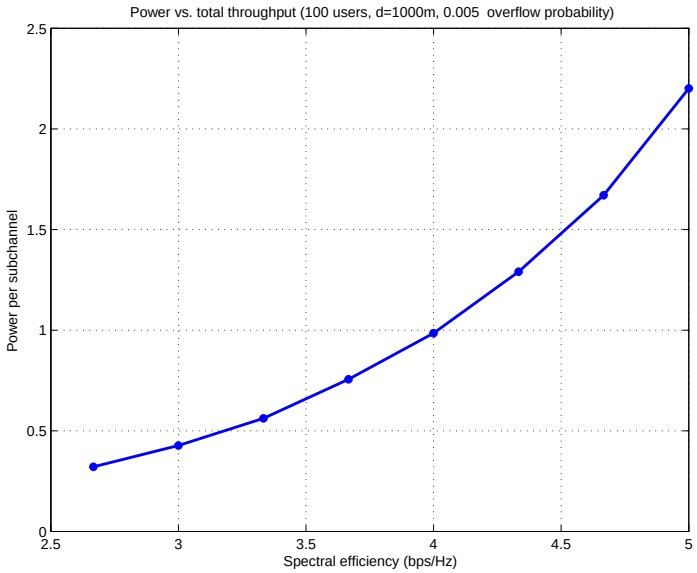

Figure 5.3: Energy-throughput trade-off

SNR, resource allocation becomes both fair and analyzable. In this case, subchannel k is allocated to the user $\arg\max_{i\in\mathcal{N}} \frac{\gamma_{i,k}}{\gamma_0^i}$. Since the SNR of a users is the product of normalized SNR and a random variable that is i.i.d. for each user and subchannel, previous results on extreme value statistics and subsequent queueing analyses still holds. If user i is allocated a subchannel, then expected number of packets that it can transmit is $R^i_{max,N}$, which is found by replacing γ_0 by γ_0^i, average SNR of the user that maximizes the normalized SNR.

In this system each user has the same channel access probability, however users with higher average SNR can support sessions with higher rates. The ratio of session rates of users i and j is, $\frac{\lambda_i}{\lambda_j} = \frac{R^i_{max,N}}{R^j_{max,N}}$. If e set the following proportionality among different user traffic rates, we can better utilize the resources.

$$\lambda_0^1 : \lambda_0^2 : \ldots : \lambda_0^N = R^1_{max,N} : R^2_{max,N} : \ldots : R^N_{max,N} \tag{5.28}$$

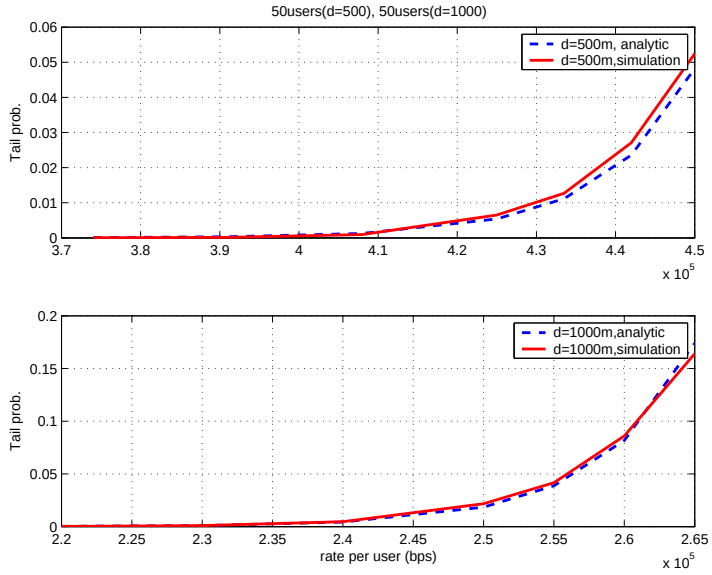

Figure 5.4: Tail probability vs. rate for heterogeneous SNR case

In Figure 5.4 we considered a system of 50 users at 500m and 50 users at 1000m distances. For near users $R^i_{max,N} = 16.6871$ and for far users $R^i_{max,N} = 9.7777$ packets/slot. The ratio is 1.7 and we increase the rate, maintaining this ratio among rates of two classes of users. We see that analytical results closely follow the simulation results.

5.5.1 Implementation of the system

A realistic system has to support users with different average SNRs and demanding services with different QoS requirements. For example data services have very loose delay requirements. Besides these sessions can use whatever rate that is available to them. On the other hand video streaming sessions have stricter delay requirements and they can be transmitted in varying quality levels (e.g. 128,256,512,1024Kbps). Since we

can estimate maximum supportable rate through $R^i_{max,N}$ for all users, a video user can choose one of the available levels based on this estimate and its QoS requirements. This system is suitable for implementation especially in uplink transmission, since it is easier for a user to control the traffic it generates. On the other hand voice sessions (e.g. VoIP) have a single rate level (e.g. 32kbps), therefore for these sessions overutilization may occur. This problem can be relieved if a voice user doesn't enter the *competition* if it doesn't have any packets in its buffer.

5.6 Summary

In this chapter we studied queueing analysis of an OFDMA based resource allocation scheme using extreme value theory and generating function approach. We performed a queueing analysis to estimate the tail probability of queue size distribution for this system. We tested the accuracy of the estimates by simulations and observed that estimates are quite accurate. We both considered systems where users have same average SNR and different average SNRs. The analysis we performed can be used to easily estimate the probability of quality of service violation given the system parameters and to adjust the session rate or transmission power to improve the utilization.

Chapter 6

Conclusions

In this dissertation we focused on resource allocation in Orthogonal Frequency Division Multiple Access systems that support users with heterogeneous quality of service requirements. In Chapters 2 and 3 we proposed joint power/bandwidth allocation algorithms that are suitable for transmission of data, voice and video sessions from Base Stations to mobile users. We consider systems in which the subcarriers grouped into subchannels by taking samples across the frequency spectrum in a distributed manner. This way we can assume that each subchannel experiences the same average fading with respect to a user. We assumed bandwidth as a continuously divisible quantity and formulated constrained optimization problems that can be solved by relatively simple algorithms. We converted the delay requirements of voice and video sessions into rate requirements at each frame. Our objective is maximizing proportional fair capacity of data users subject to rate constraints for voice and video sessions. Simulation results showed that our algorithms perform significantly better than a multichannel version of M-LWDF, which is a well known algorithm that can support heterogeneous traffic. In Chapter 3 we also distinguished video and voice sessions in terms of elasticity. Using a simple video rate control scheme for both our algorithm and benchmark algorithm we observed that the proposed algorithm can provide more rate for video users than the benchmark algorithm.

In Chapter 4, we considered the use of low-cost Relay Stations (RSs), that are

able to improve the cell coverage by relaying the information coming from the Base Station(BS) to mobile stations (MS). Such networks are recently gaining interest along with the IEEE 802.16j standard that is being developed. Low-cost nature of the relay station equipment doesn't allow simultaneous transmission and reception, therefore we need to divide the frame into TDMA subframes, in which different $BS-RS$, $BS-MS_{BS}$, and $RS-MS_{RS}$ pairs (i.e. composite links) schedule their transmissions. Resource allocation comes in three dimensions, power, bandwidth and time. We proposed an efficient algorithm that first allocates the TDMA subframes and then performs joint power-bandwidth allocation for each BS-RS-MS pair. Simulation results show that using RSs provides significant performance improvement especially for Video and Voice sessions at the cell edge and that it is possible to increase the cell size and decrease the number of BSs in a multicell environment by the use of RSs, which makes mobile multihop relay networks a promising approach.

The work we did in Chapter 5 presents a different approach. In this Chapter we addressed frequency selective fading channels, unlike previous chapters, and considered a simple subchannel allocation scheme that allocates each subchannel to user with maximum normalized SNR. Although this scheme doesn't guarantee any performance objectives as in our previously proposed algorithms, it exploits multiuser diversity and it can be theoretically analyzed. Using extreme value theory and generating functions approach we analyzed the tail distribution of the queue sizes in this system. Simulation results show that our estimates are quite close to the actual values. The proposed method can be used for admission control and rate control in the presence of QoS constraints.

6.1 Future Work

We can extend the work done in each chapter in order to make them more suitable to use in real-time environments. Here are some possible directions.

6.1.1 Realistic evaluation and comparison of resource allocation algorithms

Resource allocation algorithms that we proposed in the first three contributions are especially suitable for mobile networks with fast fading. Since we assume distributed subcarrier grouping (e.g. PUSC in WiMax), frequency selectivity in fading is eliminated and base station doesn't need to estimate the fading level in each subchannel separately. Another advantage of this way of subchannelization is that each subchannel is equivalent with respect to a user, therefore we are able to propose less complex algorithms that treat the entire frequency spectrum as a continuously divisible quantity.

Although we equalize the average fading level in each subchannel by distributed subcarrier grouping, frequency selectivity is still there among the subcarriers in a subchannel. It would be interesting to create a realistic simulation environment and test our algorithms. It would be also interesting to compare our algorithms with the algorithms in the literature that are proposed for the frequency selective fading. Our algorithms are less complex and they are supposed to perform well under fast fading.

6.1.2 Frequency reuse and cooperation in multihop relay networks

We proposed a joint power, bandwidth and time allocation algorithm for multihop relay networks. In this setting transmissions of relay and base stations are scheduled in a TDMA fashion. It is possible to increase the network capacity through frequency reuse. Depending on the path losses and fading between relay stations and users, two or even more relay stations can transmit simultaneously. Location of the relay stations are also important in frequency. Intercell interference also depends on the location of relay stations, therefore network topology management should also be studied.

Cooperation in relay channels was extensively studied in the literature. Relays can take the advantage of statistical dependence between their channel outputs and destination channel outputs [65], [66]. In our system model we did not consider cooperation. In fact cooperation may provide significant room for improvement and it is a direction of future research.

6.1.3 Extensions for queueing analysis of OFDMA-based system

In Chapter 5 we made a queueing analysis for an OFDMA based subchannel allocation scheme, in the presence of frequency selective fading. We saw that our tail probability estimates for the queue sizes are quite accurate, however it is also important to investigate the block fading case. If the fading level is fixed for several time slots, service process for a node becomes more bursty and packet delays are supposed to increase. It is also a future research direction to pursue the analysis for different arrival processes and see the performance of the algorithm for heterogeneous traffic.

Research directions listed above are possible extensions of the work we did in this thesis. Besides, there are more diverse future directions. Investigating the use of multiple antennas is one of them. We also assumed fixed number of sessions in our simulations. It is also important to consider admission control and rate control and investigate possible interactions between MAC Layer and Network and Transport Layers to improve the performance.

Appendix A

Proof of Lemma 2.1

Lemma A.1 *The reward function*

$$C(\mathbf{w}^n, \mathbf{p}^n) = \sum_{i \in U} \log \left(\alpha_i R_i + (1 - \alpha_i) \left(w_i log(1 + \frac{p_i}{n_i w_i}) - r_i^0 \right) \right) \tag{A.1}$$

is a concave function of w_i and p_i for all $i \in U_D$.

Proof A.1 *If we take the Hessian $\mathbf{H_g}$ of $g(w,p) = \log\left(\alpha R + (1-\alpha)\left(w log(1+\frac{p}{nw}) - r^0\right)\right)$,*

$$\mathbf{H_g} = \frac{-(1-\alpha)}{(p+nw)^2} \begin{bmatrix} w & -p \\ p & \frac{p^2}{w} \end{bmatrix} \tag{A.2}$$

we see that it is negative definite, therefore the function is strictly concave. Therefore the linear combination (A.1) is also concave.

A.0.4 Convexity of the Feasible Set

Lemma A.2 *The feasible set of power and bandwidth levels (w,p) defined by (2.21), (2.22) and (2.23) defines a convex set.*

Proof A.2 *Consider two power-bandwidth vectors $(\mathbf{w}^1, \mathbf{p}^1)$ and $(\mathbf{w}^2, \mathbf{p}^2)$ that are in the feasible set. Now let us consider power-bandwidth vector $(\lambda \mathbf{w}^1 + (1-\lambda)\mathbf{w}^2, \lambda \mathbf{p}^1 + (1-\lambda)\mathbf{p}^2)$. It is clear that this vector satisfies the feasibility constraints in (2.23).*

Now consider a user $i \in U_V$. This user has a rate constraint r_i^0 in (2.22). If (w_i^1, p_i^1) and (w_i^2, p_i^2) both satisfy constraint (2.22):

$$r(w_i^1, p_i^1) = w_i^1 log(1 + \frac{p_i^1}{n_i w_i^1}) = r_i^0 \quad (A.3)$$

$$r(w_i^2, p_i^2) = w_i^2 log(1 + \frac{p_i^2}{n_i w_i^2}) = r_i^0, \forall i \in U_V \quad (A.4)$$

From the concavity of the Shannon capacity with respect to w_i and p_i, we can write ($\overline{\lambda} = 1 - \lambda$):

$$r(\lambda w_i^1 + \overline{\lambda} w_i^2, \lambda p_i^1 + \overline{\lambda} p_i^2) = (\lambda w_i^1 + \overline{\lambda} w_i^2) log(1 + \frac{\lambda p_i^1 + \overline{\lambda} p_i^2}{n_i(\lambda w_i^1 + \overline{\lambda} w_i^2)}) \quad (A.5)$$

$$\geq \lambda w_i^1 log(1 + \frac{p_i^1}{n_i w_i^1}) + \overline{\lambda} w_i^2 log(1 + \frac{p_i^2}{n_i w_i^2}) \quad (A.6)$$

$$= \lambda r_i^0 + \overline{\lambda} r_i^0 = r_i^0 \quad (A.7)$$

Hence the power bandwidth values $(\lambda w_i^1 + \overline{\lambda} w_i^2, \lambda p_i^1 + \overline{\lambda} p_i^2)$ also satisfy the rate constraints for users $i \in U_V$.

For users $i \in U_D$ the same method can be used, only by replacing r_i^0 by ρ_i^0 in feasibility condition (2.21). Hence it is proven that the feasible set of power and bandwidth levels $(\mathbf{w}, \mathbf{p})$ defines a convex set.

Appendix B

Proof of Lemma 2.2

Lemma B.1 *The following properties hold:*

1. *Effective SINR ($x_i(\Lambda_x)$) is a monotonic increasing function of Λ_x for users $i \in U_D \cup U_R'$.*

2. *If $n_i < n_j$ then $x_i(\Lambda_x) > x_j(\Lambda_x)$*

3. *If $n_i > n_j$ then $x_i(\Lambda_x)n_i > x_j(\Lambda_x)n_i$*

Proof B.1 1. *The derivative of function $f_x(x)$ is:*

$$f_x'(x) = log(1+x) > 0$$

for $x > 0$. Therefore $f_x(x_i)$ is a strictly increasing function of x_i for all users i. Hence the inverse $x_i(\Lambda_x) = f_x^{-1}(\Lambda_x/n_i)$ is also increasing in Λ_x.

2. *Since $f_x^{-1}(\Lambda_x/n_i)$ is a monotonic increasing function of Λ_x, it is a monotonic decreasing function of n_i, therefore the property holds.*

3. *For a fixed Λ_x let us define $a_i(n_i) = f_x^{-1}(\Lambda_x/n_i)n_i$. Then after some derivations*

$$\frac{da_i}{dn_i} = -1 + \frac{a_i/n_i}{\log(1+a_i/n_i)} > 0 \tag{B.1}$$

The derivative is greater than zero because of the logarithmic identity $x > \log(1+x)$. Therefore if $n_i > n_j$ then $x_i(\Lambda_x)n_i > x_j(\Lambda_x)n_i$

Appendix C

Proof of Lemma 2.3

i. For any $\Delta_p > 0$, we can write the following,

$$[\Lambda_p + \Delta_p - n_i(1+x_i)R_i\widetilde{\alpha}_i]^+ \geq [\Lambda_p - n_i(1+x_i)R_i\widetilde{\alpha}_i]^+, \forall i \in U_D$$

$$S_w(\Lambda_x, \Lambda_p + \Delta_p) - S_w(\Lambda_x, \Lambda_p) = \sum_{i \in U_D} \frac{[\Lambda_p + \Delta_p - n_i(1+x_i)R_i\widetilde{\alpha}_i]^+}{log(1+x_i)(1+x_i)n_i} - \sum_{i \in U_D} \frac{[\Lambda_p - n_i(1+x_i)R_i\widetilde{\alpha}_i]^+}{log(1+x_i)(1+x_i)n_i} \geq 0$$

Hence $S_w(\Lambda_x, \Lambda_p)$ is nondecreasing in Λ_p. Also,

$$\lim_{\Lambda_p \to \infty} [\Lambda_p - n_i(1+x_i)R_i\widetilde{\alpha}_i]^+ = \infty, \forall \Lambda_x$$

Therefore $\lim_{\Lambda_p \to \infty} S_w(\Lambda_x, \Lambda_p) = \infty$

We can similarly verify that $S_p(\Lambda_x, \Lambda_p)$ is nondecreasing in Λ_p and $\lim_{\Lambda_p \to \infty} S_p(\Lambda_x, \Lambda_p) = \infty$ for all Λ_x.

ii. We know for all users $i \in U_D \cup U_R'$ that:

- $x_i = f_a^{-1}(\Lambda_x / n_i)$ is increasing in Λ_x (From Lemma 2.2).
- The expression $[\Lambda_p - n_i(1+x_i)R_i\widetilde{\alpha}_i]^+$ is nonincreasing in x_i for any Λ_p. It goes to zero as x_i goes to infinity for all Λ_p. It is equal to $[\Lambda_p - n_i R_i\widetilde{\alpha}_i]^+$ at $x_i = 0$.
- The expressions $\frac{1}{\log(1+x_i)}$ and $\frac{1}{\log(1+x_i)(1+x_i)}$ are decreasing in x_i and both go to zero as x_i goes to infinity and they go to infinity as x_i goes to zero.

From these properties we can deduce that $S_w(\Lambda_x, \Lambda_p)$ in (2.38) is a decreasing function of Λ_x and $\lim_{\Lambda_x \to 0} S_w(\Lambda_x, \Lambda_p) = \infty$, and $\lim_{\Lambda_x \to \infty} S_w(\Lambda_x, \Lambda_p) = 0$ for all Λ_p.

iii. We know for all users $i \in U_D \cup U'_R$ that:

- $x_i = f_a^{-1}(\Lambda_x / n_i)$ is increasing in Λ_x.
- The expression $\frac{x_i}{\log(1+x_i)}$ is strictly increasing in x_i.
- The expression $\frac{[-R_i \tilde{\alpha}_i]^+ x_i}{\log(1+x_i)}$ is equal to zero since $R_i \geq 0$.

From these properties we can deduce that $S_p(\Lambda_x, 0)$ in (2.44) is an increasing function of Λ_x.

iv. There exists such a Λ_x^0 because $S_w(\Lambda_x, 0)$ is a strictly decreasing function which is infinity for $\Lambda_x = 0$ and zero for $\Lambda_x = \infty$ as a corollary of Lemma 2.3.ii.

- $\Rightarrow$: If $S_p(\Lambda_x^0, 0) \leq P$ then both feasibility conditions (2.43) and (2.44) hold, therefore the problem is feasible.
- $\Leftarrow$: If the problem is feasible, then there exists Λ_x such that both (2.43) and (2.44) hold. Now let's assume that $S_p(\Lambda_x^0, 0) > P$, then from Lemma 2.3.i and iii $S_p(\Lambda_x^0, \Lambda_p) > P$ for all $\Lambda_x \geq \Lambda_x^0$ and $\Lambda_p \geq 0$. Note that $S_w(\Lambda_x^0, \Lambda_p) > W$ for $\Lambda_x < \Lambda_x^0$, $\Lambda_p \geq 0$ from Lemma 2.3.i. This means that there is no Λ_x such that both (2.43) and (2.44) hold and the problem is infeasible. This is a contradiction, therefore $S_p(\Lambda_x^0, 0) \leq P$. The property holds.

v. We can derive $\frac{d\Lambda_p(\Lambda_x)}{d\Lambda_x}$ as follows:

$$\Lambda_p^*(\Lambda_x)|U_D'| = \Lambda_x W + P + \sum_{U_D'} n_i(1+f_x^{-1}(\Lambda_x^*/n_i))R_i\widetilde{\alpha}_i - \sum_{U_R'} r_i^c n_i(1+f_x^{-1}(\Lambda_x^*/n_i)) \tag{C.1}$$

$$|U_D'|\frac{d\Lambda_p^*(\Lambda_x)}{d\Lambda_x} = W + \sum_{U_D'} \frac{R_i\widetilde{\alpha}_i}{\log(1+f_x^{-1}(\Lambda_x^*/n_i))} - \sum_{U_R'} \frac{r_i^c}{\log(1+f_x^{-1}(\Lambda_x^*/n_i))} \tag{C.2}$$

We know that for $\Lambda_x > \Lambda_x^0$, the problem is feasible and $S_w(\Lambda_x,0) = \sum_{U_R'} \frac{r_i^c}{\log(1+f_x^{-1}(\Lambda_x^*/n_i))} < W$. Therefore the right hand side of (C.2) is greater than zero, which obviously means that $\frac{d\Lambda_p^*(\Lambda_x)}{d\Lambda_x} > 0$. Hence the function $\Lambda_p^*(\Lambda_x)$ is an increasing function of Λ_x.

vi. For a feasible problem, using (2.42) and the fact $[\Lambda_p^*(\Lambda_x) - n_i(1+f_x^{-1}(\frac{\Lambda_x}{n_i}))R_i\widetilde{\alpha}_i]^+ \geq \Lambda_p^*(\Lambda_x) - n_i(1+f_x^{-1}(\frac{\Lambda_x}{n_i}))R_i\widetilde{\alpha}_i$ the following can be written:

$$\begin{aligned}
\Lambda_x W + P &\geq \sum_{i\in U_D} \Lambda_p^*(\Lambda_x) - n_i(1+f_x^{-1}(\frac{\Lambda_x}{n_i}))R_i\widetilde{\alpha}_i + \sum_{i\in U_R'} r_i^c n_i(1+f_x^{-1}(\frac{\Lambda_x}{n_i})) \\
\Lambda_p^*(\Lambda_x) &\leq \frac{\Lambda_x W + P - \sum_{i\in U_R'} r_i^c n_i(1+f_x^{-1}(\frac{\Lambda_x}{n_i})) + \sum_{i\in U_D} n_i(1+f_x^{-1}(\frac{\Lambda_x}{n_i}))R_i\widetilde{\alpha}_i}{|U_D|}
\end{aligned}$$

Hence the inequality is proved.

We can prove the inequalities for the optimal Λ_x^* using contradiction. Suppose that $\Lambda_x^* > \max_{i\in U_D\cup U_R'}\left\{n_i f_x\left(\frac{P}{n_i W}\right)\right\}, \forall i \in U_D \cup U_R'$, then $f_x^{-1}(\Lambda_x^*/n_i) > \frac{P}{n_i W}, \forall i$, from the monotonicity property. Then the total power is greater than $\sum_{i\in U_D'\cup U_R'} w_i^* \frac{P}{n_i W} = P$, which contradicts with the power constraint, therefore the upper bound is proven. For the lower bound assume that $\Lambda_x^* < \min_{i\in U_D\cup U_R'}\left\{n_i f_x\left(\frac{P}{n_i W}\right)\right\}, \forall i \in U_D \cup U_R'$, then $f_x^{-1}(\Lambda_x^*/n_i) < \frac{P}{n_i W}, \forall i$, from the monotonicity property. Then the total power is smaller than $\sum_{i\in U_D'\cup U_R'} w_i^* \frac{P}{n_i W} = P$. This is not optimal because proportional fair capacity can be increased by using the residual power, therefore the lower bound is also proven.

vii. For a feasible problem $S_w(\Lambda_x,0)\leq W$ for $\Lambda_x\geq\Lambda_x^0$ (Lemma 2.3.iv). Since $S_w(\Lambda_x,\Lambda_p)$ is a nondecreasing function of Λ_p and goes to infinity as Λ_p goes to infinity (Lemma 2.3.i) there exists a $\Lambda_p^*(\Lambda_x)$ such that $S_w(\Lambda_x,\Lambda_p^*(\Lambda_x))=W$.

Since Λ_x^0 is a feasible Λ_x value, the sum of user powers is smaller than P for Λ_x^0 from Lemma 2.3.iv. As Λ_x goes to infinity x_i goes to infinity for all i. Since $0\leq w_i(\Lambda_x,\Lambda_p^*(\Lambda_x))\leq W$ for all users $S_p(\Lambda_x,\Lambda_p^*(\Lambda_x))=\sum_i w_i(\Lambda_x,\Lambda_p^*(\Lambda_x)))x_i(\Lambda_x)n_i$ goes to infinity as Λ_x goes to infinity.

viii. Using 2.42 the relation between $\Lambda_p^*(\Lambda_x)$ and Λ_x is as follows:

$$W=\sum_{i\in U_D}\left[\frac{\Lambda_p^*(\Lambda_x)}{\Lambda_x+P/W}-\frac{n_i(1+f_x^{-1}(\frac{\Lambda_x^*}{n_i}))R_i\widetilde{\alpha}_i}{\Lambda_x+P/W}\right]^+ +\sum_{i\in U_R'}\frac{r_i^c n_i(1+f_x^{-1}(\frac{\Lambda_x^*}{n_i}))}{\Lambda_x+P/W} \quad \text{(C.3)}$$

Also from 2.38,

$$S_w(\Lambda_x,\Lambda_p^*(\Lambda_x))=\sum_{i\in U_D\cup U_R'}a_i(\Lambda_x,\Lambda_p^*(\Lambda_x))\frac{\Lambda_x+P/W}{\Lambda_x+f_x^{-1}(\Lambda_x/n_i)n_i} \quad \text{(C.4)}$$

where $a_i(\Lambda_x,\Lambda_p^*(\Lambda_x))=\left[\frac{\Lambda_p^*(\Lambda_x)}{\Lambda_x+P/W}-\frac{n_i(1+f_x^{-1}(\frac{\Lambda_x^*}{n_i}))R_i\widetilde{\alpha}_i}{\Lambda_x+P/W}\right]^+$ for $i\in U_D$ and $a_i(\Lambda_x,\Lambda_p^*(\Lambda_x))=\frac{r_i^c n_i(1+f_x^{-1}(\frac{\Lambda_x^*}{n_i}))}{\Lambda_x+P/W}$ for $i\in U_R'$. Combining the two equations we obtain

$$S_w(\Lambda_x,\Lambda_p^*(\Lambda_x))-W=\sum_{i\in U_D\cup U_R'}a_i(\Lambda_x,\Lambda_p^*(\Lambda_x))\frac{P/W-f_x^{-1}(\Lambda_x/n_i)n_i}{\Lambda_x+f_x^{-1}(\Lambda_x/n_i)n_i} \quad \text{(C.5)}$$

The function $\frac{n_i(1+f_x^{-1}(\frac{\Lambda_x^*}{n_i}))R_i\widetilde{\alpha}_i}{\Lambda_x+P/W}$ takes value $R_i\widetilde{\alpha}_i\frac{Wn_i}{P}$ at $\Lambda_x=0$. It is increasing for $0<\Lambda_x<n_i f_x(P/Wn_i)$ and takes value $\frac{R_i\widetilde{\alpha}_i}{\log(1+P/Wn_i)}$ at $\Lambda_x=n_i f_x(P/Wn_i)$. It is decreasing at $n_i f_x(P/Wn_i)\leq\Lambda_x<\infty$ and goes to zero at $\Lambda_x\to\infty$.

The function $\frac{P/W-f_x^{-1}(\Lambda_x/n_i)n_i}{\Lambda_x+f_x^{-1}(\Lambda_x/n_i)n_i}$ is greater than zeros for $0<\Lambda_x<n_i f_x(P/Wn_i)$ and smaller than zeros for $n_i f_x(P/Wn_i)<\Lambda_x<\infty$. It goes to zero at $\Lambda_x\to\infty$.

Using (C.3) and plugging the expression for $\Lambda_p^*(\Lambda_x)$ into (C.4) we get the following expression for $S_w(\Lambda_x, \Lambda_p^*(\Lambda_x))$,

$$
\begin{aligned}
S_w(\Lambda_x, \Lambda_p^*(\Lambda_x)) &= \frac{1}{|U_D'|} \sum_{i \in U_D'} \frac{W\Lambda_x + P}{\Lambda_x + n_i f_x^{-1}(\Lambda_x/n_i)} \\
&+ \frac{1}{|U_D'|} \sum_{i \in U_D'} \left(\sum_{j \in U_D'} R_j \widetilde{\alpha}_j \left(\frac{n_j (1 + f_x^{-1}(\frac{\Lambda_x}{n_j}))}{\Lambda_x + n_i f_x^{-1}(\Lambda_x/n_i)} - \frac{1}{\log(1 + f_x^{-1}(\frac{\Lambda_x}{n_j}))} \right) \right) \\
&- \frac{1}{|U_D'|} \sum_{i \in U_D'} \left(\sum_{j \in U_R'} r_j^c \left(\frac{n_j (1 + f_x^{-1}(\frac{\Lambda_x}{n_j}))}{\Lambda_x + n_i f_x^{-1}(\Lambda_x/n_i)} - \frac{1}{\log(1 + f_x^{-1}(\frac{\Lambda_x}{n_j}))} \right) \right) \qquad \text{(C.6)}
\end{aligned}
$$

where the set U_D' is defined as $U_D' = \{i \in U_D | \Lambda_p^*(\Lambda_x) - n_i(1 + f_x^{-1}(\frac{\Lambda_x}{n_i}))R_i\widetilde{\alpha}_i\}$

$$
\begin{aligned}
S_w(\Lambda_x, \Lambda_p^*(\Lambda_x)) &= W + \frac{1}{|U_D'|} \sum_{i \in U_D'} \frac{P - W n_i f_x^{-1}(\Lambda_x/n_i)}{\Lambda_x + n_i f_x^{-1}(\Lambda_x/n_i)} \\
&+ \frac{1}{|U_D'|} \sum_{i \in U_D'} \left(\sum_{j \in U_D'} R_j \widetilde{\alpha}_j \left(\frac{n_j f_x^{-1}(\Lambda_x/n_j) - n_i f_x^{-1}(\Lambda_x/n_i)}{(\Lambda_x + n_i f_x^{-1}(\Lambda_x/n_i)) \log(1 + f_x^{-1}(\frac{\Lambda_x}{n_j}))} \right) \right) \\
&- \frac{1}{|U_D'|} \sum_{i \in U_D'} \left(\sum_{j \in U_R'} r_j^c \left(\frac{n_j f_x^{-1}(\Lambda_x/n_j) - n_i f_x^{-1}(\Lambda_x/n_i)}{(\Lambda_x + n_i f_x^{-1}(\Lambda_x/n_i)) \log(1 + f_x^{-1}(\frac{\Lambda_x}{n_j}))} \right) \right) \qquad \text{(C.7)}
\end{aligned}
$$

Appendix D

Energy Efficient Power and Rate Control Fading Channels

D.1 Introduction

A key concern for uplink transmission in wireless networks is energy efficiency. Limited and non-renewable battery supplies in most of the wireless devices require some adaptive transmission schemes that efficiently use these resources. Power control is one of those adaptive schemes. Choice of transmission power has many implications in wireless networking, such as interference, success probability, energy, delay and buffer overflow. The main motivation in the past work on power control was mitigating the effects of interference and fading in order to maximize the achievable capacity (e.g. [67],[68]). The previous studies on power control assumed that there is an infinite number of packets waiting to be transmitted and they concentrated on maximizing the throughput. An important issue that is not considered in the traditional studies on power control is the random characteristic of packet arrivals to the buffer. For instance, considering a limited buffer capacity, if the transmission power is lowered or channel conditions worsen, transmission success rate decreases. When the queue length is close to the capacity, a burst of buffer overflows occurs in case of an arrival burst. In order to minimize energy expenditure in the presence of queueing related constraints such as queueing delay or buffer overflow, power control decisions must also be a function of the queue size, traffic and channel conditions.

In this work we studied power control for transmission through a fading channel. Data packets come randomly from higher network layers and are held in an infinite capacity buffer until they are transmitted. Channel gain is constant during a time slot and varies i.i.d. according to Rayleigh distribution from slot to slot. The node sends its queue size to the base station as a feedback at every time slot. Then the base station decides the optimal number of packets to be transmitted and the node transmits accordingly. The transmitter is able choose from a set of modulation and coding pairs. The performance considerations are average queue size and energy expenditure. Energy efficient transmission has been studied previously for a single user system. For example in [38], the authors studied the problem of minimizing energy expenditure of transmitting randomly arriving packets subject to a transmission deadline constraint in a fading channel. The paper [40] is an extension of [38] that studies joint minimization of delay and energy. In [41] Berry and Gallager obtain structural results that points out a tradeoff between delay and energy in a single user transmission. They show that the optimal power delay curve is convex. They also proposed simple buffer control policies that achieve points on this curve. We have previously considered such a setting and studied optimal power control in a single user channel [69]. The transmitter has two transmission power levels and we proved that the relation between queue size and optimal power control policy is of threshold type. That is, in order to jointly optimize energy expenditure and buffer overflow, the transmitter has to transmit with the higher power level if the queue size is greater than a threshold. The work in [42] extends [41] and finds a closed form expression of optimal policy in terms of the optimal policy when the signal to noise ratio is one. They also find some structural results for the optimal policy and bounds for the optimum cost. In this work we perform

a numerical study based on [42] and investigate the optimal rate control as a function of queue size and channel condition.

D.2 Single User System Model

We assume a user transmitting to a Base Station. We assume an AWGN channel with p.s.d. equal to N_0. The system bandwidth is W Hz. Signal attenuation consists of a constant path loss g and fading. Fading gain process $h(t) \in [0,\infty)$ remains fixed over a time slot and varies i.i.d. according to a Rayleigh distribution with mean μ from slot to slot. Let us quantize fading with thresholds $\{0 = h_1 < h_2, \ldots < h_K = \infty\}$, where $p(k)$ denotes the probability that $h_k \leq h(t) < h_{k+1}$.

We consider a random traffic, where a number of packets of length L_p bits arrive each time slot. Number of packet that arrive in a time slot is Poisson distributed with mean λ. Let $A(a)$ be the probability that a packets arrive. Then $A(a) = \frac{e^{-\lambda}\lambda^k}{k!}$. Let $q(t)$ be the number of packets in the buffer at time slot t, and let $r(t)$ be the amount of packets transmitted in time slot t. Considering the constraint $r[n] \leq s[n]$, the evolution equation can be written as

$$q(t+1) = q(t) + a(t) - r(t) \tag{D.1}$$

Our aim in this system to use minimum power , while achieving a mean delay constraint.

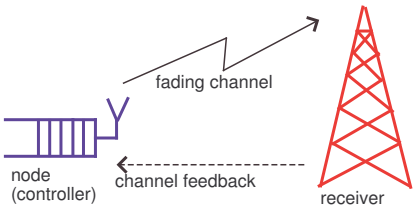

Figure D.1: System Model

D.3 Markov Decision Process Model

D.3.1 Single stage Cost function

The cost of transmitting $r(t)$ units of data ($\frac{r(t)}{T_s}$ rate) at time slot t is a combination of total amount of energy required for transmission, queue size and buffer overflow cost. Let W be the system bandwidth. We use the following rate function.

$$r(t) = T_s W \log\left(1+\beta\frac{p(t)gh(t)}{N_0 W}\right) \tag{D.2}$$

From this formula the power required to transmit r units of packets is

$$P_t(h,r) = \frac{N_0 W}{\beta g h(t)}\left(2^{\frac{r(t)}{T_s W}}-1\right) \tag{D.3}$$

Let $X(t), t \in \mathcal{X} = \{0,1,\ldots\}$ denote a controlled Markov chain with state space $(q(t),h(t)) = X \in \{0,1,2,\ldots\} \times \{h_1,h_2,\ldots,h_K\}$. and let the action space be $r \in \mathcal{R} = \{0,1,2,\ldots,q\} \cap P_t(h,r) \leq P^{max}$, when the queue state is q. We consider the following constrained optimization problem. Find

$$\arg\min_{r\in\mathcal{R}} \sum_{t=1}^{\infty} \alpha^t \frac{N_0 W}{\beta g h(t)}\left(2^{\frac{r(t)}{T_s W}}-1\right) \tag{D.4}$$

subject to

$$\sum_{t=1}^{\infty} \alpha^t q(t) \leq D^{max} \tag{D.5}$$

$$q(t+1) = q(t) - r(t) + a(t) \tag{D.6}$$

Objective function above is convex , while the constraint set is also convex. Therefore writing the Langrange multiplier the single stage cost $c(X,r)$ of transmitting r packets at state $X = (q,h)$ is:

$$c(X,r) = \lambda_d q + \frac{N_0 W}{\beta g h(t)} \left(2^{\frac{r(t)}{T_s W}} - 1 \right) \tag{D.7}$$

Here λ_d is the coefficient of energy cost which is used to adjust its weight in the overall cost. Let π be a policy that generates at time slot t , an action $r(t)$ depending on the history of the process (i.e. decisions at instants $t \in \{1,2,\ldots\}$) , that is a mapping from the state space to the action space. Let Π be the set of all those policies. For a policy $\pi \in \Pi$ and initial state $x \in X$, we define the discounted cost problem with discount factor α. For initial state $x = (s,h)$ define:

$$V_\alpha(X) = \min_{\pi \in \Pi} E_X^\pi \left[\sum_{t=1}^{\infty} \alpha^t g(X(t), r(t)) \right] \tag{D.8}$$

for every $X = (q,h)$ in $\{0,1,2,\ldots,L\} \times \{h_1, h_2, \ldots, h_K\}$ and policy π. It is worth noting that the discount factor a has a practical meaning in the system. Since we have a delay constraint, we need to satisfy a short term rate constraint. Therefore, for a delay constraint D^{max}, choosing $\alpha \leq 1 - \frac{D^{max}}{T_s}$, is reasonable. We can also interpret α as the probability that the communication session terminates in the current time slot. Therefore session duration becomes geometrically distributed.

In order to ensure the existence of the expected infinite horizon discounted cost, it is sufficient that the cost-per-stage function is uniformly bounded, that is $|c(X(t))| < B < \infty$ for all t and $0 <$

$lambda < 1$ [70]. Looking at the single stage cost function (D.7), we need that the system is stable (q(t) finite for all t). In order to satisfy this, it is sufficient that the system is stable if the maximum power is used at all time slots. the condition $E_h\left(T_s W \log\left(1+\beta\frac{P^{max}gh(t)}{N_0 W}\right)\right) <$ $E[a]L$. If $q < \infty$ the following inequality holds:

$$|g(X(t))| \leq \lambda_d q(t) + P^{max} \forall t \tag{D.9}$$

This set of conditions is sufficient for the existence of the solution of the problem in equation (D.8). Well known result in [70] states that optimum discounted cost value function V(.) satisfies the following discounted cost optimality equation:

$$V_\alpha = \min_{0 \leq r \in \mathcal{R}} \left\{ g(X,r) + \alpha \sum_{a=0}^{\infty} \sum_{k=1}^{K} A(a) p_k V_\alpha(q-r+a, h_K) \right\} \tag{D.10}$$

where $X = (q,h)$ is the initial state of the system, $p(k)$ is the probability that $h_k \leq h < h_{k+1}$ and $A(a)$ is the probability that a packets arrive in a time slot. According to (D.10), the cost incurred by choosing an action r, is the sum of the instantaneous cost $g(X,r)$ and the expected cost for the future $\sum_{a=0}^{\infty}\sum_{k=1}^{K} A(a)p_k V_\alpha(q-r+a,h_K)$, multiplied by the discount factor α.

D.4 Analysis of the Discounted Cost Function

We can write the discounted cost optimality equation as:

$$V_\alpha(s,h) = \min_{r\in\mathcal{R}} \left\{ \lambda_d q + \frac{N_0 W}{\beta g h} \left(2^{\frac{rL}{T_s W}} - 1 \right) + \alpha H(u) \right\} \tag{D.11}$$

where $u = q - r$ is the number of packets remaining in the queue after the packets to be transmitted are removed from the buffer and before a new arrival. The function $H(u)$ is defined as

$$H(u) = \sum_{a=0}^{\infty} \sum_{k=1}^{K} A(a) p_k V_\alpha(u+a, h_K) \tag{D.12}$$

Theorem D.1 *H(u) is a convex function*

Proof D.1 *$H(u)$ is a convex combination of $V(u+a,h)$ for different values of a, and h, therefore it is sufficient to show the convexity of this function. Optimum value of the cost function and the optimal policy can be found by the following value iteration.*

$$V_n(s,h) = \min_{0\le r\in\mathcal{R}} \{ \lambda_d q + \frac{N_0 W}{\beta g h} \left(2^{\frac{rLp}{T_s W}} - 1 \right) + \alpha \sum_{a=0}^{\infty} \sum_{k=1}^{K} A(a) p(k) V_{n-1}(q-r+a,h) \} \tag{D.13}$$

We will show it through induction that at every step of the iteration , the value function stays convex.

1. *For $n = 0$, for any r, $V_0(q,h)$ is a convex function. This is because q is convex, and the energy cost is an increasing exponential function of q, which is convex.*

2. *Assume that $V_{n-1}(q,h)$ is convex in q for each h. For a fixed fading level h, let $u(q) = q - r(q)$ be the optimal policy in state $X = (s,h)$ in the n^{th} iteration. Define*

$1-\lambda=\bar{\lambda}$ *and* $q=\lambda q_1+\bar{\lambda}q_2$. *Let the operator* $E_{a,h}(.)$ *denote the averaging with respect to arrivals a and fading in the next slot h, given the current queue size q and fading h.*

We can write the following,

$$\lambda V_n(q_1,h)+\bar{\lambda}V_n(q_2,h)$$
$$=\lambda_d q+\frac{N_0W}{\beta gh}(\lambda 2^{\frac{L_p(q_1-u(q_1))}{T_sW}}+\bar{\lambda}2^{\frac{L_p(q_2-u(q_2))}{T_sW}}-1)$$
$$+\alpha E_{a,h}[\lambda V_{n-1}(u(q_1)+a,h)+\bar{\lambda}V_{n-1}(u(q_2)+a,h)] \quad \text{(D.14)}$$

$$\geq \lambda_p q+\frac{N_0W}{\beta gh}(2^{\frac{L_p}{T_sW}(\lambda(q_1-u(q_1))+\bar{\lambda}(q_2-u(q_2))}-1)$$
$$+\alpha E_{a,h}[V_{n-1}(\lambda(u(q_1)+a)+\bar{\lambda}(u(q_2)+a),h)] \quad \text{(D.15)}$$

$$=\lambda_p q+\frac{N_0W}{\beta gh}(2^{\frac{L_p}{T_sW}(q-\lambda u(q_1)-\bar{\lambda}u(q_2))}-1)$$
$$+\alpha E_{a,h}[V_{n-1}(\lambda u(q_1)+\bar{\lambda}u(q_2)+a,h)] \quad \text{(D.16)}$$

$$\geq \lambda_p q+\frac{N_0W}{\beta gh}(2^{\frac{L_p}{T_sW}(q-u(q))}-1)$$
$$+\alpha E_{a,h}[V_{n-1}(u(q)+a,h)] \quad \text{(D.17)}$$

$$=V_n(\lambda q_1+\bar{\lambda}q_2,h)=V_n(q,h) \quad \text{(D.18)}$$

Here the inequality (D.15) comes from the convexity of the functions 2^x *and* $V_{n-1}(s,h)$ *and the fact that the arrival probability* $A(a)$ *is the same for the states* (q_1,h) *and* (q_2,h). *The inequality (D.17) comes from the optimality of* $u(q)$ *for the state* (q,h). *The last equality comes from the definition. Hence we proved that the function* $V(q,h)$ *is convex*

in q for all h. Therefore as a linear combination of $V(u+a,h)$, $H(u)$ is also a convex function.

Theorem D.2 *The optimal rate allocation policy $r(q,h) = q - u(q,h)$ is nondecreasing in q.*

Proof D.2 *We prove this by contradiction. Assume that $q_1 < q_2$ but $r(q_1) > r(q_2)$. From the optimality equations it follows that:*

$$\lambda_d q_1 + \frac{N_0 W}{\beta g h}(2^{\frac{L_p}{T_s W} r(q_1)} - 1) + \alpha H(q_1 - r(q_1)) < \lambda_d q_1 + \frac{N_0 W}{\beta g h}(2^{\frac{L_p}{T_s W} r(q_2)} - 1) + \alpha H(q_1 - r(q_2)) \quad \text{(D.19)}$$

$$\lambda_d q_2 + \frac{N_0 W}{\beta g h}(2^{\frac{L_p}{T_s W} r(q_2)} - 1) + \alpha H(q_2 - r(q_2)) < \lambda_d q_2 + \frac{N_0 W}{\beta g h}(2^{\frac{L_p}{N_0 W} r(q_1)} - 1) + \alpha H(q_2 - r(q_1)) \quad \text{(D.20)}$$

Adding the two equations, we get,

$$\alpha(H(q_1 - r(q_1)) + H(q_2 - r(q_2))) < \alpha(H(q_1 - r(q_2)) + H(q_2 - r(q_1))) \quad \text{(D.21)}$$

Here there is a contradiction because if the inequalities $q_1 < q_2$ and $r(q_1) > r(q_2)$ are true then the function $\alpha H(u)$ cant be convex, therefore the inequality $r(q_1) > r(q_2)$ is wrong. From this contradiction it is proved that if $q_1 < q_2$ then $r(q_1) \leq r(q_2)$ for all q_1 and q_2. Hence we proved that optimal number of packets to be transmitted is a nondecreasing function of queue size.

The optimal policy π^* can be found by initializing with $V_0(q,h) = 0, \forall q,h$ and finding the maximum in (D.13). Because of the monotonicity of $H(u)$, value iteration converges.

Theorem D.3 *The optimal rate allocation policy $r(q,h) = q - u(q,h)$ is nondecreasing in h.*

Proof D.3 *We prove this by contradiction. Assume that $h_i < h_j$ but $r(q,h_i) > r(q,h_j)$. From the optimality equations it follows that:*

$$\lambda_d q + \frac{N_0 W}{\beta g h_i}(2^{\frac{L_p}{T_s W} r(q,h_i)} - 1) + \alpha H(q - r(q,h_i)) < \lambda_d q + \frac{N_0 W}{\beta g h_i}(2^{\frac{L_p}{T_s W} r(q,h_j)} - 1) + \alpha H(q - r(q,h_j)) \quad \text{(D.22)}$$

$$\lambda_d q + \frac{N_0 W}{\beta g h_j}(2^{\frac{L_p}{T_s W} r(q,h_j)} - 1) + \alpha H(q - r(q,h_j)) < \lambda_d q + \frac{N_0 W}{\beta g h_j}(2^{\frac{L_p}{N_0 W} r(q,h_i)} - 1) + \alpha H(q - r(q,h_i)) \quad \text{(D.23)}$$

Adding these two equations, some of the terms cancel, and we get,

$$\frac{2^{\frac{r(q,h_i)L}{T_s W}} - 1}{h_i} + \frac{2^{\frac{r(q,h_j)L}{T_s W}} - 1}{h_j} < \frac{2^{\frac{r(q,h_j)L}{T_s W}} - 1}{h_i} + \frac{2^{\frac{r(q,h_i)L}{T_s W}} - 1}{h_j} \quad \text{(D.24)}$$

$$h_j \left(2^{\frac{r(q,h_i)L}{T_s W}} - 2^{\frac{r(q,h_j)L}{T_s W}}\right) < h_i \left(2^{\frac{r(q,h_i)L}{T_s W}} - 2^{\frac{r(q,h_j)L}{T_s W}}\right) \quad \text{(D.25)}$$

$$h_j < h_i, \quad \text{(D.26)}$$

which contradicts with $h_i < h_j$, therefore we can conclude that if $h_i < h_j$ then $r(q,h_i) < r(q,h_j)$ and hence $r(q,h)$ is nondecreasing in h.

D.5 Computational Results

In the previous section we found some structural results on the cost function and optimal policy. In this section, we present some computational results for the solution of the power control that verify the above results. In these simulations, value iteration method in (D.13) is used to solve the dynamic programming equation. We consider a single user system with AWGN channel with psd $N_0 = -174dBm$ and Rayleigh fading with mean 1. Transmitter power is $P = 1$ Watt and path loss (in dB) is $-31.5 - 35\log_{10} d + \psi_{dB}$, where d is the distance in meters. We assume a distance of 900m. As for the bandwidth, we consider a single subchannel of an OFDMA system with bandwidth 250 KHz. We consider arrivals of 250-bit packets arriving according to a poisson distribution with rate 0.8 packets/slot (corresponds to 200 kbps).

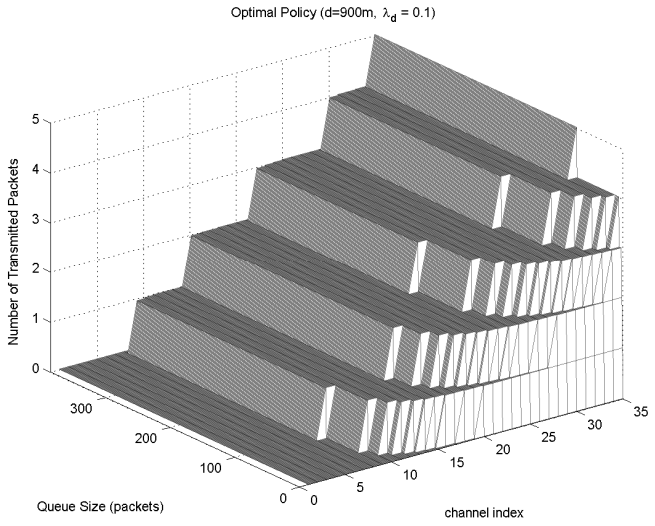

Figure D.2: Optimal number of packets transmitted. Parameters, $\lambda_d = 0.1$

In Figures D.2 and D.3 we observe the result of value iteration for $\lambda_d = 0.1$ and $\lambda_d - 0.12$. We see that optimal rate is nondecreasing w.r.t. queue size and channel gain.

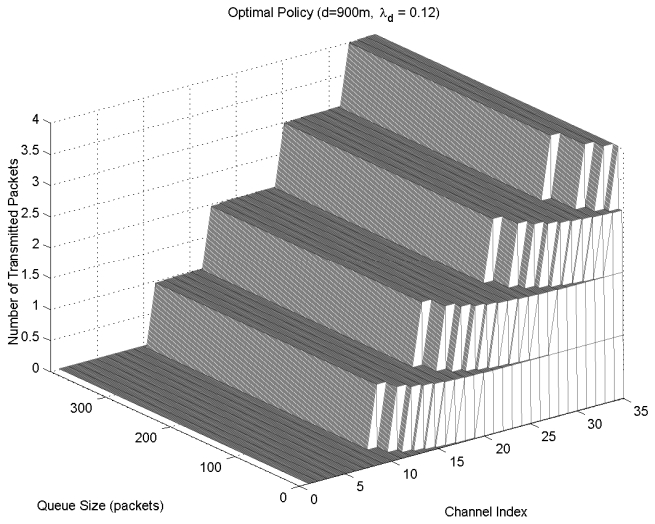

Figure D.3: Optimal number of packets transmitted. Parameters, $\lambda_d = 0.12$

When we increase l_d from 0.1 to 0.12 optimal number of transmitted packets decrease for all queue sizes and channel conditions.

In Figure D.4 we see the average power versus average delay for two different distances. We observe that average power is a decreasing convex function of average delay.

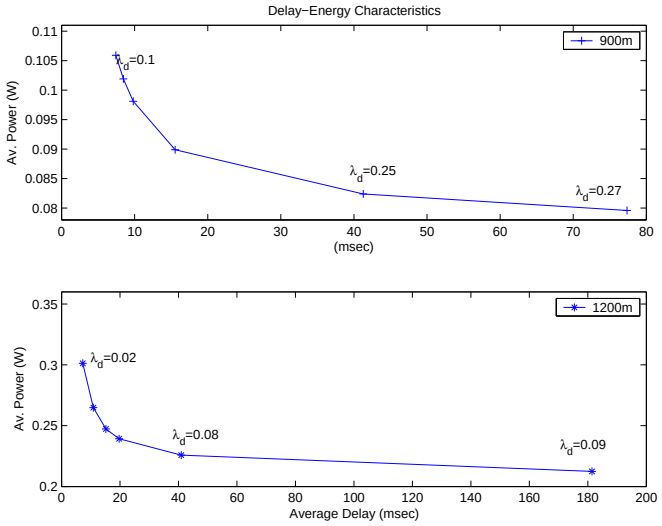

Figure D.4: Average power versus average delay for different λ_d values.

Bibliography

[1] D. Tse, *Forward Link Multiuser Diversity Through Rate Adaptation and Scheduling*, Submitted to IEEE Journal of Selected Areas in Comm. , 2001.

[2] R. Knopp, P.A. Humblet,*Information Capacity and Power Control in Single-Cell Multiuser Communications*, IEEE International Conference on Communications, p.p. 131-135, 1995.

[3] A. Jalali, R. Padowani, R. Pankaj, *Data Throughput of CDMA-HDR: A High Efficiency-High Data Rate Personal Communication Wireless System*, Proc. IEEE Semiannual Vehicular Tech. Conf., May 2000.

[4] *Wimax*, Intel Technology Journal, Vol. 8(3) Aug. 2004. `http://www.intel.com/technology/itj/2004/volume08issue03`

[5] R. Agrawal, R. Berry, *Optimal Scheduling for OFDM Systems*, Presentation, The Maryland Hybrid Networks Center, 2006

[6] M. Einhaus, O. Klein, B. Walke, R. Halfmann, *MAC Level Performance Comparison of Distributed and Adjacent OFDMA Subchannels in IEEE 802.16*, Proc.European Wireless 07,Paris, France, Apr 1-4 2007.

[7] M. Einhaus, O. Klein, B. Walke, R. Halfmann, *Simulative MAC Level Performance Evaluation of an OFDMA System under the Consideration of Frequency Correlated Fading*, Proc. WCNC2007, Hong Kong, Mar. 11-15 2007.

[8] C. Eklund,R. B. Marks, K.L. Stanwood, S. Wang, IEEE *Standard 802.16: A Technical Overview of the Wireless*MAN *Air Interface for Broadband Wireless Access*, IEEE Communications Magazine, June 2002.

[9] IEEE 802.16 2004.,*Amendment to IEEE Standard for Local and Metropolitan Area Networks - Part 16: Air Interface for Fixed Broadband Wireless Access Systems.*, IEEE, Oct. 2004.

[10] A. Ghosh, D. Wolter,J. G. Andres, R. Chen, *Broadband Wireless Access with WiMax/802.16: Current Performance Benchmarks and Future Potential*, IEEE Communications Magazine, Feb. 2005.

[11] IEEE 802.16e, *IEEE Standard for Local and Metropolitan Area Networks, Part 16: Air Interface for Fixed and Mobile Broadband Wireless Access Systesm, Amendment*

2: Physical and Medium Access Control Layers for Combined Fixed and Mobiel Operation in Licensed Bands and Corrigendum 1., IEEE, Feb. 2006.

[12] Wimax Forum, *Mobile WiMAX Part I:A Technical Overview and Performance Evaluation*, Prepared on Behalf of the WiMAX Forum, Mar. 2006.

[13] Kwon, T. and Lee, H. and Choi, S. and Kim, J. and Cho, K. and Cho, S. and Yun, S. and Park, W. and Kim, K.,*Design and Implementation of a Simulator Based on a Cross-Layer Protocol between MAC and PHY Layers in a WiBro Compatible IEEE 802.16e OFDMA System*, IEEE Communications Magazine, Dec. 2005.

[14] Z. Shen, J. G. Andrews, B. L. Evans, *Adaptive resource allocation in multiuser OFDM systems with proportional rate constraints*, IEEE Transactions on Wireless Communications, p.p. 2726-2737, Nov. 2005.

[15] H. Kim, Y. Han, *A Proportional Fair Scheduling for Multicarrier Transmission Systems*, IEEE Communication Letters, p.p. 210-212, Mar. 2005.

[16] G. Song, G. Li, *Utility-Based Resource Allocation and Schedulng in* OFDM-*Based Wireless Broadband Networks*, IEEE Communications Magazine, Dec. 2005.

[17] G.-C. Song and Y. (G.) Li, *Cross-layer optimization for OFDM wireless networks Part I: theoretical framework*, IEEE Transactions on Wireless Communications, vol. 4, no. 2, pp. 614 624, March 2005.

[18] C. Zhu, J. Agre, *Proportional-Fair Scheduling Algorithms for OFDMA-based Wireless Systems*, Preprint, Fujitsu Labs, 2006.

[19] G. Song, Y. (G.) Li, L. J. Cimini, Jr., and H. Zheng, *Joint channel-aware and queue-aware data scheduling in multiple shared wireless channels*, IEEE Wireless Communications and Networking Conference 2004 (WCNC 2004), Mar. 2004, pp.1939 - 1944.

[20] M. Andrews, *Instability of the Proportional Fair Scheduling Algorithm in HDR*, IEEE Trans. on Wireless Communications, p.p. 1422-1426, Sep. 2004.

[21] C. Wong, R. S. Cheng, K. B. Letaief, R. D. Murch, *Multiuser Subcarrier Allocation for OFDM Transmission using Adaptive Modulation*, 49th IEEE Vehicular Technology Conference, p.p. 479-483, 16-20 May 1999.

[22] J. Jang, K. B. Lee, *Transmit power adaptation for multiuser OFDM systems* , IEEE Journal on Selected Areas in Communications, 21(2), Feb 2003 p.p. 171 - 178

[23] W. Rhee, J. M. Cioffi, *Increase in Capacity of Multiuser OFDM System Using Dynamic Subchannel Allocation*, Proc. of 51st IEEE Vehicular Technology Conference, p.p. 1085-1089, 15-18 May 2000.

[24] M. Ergen, S. Coleri, P. Varaiya, *QoS Aware Adaptive Resource Allocation Techniques for Fair Scheduling in OFDMA Based Broadband Wireless Access Systems*, IEEE Transactions on Broadcasting, p.p. 362-270, Dec. 2003.

[25] H. Kim, Y. Han, S. Kim, *Joint subcarrier and power allocation in uplink OFDMA systems*, IEEE Communication Letters, p.p. 526-528, June 2005.

[26] Cendrillon, Raphael and Moonen, Marc, *Dual Optimization Methods for Multiuser OFDM Systems*, In IEEE Global Telecommunications Conference (GLOBECOM), p.p 2334-2338, Dallas, Texas, Nov. 2004

[27] K. H. Kwon, Y. Han and S. Kim, *Optimal Resource Allocation for OFDMA Downlink Systems*, IEICE Transactions on Communications, 2007 E90-B(2):368-371

[28] W. Yu, Lui, R., *Dual methods for nonconvex spectrum optimization of multicarrier systems*, IEEE Transactions on Communications, July 2006, 54(7),p.p. 1310- 1322.

[29] K. Seong , M. Mohseni, J. M. Cioffi, *Optimal Resource Allocation for OFDMA Downlink Systems*, IEEE International Symposium on Information Theory, July 2006, p.p. 1394-1398, Seattle, WA,

[30] A. Stamoulis, N. D. Sidiropoulos, G. B. Giannakis, *Time-varying fair queueing scheduling for multicode CDMA based on dynamic programming*,IEEE TRansactions on Wireless Communications, p.p. 512-523, Mar. 2004.

[31] R. Agrawal, V. Subramanian, R.Berry, *Joint Scheduling and Resource Allocation in CDMA systems*, 2nd Workshop on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt '04), Cambridge, UK, March 24-26 2004.

[32] Abedi, S., *Efficient radio resource management for wireless multimedia communications: a multidimensional QoS-based packet scheduler*, Wireless Communications, IEEE Transactions on, Vol. 4(6), p.p. 2811- 2822, Nov. 2005.

[33] M. Alouini, A. J. Goldsmith,*Adaptive Modulation over Nakagami Fading Channels*, Wireless Personal Communications, May 2000.

[34] X. Qiu, K. Chawla, *On the Performance of Adaptive Modulation in Cellular Systems*, IEEE Trans. on Communications, Vol. 47, p.p. 884-895, June 1999.

[35] Li, L and Sarca, O. FEC Performance with ARQ and Adaptive Burst Profile Selection. IEEE 802.16 Broadband Wireless Access Working Group, Nov. 2001.

[36] A. El-Gamal, C. Nair and Prakbahar,B. and Uysal-Biyikoglu,E. and Zahedi,S., *Energy-Efficient Scheduling of Packet Transmissions over Wireless Networks*IEEE INFOCOM, Jun 2002.

[37] Uysal-Biyikoglu,E. and Prakbahar,B., *Energy-Efficient Packet Transmission Over a Wireless Link*, IEEE Trans. on Networking, Vol. 10, Aug. 2002.

[38] Uysal-Biyikoglu,E. and El Gamal, A. and Prakbahar,B.,*Adaptive Transmission of Variable-Rate Data over a Fading Channel for Energy Efficiency*, In Proc. of IEEE GLOBECOM, p.p. 97-101, 2002.

[39] Nuggehalli,P. and Rao,R. R., *Delay Constrained Energy-Efficient Transmission Strategies for Wireless Devices*, IEEE INFOCOM, Jun. 2002.

[40] E. Yeh and T. Klein, *Optimal trade-off between energy efficiency and average delay*, WiOpt03, Mar. 2003.

[41] R. Berry and R.G. Gallager, *Communication over fading channels with delay constraints*, IEEE Transactions on Information Theory, pages 11351149, May 2002.

[42] Goyal,M. and Kumar, A. and Sharma,V., *Energy Constrained and Delay Optimal Policies for Scheduling Transmission over a Fading Channel*, In Proc. of IEEE INFOCOM, 2003.

[43] I. Emre Telatar and R. G. Gallager, *Combining Queueing Theory with Information Theory for Multiaccess*, IEEE Journal on Selected Areas in Communications 13(6): 963-969 (1995).

[44] P. Gao , S. Wittevrongel, H. Bruneel,*Discrete-time Multiserver Queues with geometric Service Times*, Computers and Operations Research, Elsevier Science Ltd. Oxford, UK, Vol. 31(1), p.p. 81-99, Jan. 2004.

[45] A. J. Goldsmith, S.G. Chua,*Variable Rate Variable Power MQAM for Fading Channels*, IEEE Trans. Comm., 45(10),p.p. 1218-1230, Oct 1997.

[46] D. Bertsekas, R. Gallager, *Data Networks*, Prentice Hall, 1992.

[47] S. Boyd, L. Vanderberghe, *Convex Optimization*, Cambridge University Press, March 8, 2004

[48] IEEE Working Group 802.20, *802.20 Channel Models Document for IEEE 802.20 MBWA System Simulations - 802.20-PD-08*, http://grouper.ieee.org/groups/802/20/Contributions.html, Sep. 2005.

[49] J. B. Andersen, T. S. Rappaport, S. Yoshida, *Propagation Measurements and Models for Wireless Communication Channels*, IEEE Communication Magazine, p.p. 42-49, Jan 1995.

[50] M. Andrews, K. Kumaran, K. Ramanan, S. Stolyar, R. Vijayakumar, P. Whiting, *Providing Quality of Service over a Shared Wireless Link*, IEEE Communications Magazine, February 2001.

[51] IEEE 802.16j PAR, *Amendment to IEEE Standard for Local and Metropolitan Area Networks - Part 16: Air Interface for Fixed and Mobile Broadband Wireless Access Systesm - Multihop Relay Specification*, IEEE, Mar. 2006.

[52] S. Gitzenis, N. Bambos, *Power Controlled Packet Relays in Wireless Data Networks* Globecom 2003, 2003.

[53] A.K. Dinnis, J.S. *Thompson Upper Limits on Performance from two hop relaying in a high data rate cellular system*, Int. Zurich Seminar on Communications (IZS), Feb 22-24 2006.

[54] Z. Jingmei, S. Chunju, W. Ying, Z. Ping,*Performance of a two hop Cellular System with Different Power Allocation Schemes* IEEE Vehicular Technology Conference, 2004.

[55] IEEE 802.16j Relay Task Group,*Channel Models and Performance Metrics for IEEE 802.16j Relay Task Group* IEEE, 2006-05-01.

[56] IEEE 802.16 Broadband Wireless Access Working Group, *Multihop System Evaluation Methodology: Traffic Models* IEEE, 2006-05-01.

[57] G. Song, *Cross-Layer Resource Allocation and Scheduling in Wireless Multicarrier Networks* Ph.D. Dissertation, Georgia Institute of Technology, Apr. 2005.

[58] G. Song, G. Li, *Asymptotic Throughput Analysis for Channel-Aware Scheduling*, IEEE Transactions on Communications, Vol. 54, No. 10, Oct 2006, p.p. 1827-1834

[59] J. Galambos, *The Asymptotic Theory of Extreme Order Statistics*, 1978.

[60] E. Castillo, A. S. Hadi, N. Balakrishnan, J. M. Sarabia,*Extreme Value and Related Models with Applications in Engineering and Science*, Wiley-Interscience, 2004.

[61] L. Kleinrock, *Queueing Systems. Volume I: Theory*, John Wiley & Sons, New York, 1975.

[62] P. V. Mieghem, *The Asymptotic Behavior of Queueing Systems: Large Deviations Theory and Dominant Pole Approximation*, Queueing Systems, Vol. 23, p.p. 27-55, 1996.

[63] C. Bisdikian, J. S. Lew , A. N. Tantawi, *On the tail approximation of the blocking probability of single server queues with finite buffer capacity*, Proc. of the Second International Conference on Queueing Networks with Finite Capacity, p.p. 267-280, 1993.

[64] C. M. Woodside, E. D. S. Ho, *Engineering calculation of overflow probabilities in buffers with Markov-interrupted service*, IEEE Transactions on Communications, Vol. 35, p.p. 1272-1277, 1987.

[65] G. Kramer, M. Gastpar, P. Gupta, *Cooperative Strategies and Capacity Theorems for Relay Networks*, IEEE Trans. Inform. Theory, 2004, p.p. 2020-2040, Jun. 2005

[66] Y. Liang V. V. Veeravalli, *Cooperative relay broadcast channels*, IEEE Trans. Inform. Theory, 53(3) p.p. 900-928, Mar. 2007

[67] R. Yates, *A framework for uplink power control in cellular radio systems. IEEE Journal on Selected Areas in Communications*, 13:13411348, Sep. 1995.

[68] A. Goldsmith and P. Varaiya, *Capacity of fading channels with channel side information*, IEEE Transactions on Information Theory, 43, Nov. 1997.

[69] T. Girici and A. Ephemides, *Optimal Power Control for Wireless Queueing Networks*, 38^{th} Conference on Information Sciences and Systems, March 2004 Princeton,NJ.

[70] Bertsekas,D., *Stochastic Control and Dynamic Programming*, Athena Scientific, 1999.