ABSTRACT

Title of Dissertation: CROSS-LAYER RESOURCE ALLOCATION ALGORITHMS IN WIRELESS NETWORKS WITH ANTENNA ARRAYS

Tianmin Ren, Doctor of Philosophy, 2005

Dissertation directed by: Professor Leandros Tassiulas
Department of Electrical and Computer Engineering

The application of antenna array is a promising approach to improving the capacity of a wireless network. In this dissertation, we study the application of antenna arrays at the base stations (BSs) in a wireless cellular network. We focus on the downlink transmission. This application requires the BSs be aware of the locations and channel conditions of the mobile users. Towards this end, we propose a family of MAC layer protocols that enable a base station to learn the locations and channel conditions of a number of intended users.

Our simulation results demonstrate that the inter-cell interference significantly degrades the system performance of the previously proposed beamforming algorithms in terms of packet loss probability (PLP) in a multi-cell environment. To cope with inter-cell interference, we propose beamforming algorithms that achieve target PLP in the presence of random inter-cell interference.

The application of antenna array on the physical layer has great impact on the protocols of higher layers. Novel MAC algorithms and protocols need to be designed to take advantage of the capacity enhancement provided by antenna array on the physical layer. In this dissertation, the issue of designing a downlink scheduling policy with base station antenna arrays is studied. We derive an optimal scheduling policy that achieves the throughput region. Then, based on the structure of the derived optimal policy, we propose two heuristic scheduling algorithms.

The interference experienced by each node in an ad-hoc network exhibits stochastic nature similar to the inter-cell interference in a cellular network. We propose a power control algorithm in a distributed scheme to achieve target PLP. Furthermore, the proposed power control algorithm is shown to minimize the aggregate transmission power given the PLP constraint.

In the above problems, we mainly consider the non-real-time traffic where throughput is the QoS parameter of concern. On the other hand, delay is an important QoS parameter for real time traffic. In this dissertation, we also consider the scheduling of real time packets by a BS with awareness of physical layer channel conditions of different users.

CROSS-LAYER RESOURCE ALLOCATION ALGORITHMS IN WIRELESS NETWORKS WITH ANTENNA ARRAYS

by

Tianmin Ren

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
2005

Advisory Committee:

Professor Leandros Tassiulas, Chairman
Professor Richard J. La
Professor K. J. Ray Liu
Professor Mark Shayman
Professor Udaya Shankar

©Copyright by

Tianmin Ren

2005

DEDICATION

To My Grandparents

ACKNOWLEDGEMENTS

I would like to thank my advisor, Professor Leandros Tassiulas, for his continuous support and encouragement during my Ph.D study, for leading me into the field of wireless networking and the area of application of antenna arrays.

My thanks also go to my co-advisor, Professor Richard J. La, for his time and great efforts on this research. We worked closely throughout the progress of the problems that are most significant in this thesis. I learned tremendously from his insights, working ethics and endless endeavor to make sense of everything.

I would also like to thank the members of my dissertation committee, Professors K. J. Ray Liu, Mark Shayman and A. Udaya Shankar, for their time and comments.

Personally, I would like to thank my friends for making my time in graduate school enjoyable. Most importantly, I express my deepest gratitude to my family. Their continuous support and patience made this dissertation possible.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

Wireless communication has been experiencing rapid development during the past decade. Increasing demand for fast wireless access and high data rate services has been the driving force for active research in the telecommunications area. Wireless communication systems have been undergoing a transition from the traditional circuit switched voice services to packet switched data services. A variety of data applications have been implemented or proposed to provide mobile users with a ubiquitous access to information.

New network architectures and protocols are proposed to support data applications in wireless networks. A typical architecture in many of current wireless systems, especially cellular networks, provides a wireless access to mobile users through base stations (BSs) or access points (APs) that are connected to the core wireline network. For instance, 3G protocols have been standardized and are being implemented to provide mobile users with wireless data access.

The most challenging task in designing these wireless communication systems is to provide the quality of service (QoS) guarantees to various data applications on wireless channels with limited bandwidth and time varying characteristics. Differ-

ent notions of QoS are available in different communication layers. QoS in physical layer is expressed as an acceptable signal to interference and noise ratio (SINR) or packet loss probability (PLP) at the receiver. In the MAC layer, QoS is usually expressed in terms of achievable goodput. In higher layers QoS can be perceived as a minimum throughput or maximum delay requirement.

The ability of the network infrastructure to fulfill QoS requirements and ultimately enhance system capacity depends on procedures in several layers. In the physical layer transmission power [1], modulation level [2], or forward error correction (FEC) coding rate [3] can be adapted based on channel quality. In the MAC or network layer, QoS guarantees are provided by scheduling or efficient resource management strategies [4].

1.1 Application of base station antenna array in cellular networks

A wide spectrum of approaches have been proposed to reuse the communication resources in time, frequency and/or space domain, in order to provide the QoS guarantees to mobile users and improve the capacity of the wireless networks. Among these approaches, spatial division multiple access (SDMA) with the application of BS antenna arrays, which explores the spatial diversity of mobile users, is considered a more promising one and the last frontier for increasing capacity of wireless networks [20]. This is due to the beamforming capability of the antenna arrays that can form the beam pattern directed to a desired user. This beamforming capability is achieved by adjusting the relative amplitude and phase shift (beamforming weights) of an array of antenna elements. This helps greatly

increase the coverage area of a BS, and suppress co-channel interference such that spatially separable users can share the same channel with their QoS requirements satisfied. Here a channel can be a timeslot in a TDMA system, a subcarrier in an OFDM system, or a code in a CDMA system. In this thesis, we focus on a TDMA system.

Due to the limited computation and communication capability of mobile users, antenna arrays are typically implemented at the BS while each mobile user is equipped with single antenna element.

Sectoring is the simple form of SDMA based on the application of antenna arrays: Each cell is divided into a number of sectors angularly and the BS has a dedicated antenna array for each sector. The same channel can be utilized simultaneously by users in different sectors such that the total system capacity is increased. Sectoring system is extensively employed in practical wireless communication networks because of its simplicity of implementation. However, since the beam patterns are not optimized for each user based on co-channel user locations and current channel conditions, the capacity improvement is limited.

Dynamic beamforming is another implementation of SDMA system and achieves increased capacity by dynamically directing the beams to the scheduled users such that SINR for each user is honored. Since the beam patterns are optimized based on the current co-channel user locations and their channel conditions, the link qualities and hence the system capacity are significantly improved compared with sectoring systems.

In this thesis, we study the dynamic transmission beamforming by BS antenna arrays in a cellular network.

1.1.1 User spatial signature acquisition

Spatial signature that reflects the location and channel condition for a user is required at the BS to calculate the beam pattern in a dynamic beamforming system. This requirement demands users' spatial signatures be known before data transmissions and imposes great challenge to the implementation of dynamic beamforming systems.

In a mobile wireless communication environment where the BS can not assume the knowledge of user locations beforehand, efficient protocols are needed for a BS to acquire user spatial signatures in a timely manner. Spatial signature of a specific user can be derived by the BS through training sequences received from this user. The protocols proposed in the literature [18] [19] dealing with this spatial signature acquisition problem assumed the users are within the broadcast transmission range of the BS. Under this assumption, the BS can broadcast a polling message for a specific user. Upon receiving this broadcast message, the destined user sends back a reply message with training sequence. In this way, the BS obtains the spatial signature of the desired user. However, the assumption that each user is within broadcast coverage range may not hold and thus learning a user's spatial signature can not be achieved through broadcast polling messages, especially in an environment where users move around randomly in a large area.

Fortunately, antenna arrays are capable of significantly extending coverage range of a BS because transmission power can be concentrated in a specific direction through beamforming. The users out of the broadcast range can be reached by the BS using properly formed beam patterns. Due to the uncertainty of user location, the BS may have to send the polling message using sequentially formed narrow beams pointing to different directions until the desired user receives the

polling message and responds. After the array is trained and learns the spatial signature of a specific user then the BS can communicate with this user in distances that are much bigger than the maximum range without antenna array.

In this thesis, we describe protocols for providing media access to users residing in or out of broadcast coverage range of the BS. We consider the class of protocols that are based on directed beamforming and use contention-based or contention-free polling methods to locate users. The proposed protocols can be embedded in existing MAC protocols so as to improve performance.

1.1.2 Downlink beamforming algorithms with inter-cell interference in multiple cell networks

With an antenna array, a BS is able to transmit to a number of users in the same timeslot. For the communications system to function correctly and enhance capacity, the packet loss probability (PLP) of each user should be kept close to some reasonable target value. This is achieved by maintaining the SINR of each user around certain target value that is determined by the relation between PLP and SINR expressed in link curves.

The total interference experienced by a user is the sum of intra-cell interference and inter-cell interference. The intra-cell interference is caused by the transmissions by the same BS to other co-cell users, and is determined by the channel condition of the user to its assigned BS and the beamforming weights and transmission powers at the BS. Therefore, the BS is aware of the intra-cell interference of each user in its cell. On the other hand, the inter-cell interference is due to the transmissions by BSs in neighboring co-channel cells, and is determined by the channel conditions to these BSs and their beamforming weights and transmission powers. Since the

BSs transmit to the users in their respective cells independently, a BS is unable to predict the inter-cell interference that a user in its own cell receives. Moreover, the inter-cell interference experienced by a user is a random variable since the BSs typically select different groups of users for transmission in different timeslots and the channel conditions vary with time.

Of all the works in the literature, the inter-cell interference is ignored or assumed constant. Therefore, the SINR of each user is achieved exactly as calculated by the beamforming algorithms. However, we will show in this thesis that the performance degrades greatly in terms of much larger than target PLP by the inter-cell interference if the beamforming algorithms do not take inter-cell interference into account in a multiple-cell environment. This performance degradation calls for the design of practical beamforming algorithms that address the randomness of the inter-cell interference and achieve target PLP in the presence of inter-cell interference.

In this thesis, we will first derive the expression of the time average PLP as a function of the distribution of the inter-cell interference. Based on this expression, we propose to compensate for the random inter-cell interference by aiming at a PLP smaller than the target PLP and use the sum of average inter-cell interference and noise to replace the noise term in the beamforming algorithms. This beamforming algorithm is shown to achieve target PLP for different scheduling algorithms in various channel conditions. Furthermore, it is displayed that the inter-cell interference possesses weak temporal correlation and is closely approximated by a log-normal distribution in a wide spectrum of scheduling and beamforming algorithms with different channel conditions. From this observation, we propose the second beamforming algorithm that sequentially calculates the transmission pow-

ers of the users to achieve target PLP based on the distribution of the inter-cell interference. This algorithm is shown to achieve target PLP as well.

1.1.3 Optimal scheduling with BS antenna array

The implementation of antenna arrays in physical layer improves system capacity and raises new problems in upper layers in the meanwhile. New algorithms have to be implemented to fully exploit the potential performance improvement provided by antenna arrays.

We investigate the downlink scheduling problem with the goal of stabilizing the queues of the users served by a central controller that coordinates the transmissions of a number of BSs. Packets arrive at the central controller from backbone network for transmission to different users. With antenna array, each BS can transmit to more than one users in the same channel, provided that the SINR requirement is satisfied at the receiver of each scheduled user. From the upper layer point of view, this system can be modeled as a queueing system with multiple servers, and the scheduling policy is the decision rule to select one feasible set of users to serve in each channel, with the goal to stabilize the system.

We will establish the conditions under which the system is stabilizable by some scheduling policy. Furthermore, we will rely on the negative drift of Lyapunov function to prove the optimalily of a scheduling policy that achieves stability if the system can be stabilized. However, the complexity of this optimal scheduling policy is exponential in the number of users. To overcome implementation difficulties, we propose two heuristic scheduling policies that achieve satisfactory performance with significantly lower complexity, and study the complexity vs. performance tradeoff in both single cell and multiple-cell environments.

1.2 Power control with distributed scheduling in ad-hoc networks

The stochastic nature of interference exists in the ad-hoc networks as well. In an ad-hoc network, centralized scheduling and power control is difficult to achieve because of the time varying network topology and traffic condition. The computation and communication overhead is prohibitive for centralized coordination. Therefore, the nodes have to carry out scheduling and power control in a distributed manner, giving rise to random, unpredictable interference experienced at each node.

In this thesis we first illustrate the shortcomings of previous physical layer models for simulation in the presence of unknown interference at the receivers. Using the physical layer model based on link curves, we develop a new power control algorithm that can provide physical layer quality-of-service (QoS) in the form of PLP.

We then formulate the problem of minimizing the average aggregate transmission power as an optimization problem, and show that our proposed power control algorithm converges to a solution of the optimization problem.

1.3 QoS provisioning to real-time traffic in wireless networks

For non real-time traffic, throughput is the most significant measure of performance and the non real-time applications are considered to be delay-tolerant. Algorithms aimed at throughput maximization are studied extensively. On the contrary, for real-time traffic delay is the most important QoS measure and real-time applica-

tions demand timely delivery of packets. For wireless communications systems, the time varying nature of wireless channels makes this requirement even more difficult to satisfy.

Users in a wireless communications system experience significantly different channel conditions due to different distances to the BS, multi-path fading and shadowing effect. In addition, link quality varies with time for each user due to environment change or user mobility. This multi-user diversity can be exploited to enhance system capacity. Because of the independence of the wireless links and the asynchronous nature of channel variations for different users, the BS is able to select the users with relatively good instant link quality to serve [8]. Hence, efficient QoS provisioning requires that MAC layer functions be aware of the physical layer characteristics. On the other hand, the scheduling function in the MAC layer determines the bandwidth sharing on the packet level. This sharing should reflect the higher layer QoS requirement in terms of bounded delay or guaranteed throughput. Therefore, efficient scheduling strategy has to take both upper layer QoS requirement and physical link characteristics into consideration.

In this thesis, we study the scheduling of real time packets to multiple users over time varying wireless channels, subject to packet delivery deadline constraints [7]. We show that this problem can be cast as a Markov decision process (MDP). Performance bounds and design guidelines for practical scheduling algorithms are obtained through analysis and simulations. Moreover, we will show that the asynchronous time variance of wireless channels is beneficial for performance enhancement.

1.4 Organization

The rest of this thesis is organized as follows. In Chapter 2, we propose the spatial signature acquisition protocols and analyze their performance. We study the beamforming algorithms in a multi-cell network in Chapter 3. Chapter 4 investigates the downlink scheduling algorithms with BS antenna arrays. The power control algorithm with distributed scheduling in ad-hoc networks is presented in Chapter 5. We present the problem of real-time packet scheduling in Chapter 6. We conclude the thesis and identify directions for future research in Chapter 7.

Chapter 2

Efficient Media Access Protocols for Wireless Cellular Networks with Antenna Arrays

2.1 Introduction

Space division multiple access (SDMA) with antenna arrays at the base stations constitutes perhaps the most promising means for ensuring QoS and increasing system capacity [20]. SDMA enables intra-cell channel reuse by several spatially separable users by pointing a beam towards the direction of an intended user and nulling out other users.

The employment of antenna arrays at the physical layer affects resource allocation methods and protocols of higher layers, *e.g.,* MAC layer. In order to exploit the benefits of SDMA, the base station needs to know the location and channel condition of each user, which are captured by its spatial signature.

In reception mode on the uplink, the base station obtains the information about

spatial signatures of the users by making use of the preamble of received packets. Each preamble is used by the base station to train the antenna array to compute the beamforming weights that effectively steer the beam towards the intended user.

In [21], the authors described a slotted ALOHA system with a single-beam adaptive antenna array at the base station. The users that attempt to access the channel in a timeslot start transmission with random time offset. By exploiting a pseudo-random sequence in the packet preamble, the base station computes a beam and locks it onto the first received packet in a timeslot, while nulling out the subsequently received packets in the same timeslot. A similar system with multi-beam capabilities is presented in [22]. The base station again uses packet preambles to form a beam for each received packet from a different user, so that several users are captured. Uplink access to a base station with an antenna array with the help of a modified carrier sense multiple access (CSMA) protocol is proposed in [23].

In transmission mode on the downlink, the base station can request information about the spatial signature of a user by broadcasting a polling message intended for that user [24]. Upon reception of the polling message, the user transmits a given sequence of symbols. The base station measures the received signal and uses it to compute the spatial signature and steer a beam towards the direction of the user.

The common characteristic of these approaches is that they are all designed for users that reside within the omni-directional transmission range of the base station. When required, the base station broadcasts polling requests to users and receives response packets from users within broadcast range by having its antenna in omni-directional mode. It then uses packet preambles to steer beams to appropriate directions. In this setting, the basic feature of SDMA to extend coverage

range essentially remains unexploited. In this thesis, we address the problem of extending the coverage range of a base station with antenna array, by devising efficient media access protocols. Such protocols are primarily meant for detecting the locations of the users that are out of broadcast range, but they can also be integrated with existing media access protocols that are designed for coping with users within broadcast range. We present protocols that use directed beamforming and employ contention-free or contention-based polling methods to acquire location information of users. In devising the protocols, some essential characteristics of the IEEE 802.11 standard for wireless local area networks (WLANs) [25] are adopted. However, our treatment is general enough to encompass other wireless networks as well. The proposed protocols are based on idealized model. The implementation of these protocols needs further consideration of practical communication environment.

The proposed protocols can also be applied to cellular networks where users experience large channel variations. When the channel condition deteriorates, it is difficult for the base station to communicate with the user in omni-directional transmission/reception mode, even though the user is in the broadcast range of the base station. However, the base station can direct its beam pattern towards the user with antenna array. In this way, the channel condition can be greatly improved such that the base station and the user with bad channel condition are able to communicate.

This chapter is organized as follows. In Section 2.2 we provide the model and main assumptions. In Section 2.3 we present our protocols and analyze their performance. Numerical results are given in Section 2.4. Finally, Section 2.5 concludes this chapter and identifies future research.

2.2 System Model

We consider the downlink of a single base station and focus on downlink access to N users. The base station is equipped with an array of M antenna elements. The broadcast range of the base station is determined by a maximum transmission power level when the array operates in simple omni-directional mode where only one antenna element is used for transmission or reception.

A beam formed by the base station is specified by its beam width δ and its angular position ϕ. The space can be covered by B beams of beam width $\delta = 360/B$ degrees and each user is covered by one beam.

The location and channel condition of a user are captured by its spatial signature. We assume that the user association phase with the base station has been completed, so that the base station knows the number of users and their identities but not their locations. Packetized data arrive from higher layer queues for transmission over the channel. If the base station uses beamforming and not broadcasting to transmit data to users, it needs to know their spatial signatures.

The base station can obtain the spatial signature of a user by using the following two methods.

1. *Contention-free polling*: The base station first sends a polling message that contains the identity of the intended user. Upon receiving the polling message, the user responds by sending a known sequence of bits on the uplink. The base station uses these bits to train the antenna array so as to steer the beam towards the direction indicated by the spatial signature of the user. This polling/response method is used to acquire the spatial signature of each user. We refer to this method as contention-free polling, since it does not involve any kind of user transmission contention.

2. *Contention-based polling*: The base station can acquire information about user spatial signature by sending a polling message that is not intended for a specific user. If the message is received by more than one users, their simultaneous responses will collide at the base station. The base station then initiates a contention resolution procedure to resolve users and obtain their spatial signatures. We refer to this method as contention-based polling.

The base station can send polling messages with omni-directional or directional transmission. After the spatial signature acquisition process is completed, data can be transmitted to users.

2.3 Media Access Protocols with Base Station Antenna Array

2.3.1 Problem statement

When the base station needs to obtain information about the spatial signature of a user residing within its broadcast range, it can poll the user by using broadcast or directed transmission with contention-free or contention-based polling. Contention-free broadcast polling is the method that results in the smallest time delay in locating the user.

However, when the user is out of broadcast range, it cannot be reached by a simple broadcast transmission. The base station needs to concentrate all transmission power into a narrower directed beam so that it reaches the user. This arising issue concerns the polling protocol that should be devised, such that the base station acquires the spatial signatures of the users out of broadcast range in a

fast and reliable manner. Towards this end, the base station can use beamforming to send polling messages with long range directed transmissions. The base station sequentially steers the beam towards different directions, so that the entire space is covered. In this case we have maximum range polling through successive directed transmissions that scan the space. The objective of the protocol is to locate all users as fast as possible.

The base station can select between contention-free and contention-based polling with directed transmissions in order to locate users out of broadcast range. In contention-free polling, the space is successively scanned by a beam until the user is located and the procedure is repeated for all users. In contention-based polling, the space is successively scanned by a beam in a different direction and the contention among users in a beam is resolved before proceeding to the next beam. The absence of contention in contention-free polling is the advantage of this method over contention-based polling. However, the time consumed in scanning the space to locate each user separately may be larger than the corresponding time with contention-based polling.

A significant issue that arises in contention-based polling is the width of the beam that scans the space. If a large beam width is used, fewer beams are needed to scan the space and the required time to scan the space with successive directed transmissions is smaller. However, with a large beam width, the number of users that receive the message is larger on average and hence the contention resolution for users in a beam lasts longer. From that point of view, a large beam width does not contribute to reduction of time delay to locate all users. A similar tradeoff holds for small beam widths as well.

We address the problem of extending the coverage range of the base station by

providing media access to users that are located out of its broadcast range. We describe contention-based and contention-free protocols and analyze their performance with respect to several involved parameters.

2.3.2 Contention-based polling with directed transmissions

When contention-based polling is employed, the base station forms successive directed beams and scans all the space. The base station attempts to locate all users within a beam before proceeding to the next beam. For now assume that the base station employs directed transmission for all users, regardless if they are in or out of broadcast range.

Time is divided into intervals that are referred to as contention resolution intervals (CRI). Each CRI consists of L timeslots. Before the beginning of a CRI, the base station sends a polling message by using directed transmission. The polling message does not contain the identity of any user. Each user that is illuminated by that beam receives the message and responds by sending back a polling acknowledgement (P_ACK) message that contains a preamble and the user identity. If only one user sends a P_ACK, the message is received correctly by the base station and the spatial signature of the user is obtained with the help of the preamble. In that case, the base station informs the user that its spatial signature is known, by sending it an ACK message with its identity. However, if there are multiple users in the beam, their P_ACK messages collide at the base station. The base station then does not issue an ACK message to the users in the beam and the users are informed about the collision and the upcoming contention resolution process.

A simple method is used for resolving the collisions: Each user with a collided P_ACK re-transmits with probability p in each of the subsequent L timeslots in

the CRI. If one user happens to transmit alone in one slot, the message of this user is resolved. Then, the base station informs the user that its P_ACK and hence its spatial signature have been obtained by sending an ACK message to the user, so that this user stops transmission in the next timeslots. If no user sends a P_ACK message in the timeslot following the polling message, the base station assumes that no user is located in the beam or that all users have been resolved in previous timeslots. It then proceeds by forming the next beam. Here we assume that the base station is able to distinguish contention from absence of transmission by measuring the received signal power. If the CRI expires and the base station does not have any indication that all users have been resolved, it initiates the next CRI by sending a polling message again. The procedure is repeated for the remaining unresolved users, until the base station has an indication that all users in the beam are resolved.

Our assumption for using a fixed re-transmission probability p in each timeslot is justified as follows. Assume that there exist n unresolved users in a beam. The probability that one user transmits in a timeslot and therefore succeeds in transmission is

$$p_s(n) = np(1-p)^{n-1}$$

. This probability is maximized for $p^* = 1/n$, which depends on the number of unresolved users. Ideally, the base station could instruct the users to re-transmit with probability p^*, so as to improve the chance of a successful transmission. The problem is that the base station is not aware of the number of users in a beam and therefore it does not know the number of unresolved users at each step of the procedure. Thus, we resort to a fixed value p.

Let us now compute the expected time $d(n)$ to obtain the spatial signatures

of n users in a beam. Let X_p, X_{p_a} and X_a denote the transmission time of poll, P_ACK and ACK message respectively. Define $p_{i,n,L}$ as the probability that i out of n users have already been resolved successfully in L contention resolution timeslots. Then, $p_{1,n,1} = p_s(n)$ and $p_{i,n,L} = 0$ if $i > n$ or $i > L$. For all other cases, $p_{i,n,L}$ can be computed with the recursive equation

$$p_{i,n,L} = p_s(n)p_{i-1,n-1,L-1} + (1 - p_s(n))p_{i,n,L-1}.$$

For the time delay $d(n)$, we have

$$d(0) = X_p$$

and

$$d(1) = X_p + X_{p\_a} + X_a$$

For $n \geq 2$,

$$d(n) = \sum_{k=0}^{n} p_{k,n,L} \left[X_p + X_{p\_a} + LX_{p\_a} + kX_a + d(n-k)\right].$$

In the beginning of a CRI, a polling message is sent and the polling response from users (collided or not) is received. If k out of n users are resolved in L contention timeslots, this means that the base station has sent k ACKs to resolved users. The term $d(n-k)$ accounts for the fact that $n-k$ users still need to be resolved.

We now compute the expected time $D(N, B)$ to resolve all N users and obtain their spatial signatures when the space is covered by B successive directed transmissions. Let $q_{i,N,B}$ be the probability that i out of N users reside in a beam. Assuming that users can reside in each of the B beams with probability $1/B$, $q_{i,N,B}$ is given by

$$q_{i,N,B} = \binom{N}{i} \left(\frac{1}{B}\right)^i \left(1 - \frac{1}{B}\right)^{N-i}$$

and the delay $D(N, B)$ can be computed recursively as

$$D(N, B) = \sum_{i=0}^{N} q_{i,N,B} \left[d(i) + D(N - i, B - 1) \right] \ .$$

where the first term in the brackets denotes the delay to resolve i users in a beam and the second term indicates that $N - i$ users need to be resolved in the remaining $B - 1$ beams.

2.3.3 Contention-free polling with directed transmissions

When contention-free polling is used, the base station again forms successive directed beams to poll users. However, polling messages now include the identity of a user and are intended for that user. The base station attempts to locate one user by sequentially scanning the space with successive directed transmissions. The base station starts by sending a polling message for a user in a beam. If the user does not reside in the beam, the base station does not receive any reply and proceeds with the formation of the next beam to locate the user. If the user is found to reside in a beam, it responds by sending a P_ACK message. Upon receiving P_ACK, the base station finds its spatial signature and sends an ACK message to the user to inform it that its location is found. The base station then starts scanning the space for another user. The order in which users are sought is arbitrary.

The advantage of this scheme is the absence of contention among users in a beam, since only one user responds to the polling message. The expected delay $D'(N, B)$ to obtain the spatial signatures of N users when covering the space with B beams is,

$$D'(N, B) = N \left(\frac{B + 1}{2} X_p + X_{p\_a} + X_a \right) \ . \tag{2.1}$$

Indeed, for each user the base station issues $(B+1)/2$ polls on average, receives one polling response when the user is located and sends one ACK to the user.

2.4 Numerical Results

2.4.1 Setup

We consider a scenario where N users are uniformly distributed in an area around a base station, so that they can reside either in or out of the base station broadcast range. The base station needs the spatial signatures of all users and can poll users with omni-directional or directional transmission. At each time one beam can be formed towards a certain direction and the area around the base station is covered by B beams. For users within broadcast range, the base station may select to poll users by broadcasting or directional beamforming and can use contention-based or contention-free polling scheme. For users out of broadcast range, the base station can use only beamforming to poll users. The transmission time of the polling, P_ACK and ACK messages are chosen to satisfy the ratios $X_p : X_{p\_a} : X_a = 1 : 2 : 1$. This selection is justified by the fact that the P_ACK message has an additional preamble for spatial signature acquisition. CRIs consist of L slots. The intervals between transmission of polling messages, reception of polling acknowledgements and transmission of ACKs are not considered in the analysis.

2.4.2 Comparative results

The performance measure is the time delay until the spatial signatures of all users are acquired. We evaluate the performance of the following four schemes for spatial signature acquisition:

- Broadcasting/Beamforming (Broad/Beam) schemes: The base station uses contention-free broadcast polling for users in broadcast range and uses polling with beamforming for users out of range.

  In the Broad/Beam schemes, the base station first broadcasts the contention-free polling messages to locate each user. A user within the broadcast range responds to the polling message destined to it by sending back P_ACK message. Then the base station acquires the spatial signature of this user and sends ACK message to acknowledge the reception of the P_ACK message. When each user is polled by broadcasting polling message, for the users that are located out of the transmission range, the base station needs to use directive transmission to resolve their locations with contention-based or contention-free polling scheme. The delays are

$$D_1 = NX_p + (N - N_{out})(X_{p\_a} + X_a) + D(N_{out}, B)$$

  and

$$D_2 = NX_p + (N - N_{out})(X_{p\_a} + X_a) + D'(N_{out}, B)$$

  for contention-based and contention-free Broad/Beam scheme respectively, where N_{out} is the number of users out of the broadcast range.

- Beamforming/Beamforming (Beam/Beam) schemes: The base station uses polling with beamforming for all users, regardless if they reside in or out of broadcast range. The polling can again be contention-based or contention-free.

In Fig. 2.1, we illustrate the performance of the aforementioned schemes for $N = 20$ users for the cases of $B = 5$ and $B = 15$ beams. The time delay is

plotted as a function of the number of users that reside out of broadcast range, N_{out}. A first observation is that the performance of the Beam/Beam contention-based and contention-free schemes is independent of N_{out}, since these schemes treat users residing in and out of the broadcast range the same. The time delay for contention-free Broad/Beam scheme increases linearly with N_{out}, as can be seen from (2.1). For $B = 5$, the Broad/Beam and Beam/Beam contention-free schemes perform better than corresponding contention-based ones. This is because the small value of B results in fast enough contention-free polling and because beams are wide enough, so that time latency due to user contention is large. When $N_{out} < 14$, Broad/Beam contention-free scheme yields the best performance, while Beam/Beam contention-free scheme is preferable in all other cases. When $B = 15$, the behavior is reversed, namely contention-based schemes incur smaller delay than the contention-free ones. The large value of B makes contention-free polling time-consuming, while at the same time user contention within each beam becomes low, since beams are narrow. Broad/Beam contention-based polling yields the smallest delay when $N_{out} < 17$, while Beam/Beam with contention achieves the best performance in all other cases.

2.5 Discussion

We addressed the problem of improving the channel quality of the users and extending the coverage range of the base station with beamforming. Our ultimate goal is to design protocols that can be integrated into existing polling protocols that were originally designed for omni-directional transmission. We considered the class of prototype media access protocols with contention-based and contention-free polling and evaluated their performance in terms of required time for the base

station to acquire the spatial signature of each user.

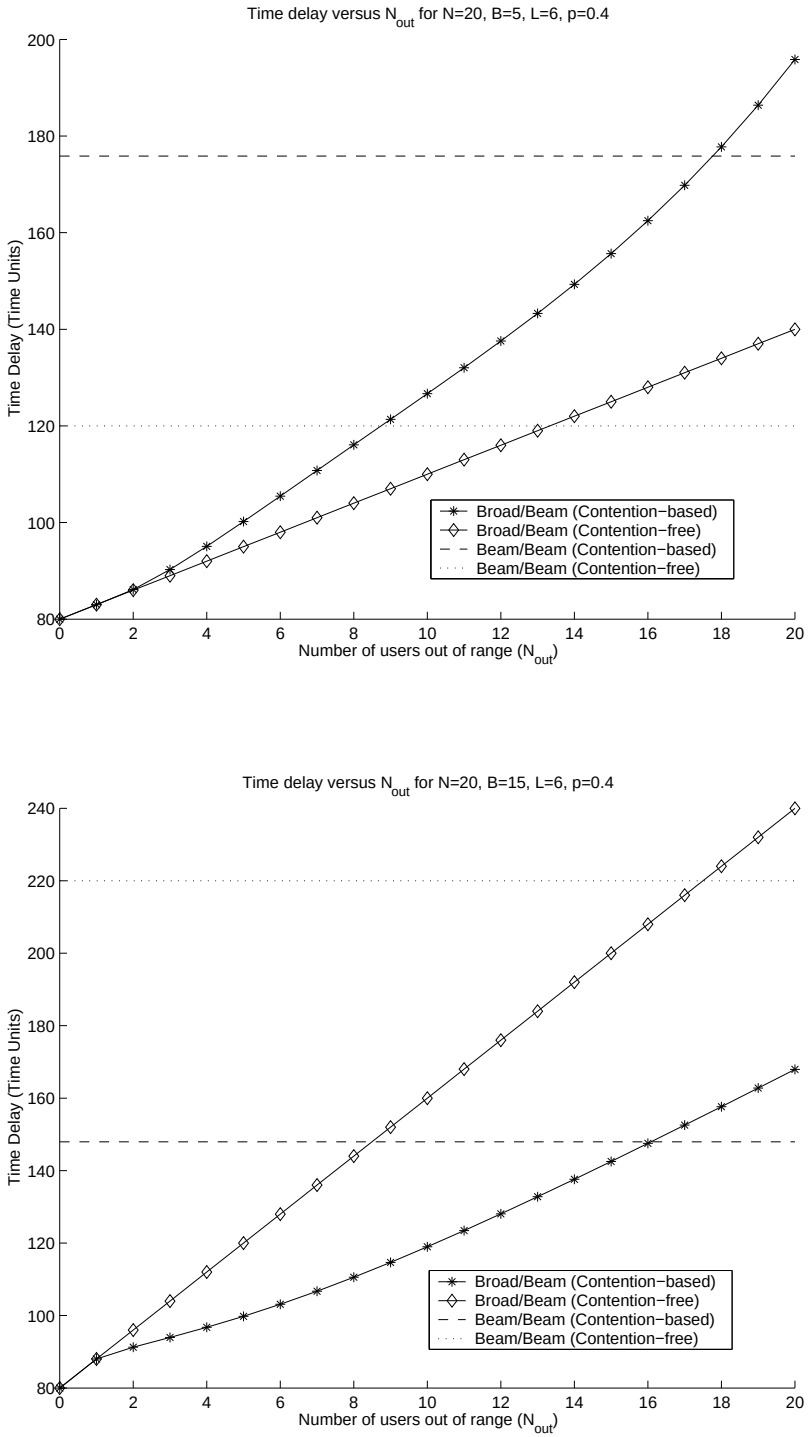


Figure 2.1: Time delay as a function of number of users out of broadcast range for $N = 20$ users when $B = 5$ and $B = 15$ beams scan the space

Chapter 3

Beamforming Algorithms with Inter-cell Interference in Multi-cell Networks

3.1 Introduction

Previous research on the downlink (from base stations to mobile users) dynamic beamforming problem in a cellular network can be categorized into two classes. The first class of research is on the physical layer: Given a set of users, the problem is to design algorithms for calculating the beamforming weights and transmission power for each user. The problem is typically modeled as an optimization problem, where the objective is to minimize the total transmission power subject to the constraint that each user's SINR requirement is satisfied. Note that this problem may be infeasible, that is, there may not exist a set of beamforming weights and transmission powers that satisfy the minimum SINR requirement of all users. In [38] iterative algorithms are proposed to minimize total transmitted power subject

to the constraint that the SINR requirement of each user is satisfied for downlink transmissions in a single cell network. Boche and Schubert propose another class of iterative algorithms in [39], where the solution is divided into two phases. In the first phase the feasibility of a set of users is tested. For a feasible set of users, the total transmission power is minimized in the second phase. In [27], the problem of joint beamforming and base station assignment is considered, where each user can be served by any base station in the network. An algorithm that assigns each user to an optimal base station and computes the corresponding transmission beam pattern for each user is designed to minimize the total transmission power of the base stations.

A similar system where a number of users can be transmitted in the same timeslot and there is interference among the selected users is studied in [64] with a stochastic programming formulation.

The second class of research focuses on MAC layer with physical layer user separability constraints. The goal of this class of research is to maximize the number of scheduled users with the SINR requirement of each scheduled user satisfied given a set of users. Algorithms aimed at maximizing the number of scheduled users are proposed in the literature [40]. These algorithms are based on the same idea of inserting users into a channel in a sequential manner, and vary in the criteria that determine the order in which users are inserted. This problem is extended to be combined with other multi-user access schemes such as TDMA, OFDM and CDMA in [31].

In all of the previous works, the focus has been on either (i) a single cell network where inter-cell interference is neglected or (ii) a centralized network where the calculation of the beamforming weights and transmission powers for a number

of base stations is conducted by a central controller. In the latter the inter-cell interference is computed by the central controller and made available for the beamforming algorithms. However, this requires the channel information of each mobile user to every base station be available at the central controller. This assumption is not practical as resource allocation algorithms become more sophisticated and computationally demanding and such computational requirement and intelligence is moved to the BSs from a *centralized* resource manager to reduce the computational burden on the resource manager and communication overhead between BSs and the resource manager.

To the best of our knowledge there is a lack of a careful investigation of the effect of the inter-cell interference on the performance in a multiple cell environment, especially in a packet switched network. In this thesis, we investigate the performance of a class of scheduling and beamforming algorithms in a multi-cell environment where each BS has only channel information of the users in its own cell. We demonstrate that the algorithms that do not account for inter-cell interference result in unacceptable performance in the presence of inter-cell interference. This calls for the design of an algorithm that can handle the time-varying random inter-cell interference.

This problem has been studied in [34] with a *single* antenna element under the assumption that a packet broken into smaller blocks requires many *consecutive* transmissions over the wireless channel, leading to strong temporal correlation. Such an approach, however, will not work when the temporal correlation is weak. For example, with increasing network capacity, the transmission of a packet may require only a few or even just one transmission in the future, resulting in considerably weaker temporal correlation.

We propose two beamforming algorithms that account for inter-cell interference without the assumption of strong temporal correlation. The first algorithm is based on the derived expression of the average packet loss probability (PLP) as a function of the target SINR. Our algorithm is shown to achieve the target PLP using general link curves and different scheduling algorithms under various channel models.

We also characterize the inter-cell interference experienced by a user under different settings. Numerical results suggest that the distribution of the inter-cell interference can be well approximated by a log-normal distribution. In addition, we demonstrate that the temporal correlation of the inter-cell interference is rather weak.

Based on the observation that the inter-cell interference can be well modeled as a log-normal random variable (rv), the second algorithm uses the estimated parameters of the log-normal distribution when computing transmission powers of the scheduled users to handle the random inter-cell interference. We show that this algorithm also achieves the target PLP.

The chapter is organized as follows. We describe the multi-cell network model under investigation in Section 3.2 and introduce optimal beamforming algorithms for a single cell in Section 3.3. Section 3.4 demonstrates the performance degradation of these algorithms in the presence of inter-cell interference. We derive the average PLP as a function of the target SINR in Section 3.5. The first proposed algorithm is outlined in Section 3.6. Section 3.7 characterizes the amplitude distribution of inter-cell interference and shows its temporal correlation. We study the trade-off between throughput and target PLP in Section 3.8. Second proposed algorithm is given in Section 3.9. We conclude in Section 3.10.

3.2 Multiple Cell Network Model

In this section we describe the multiple cell network model that is used for our analysis throughout the chapter and introduce the beamforming algorithm for a single cell.

3.2.1 Network layout

We consider a network that consists of 7 co-channel cells shown as the shaded cells in Fig. 3.1. The co-channel cells of a cell can be found by (1) moving i cells along any chain of hexagons, (2) turning 60 degrees counter-clockwise, and (3) moving j cells (pp. 28 [37]). In Fig. 3.1, $i = 2$ and $j = 1$, and a total of $i^2 + i \cdot j + j^2$ cells share the available spectrum. We call this pair (i, j) the reuse pattern. The radius of each cell is denoted by R.

Time is assumed to be divided into contiguous equal-sized timeslots. We assume that a packet can be accommodated within a timeslot, and the duration of a timeslot is assumed to equal the transmission time of a packet. In recent years, especially with the emergence of 3G/4G technologies, the capacity of a wireless system has increased significantly. This trend is likely to continue in the future, and with a high capacity a timeslot will be able to accommodate the transmission of an entire packet (*e.g.,* TCP segment). Furthermore, even when a packet does not fit into a timeslot and is segmented into several smaller Protocol Data Units (PDUs), service providers may prefer to adopt a block level scheduling algorithm that spreads out the transmission of PDUs belonging to the same packet in order to reduce the delay jitter experienced by the end users.

One base station is located at the center of a cell and transmits packets to N users that are uniformly distributed in the cell. We assume that every user always

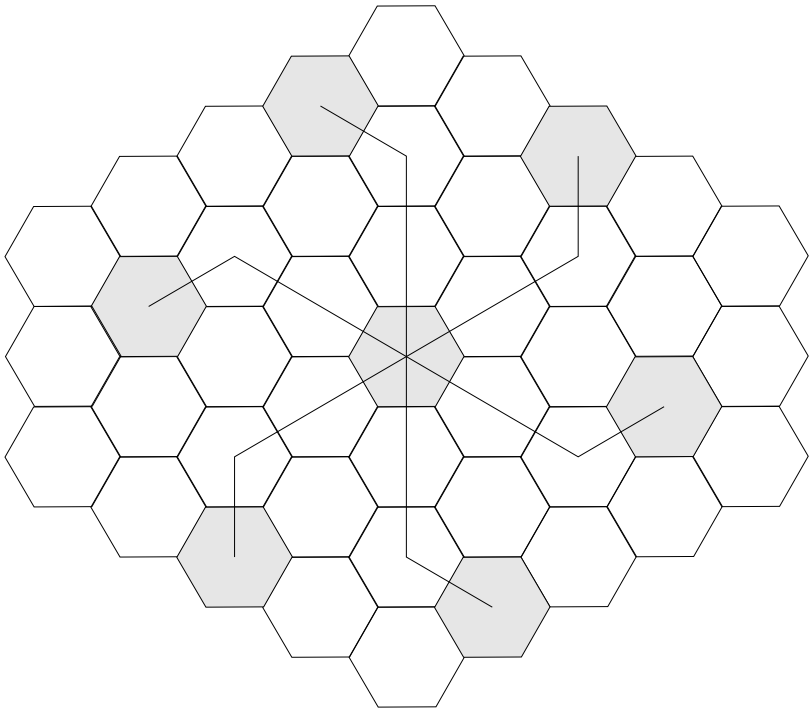


Figure 3.1: Co-channel cells

has a packet ready for transmission when scheduled, *i.e.,* infinite traffic model. In each timeslot, each base station schedules a set of users for transmission, and calculates the beamforming weights and transmission powers for the scheduled users.

Although we model a network with 7 co-channel cells, we focus on the cell at the center as we are interested in the performance of the system with inter-cell interference. The inter-cell interference experienced by a user in a cell is often approximated by the interference from the 6 closest co-channel cells multiplied by some scaling constant larger than one. In this thesis we approximate it only with the interference of the closest co-channel cells. However, as one will see, the performance of our proposed algorithms will not be affected by the assumption.

3.2.2 Channel model

In this subsection, we present the adopted multi-path wireless channel model [29]. Each base station is equipped with an antenna array, where M antenna elements are uniformly located on a circle of radius r. The multi-path channel between a given user and the m-th antenna element of a given base station is expressed as

$$h^m(t) = \sum_{\ell=1}^{L} g_\ell \, \delta\left(t - \tau_\ell + \tau_\ell^m\right),$$

where L is the number of paths, g_ℓ is the complex gain of the ℓ-th path, and τ_ℓ is the delay for that path with respect to the first antenna element with $m = 1$. The gain g_ℓ is a complex random variable with zero mean and variance A_ℓ. The term $\tau_\ell^m = (r/c)(\cos\theta_\ell - \cos(2\pi(m-1)/M - \theta_\ell))$ captures the delay to the m-th antenna with respect to the first antenna, where θ_ℓ is the angle of arrival of the ℓ-th path of the user, and c is the electro-magnetic wave propagation speed. We assume each path results from the reflection by a scatterer, and L scatterers are distributed within a circle of radius R' centered at the user, and are uniformly distributed both in distance and in angle with reference to the user.

The signal received by the user from the base station is given by

$$y(t) = \sqrt{p} \sum_{m=1}^{M} w^m \sum_{\ell=1}^{L} g_\ell e^{j\omega\tau_\ell^m} s(t - \tau_\ell)$$

where p is the transmission power, w^m is the beamforming weight of the m-th antenna element, and $s(t)$ is the signal. Note that here we implicitly assume that τ_ℓ^m is much smaller than the transmission time of one symbol. $\tau_\ell \gg \tau_\ell^m$ and the sum $\tau_\ell + \tau_\ell^m \approx \tau_\ell$. Beamforming vector $\mathbf{w} = [w^1, w^2, \cdots, w^M]^T$ satisfies $\mathbf{w}^H\mathbf{w} = 1$. The $M \times 1$ antenna steering vector $\mathbf{v}(\theta_\ell)$ at direction θ_ℓ is defined to be $[e^{j\omega\tau_\ell^m}; m = 1, \ldots, M]$, where ω is the carrier frequency. The vector $\mathbf{a} = \sum_{\ell=1}^{L} g_\ell \mathbf{v}(\theta_\ell)$ is called the spatial signature of the user and captures the spatial and multi-path

properties. The expected received signal power is given by

$$\begin{aligned}
&\mathbf{E}\left[\left|\sqrt{p}\sum_{m=1}^{M} w^m \sum_{\ell=1}^{L} g_\ell e^{j\omega\tau_\ell^m} s(t-\tau_\ell)\right|^2\right] \\
&= p\mathbf{w}^H \sum_{\ell_1=1}^{L}\sum_{\ell_2=1}^{L}\Big(\mathbf{v}(\theta_{\ell_1})\mathbf{v}^H(\theta_{\ell_2})\mathbf{E}\left[g_{\ell_1} g_{\ell_2}^*\right] \\
&\qquad \times \mathbf{E}\left[s(t-\tau_{\ell_1})s^*(t-\tau_{\ell_2})\right]\Big)\mathbf{w} \\
&= p\mathbf{w}^H \mathcal{H}\mathbf{w}
\end{aligned}$$

Observe that

$$\mathbf{E}\left[g_{\ell_1} g_{\ell_2}^*\right] = \begin{cases} 0, & \text{if } \ell_1 \neq \ell_2 \\ A_\ell, & \text{if } \ell_1 = \ell_2 = \ell \end{cases}$$

assuming that all paths are independent since $\mathbf{E}\left[g_\ell\right]$ is assumed to be zero. By assuming signal power is normalized, we have,

$$\mathcal{H} = \sum_{\ell=1}^{L} A_\ell \, \mathbf{v}(\theta_\ell)\mathbf{v}^H(\theta_\ell)\,.$$

The matrix $\mathcal{H}$ is called a spatial covariance matrix and in general has a rank larger than one.

We denote the spatial covariance matrix of user j with respect to base station i as $\mathcal{H}_j^i$ and the base station assigned to user j as i_j. The SINR of user j, denoted by $SINR_j$, is given by

$$SINR_j = \frac{S_j}{I_j^{intra} + I_j^{inter} + n_j^2} \tag{3.1}$$

where S_j, I_j^{intra} and I_j^{inter} are the signal power, intra-cell interference, and inter-cell interference received by user j, respectively, and n_j^2 is the noise power at user j. Denote the set of users in the 7 co-channel cells we consider by $\mathcal{U}^\star$. For each user

j,

$$S_j = p_j \left(\mathbf{w}_j^H \mathcal{H}_j^{i_j} \mathbf{w}_j \right)$$
$$I_j^{intra} = \sum_{\substack{k \in \mathcal{U}^\star \\ k \neq j, i_k = i_j}} p_k \left(\mathbf{w}_k^H \mathcal{H}_j^{i_k} \mathbf{w}_k \right)$$
$$I_j^{inter} = \sum_{\substack{k \in \mathcal{U}^\star \\ k \neq j, i_k \neq i_j}} p_k \left(\mathbf{w}_k^H \mathcal{H}_j^{i_k} \mathbf{w}_k \right)$$

where p_j and $\mathbf{w}_j$ are transmission power and beamforming vector, respectively, for user $j \in \mathcal{U}^\star$.

3.3 Optimal beamforming for a single cell

Since each base station has only the channel information of the users in its own cell and the base stations independently transmit to the users in their respective cells, a beamforming algorithm at a base station only can compute the received signal strength and intra-cell interference for the scheduled users. In this subsection, we describe a beamforming algorithm [39] that equalizes relative SINRs of the users in a single cell under the assumption that the noise power at each user available to the base station is time invariant, *i.e.,* constant. This algorithm consists of two phases; in the first phase, the minimum relative SINR among the scheduled users, which is the ratio of the achieved SINR to some target SINR, is maximized. This is equivalent to finding the largest common relative SINR η_c^* under a power budget constraint $||\mathbf{p}||_1 = P_{max}$ where $\mathbf{p}$ is the transmission power vector.[1] Let $\mathbf{W}$ denote the ensemble of beamforming vectors for all users, *i.e.,* $\mathbf{W} = \{\mathbf{w}_j, j \in \mathcal{U}\}$, where $\mathcal{U}$ is the set of scheduled users and $|\mathcal{U}| = U$. The problem of finding η_c^* can be

[1] Here $||\cdot||_1$ denotes a L_1 norm.

formulated as

$$\max_{\mathbf{W},\mathbf{p}} \min_{j\in\mathcal{U}} \frac{SINR_j}{\gamma_j} \tag{3.2}$$
$$\text{subject to } ||\mathbf{p}||_1 = P_{max}$$

where γ_j is the target SINR for user j and $SINR_j$ is the SINR of user j given by

$$SINR_j = \frac{p_j \mathbf{w}_j^H \mathcal{H}_j \mathbf{w}_j}{\sum_{i\in\mathcal{U}, i\neq j} p_i \mathbf{w}_i^H \mathcal{H}_j \mathbf{w}_i + n_j^2} \ . \tag{3.3}$$

For simplicity of notation we omit the superscript of the covariance matrix $\mathcal{H}_j^i$ since we focus on a single base station. A set of users can be scheduled with their respective SINR requirement satisfied if $\eta_c^* \geq 1$, and a set of users that satisfies this condition is called a *feasible set*.

The second phase of the algorithm attempts to minimize the total transmission power subject to SINR requirement given a feasible set. This will be discussed in more details shortly.

First, we discuss how to compute η_c^* and hence decide if a set of users is feasible. Define a $U \times U$ matrix $\mathbf{\Psi}(\mathbf{W}) = [\psi_{ij}, i, j \in \mathcal{U}]$ where ψ_{ij} is the interference caused by user j to user i per unit power given by

$$\psi_{ij} = \begin{cases} 0, & i = j \\ \mathbf{w}_j^H \mathcal{H}_i \mathbf{w}_j, & i \neq j \end{cases} \ .$$

Let

$$\Upsilon(\mathbf{W}, P_{max}) = \begin{bmatrix} \mathbf{D}\Psi(\mathbf{W}) & \mathbf{D}\sigma \\ \frac{1}{P_{max}} \underline{1}^T \mathbf{D}\Psi(\mathbf{W}) & \frac{1}{P_{max}} \underline{1}^T \mathbf{D}\sigma \end{bmatrix}$$

where $\sigma = [n_1^2, \cdots, n_U^2]^T$, $\underline{1} = [1, 1, 1, \cdots, 1]^T$, and

$$\mathbf{D} = \text{diag}\Big\{ \frac{\gamma_1}{\mathbf{w}_1^H \mathcal{H}_1 \mathbf{w}_1}, \cdots, \frac{\gamma_U}{\mathbf{w}_U^H \mathcal{H}_U \mathbf{w}_U} \Big\} \ .$$

The matrix $\Upsilon(\mathbf{W}, P_{max})$ is called the extended downlink coupling matrix. The last row of $\Upsilon(\mathbf{W}, P_{max})$ accounts for the maximum power constraint.

A system in which all users achieve a common SINR ratio η_c and $||\mathbf{p}||_1 = P_{max}$ is described by the set of linear equations

$$\Upsilon(\mathbf{W}, P_{max})\mathbf{p}_{ext} = \frac{1}{\eta_c}\mathbf{p}_{ext}$$

where $\mathbf{p}_{ext} = [\mathbf{p};\ 1]$. Thus, η_c is the reciprocal of an eigenvalue of $\Upsilon(\mathbf{W}, P_{max})$. The matrix $\Upsilon(\mathbf{W}, P_{max})$ has the property that only the maximum eigenvalue $\lambda_{max}(\Upsilon(\mathbf{W}, P_{max}))$ satisfies the requirement that the eigenvalue and every element of the corresponding eigenvector are strictly positive [41]. Thus, we have $1/\eta_c = \lambda_{max}(\Upsilon(\mathbf{W}, P_{max}))$. If η_c^* is the maximum possible common relative SINR, then

$$\eta_c^* = \frac{1}{\min\limits_{\mathbf{W}} \lambda_{max}(\Upsilon(\mathbf{W}, P_{max}))} = \frac{1}{\lambda_{max}^*} \tag{3.4}$$

Virtual uplink problem

The downlink SINR of a user in (3.3) depends on the beamforming weights and transmission powers of other users. Therefore, the optimization of beamforming weights and transmission powers for different users is coupled and difficult to solve directly. Alternatively, we can solve a virtual uplink problem. It has been shown [39] that if the noise power is the same on both uplink and downlink for every user, then the solution of the following optimization problem, denoted by η_c^{U*}, equals η_c^*.

$$\begin{aligned} &\max_{\mathbf{W},\mathbf{q}} \min_{j\in\mathcal{U}} \frac{SINR_j^U}{\gamma_j} \\ &\text{subject to } ||\mathbf{q}||_1 = P_{max} \end{aligned} \tag{3.5}$$

where $\mathbf{q}$ is the uplink power vector, and $SINR_j^U$ is the uplink SINR for user j defined by

$$SINR_j^U = \frac{q_j \mathbf{w}_j^H \widetilde{\mathcal{H}}_j \mathbf{w}_j}{\mathbf{w}_j^H (\sum_{i \in \mathcal{U}, i \neq j} q_i \widetilde{\mathcal{H}}_i + I) \mathbf{w}_j}$$

where I is the $M \times M$ identity matrix, and

$$\widetilde{\mathcal{H}}_j = \mathcal{H}_j / n_j^2 \ . \tag{3.6}$$

Since $SINR_j^U$ depends on other users only through the transmission power, this problem is typically easier to solve than (3.2).

The minimum total transmission power that can satisfy the SINR requirement of all scheduled users is the same for both the virtual uplink problem and the downlink problem. Hence, rather than solving (3.2) we can solve the virtual uplink problem in (3.5) for η_c^*. Moreover, the beamforming vectors that achieve η_c^{U*} also achieve η_c^* for the downlink problem in (3.2) as well [39].

Given a feasible set, we can minimize the total transmission power with the given SINR constraint on the virtual uplink. The resulting optimal beamforming weights minimize the total transmission power on the downlink as well. Using this set of optimal beamforming weights with the constraint that $\eta_c = 1$, we can calculate the optimal transmission power vector on the downlink.

We describe how one can solve the virtual uplink problem to compute η_c^*. First, define the extended uplink coupling matrix similar to $\Upsilon(\mathbf{W}, P_{max})$:

$$\Lambda(\mathbf{W}, P_{max}) = \begin{bmatrix} \mathbf{D}\Psi^T(\mathbf{W}) & \mathbf{D}\sigma \\ \frac{1}{P_{max}} \underline{1}^T \mathbf{D}\Psi^T(\mathbf{W}) & \frac{1}{P_{max}} \underline{1}^T \mathbf{D}\sigma \end{bmatrix}$$

Here normalized covariance matrices $\widetilde{\mathcal{H}}_j$, $j \in \mathcal{U}$ are used in calculating $\mathbf{D}$ and $\mathbf{\Psi}$ in place of $\mathcal{H}_j$, and σ reduces to $\underline{1}$.

If all users achieve a common relative SINR η_c^U, we have

$$\Lambda(\mathbf{W}, P_{max})\mathbf{q}_{ext} = \frac{1}{\eta_c^U}\mathbf{q}_{ext}$$

where $\mathbf{q}_{ext} = [\mathbf{q};\ 1]$, and η_c^{U*} is given by

$$\eta_c^{U*} = \frac{1}{\min\limits_{\mathbf{W}} \lambda_{max}(\Lambda(\mathbf{W}, P_{max}))}$$

Beamforming and power control algorithms

The following iterative algorithm that explores the idea of virtual uplink was proposed in [39] to find η_c^* and is referred to as the FEASIBILITY algorithm, where the beamforming weights and transmission powers are calculated alternatively to improve η_c^U until it reaches η_c^{U*} (or η_c^*) on the virtual uplink.

ALGORITHM I: FEASIBILITY(P_{max}; $\mathcal{H}_j, n_j^2, \gamma_j$, $\forall j \in \mathcal{U}$)

STEP 1: Set $t = 0$, $\mathbf{q}^{(0)} = [0, \cdots, 0]^T$, and $\lambda_{max}^{(0)} = \infty$.

STEP 2: While 1, do

- Set $t \leftarrow t + 1$. Solve a set of U generalized eigenproblems:

$$\mathbf{w}_j^{(t)} = \arg \max_{||\mathbf{w}_j||=1} \frac{\mathbf{w}_j^H \widetilde{\mathcal{H}}_j \mathbf{w}_j}{\mathbf{w}_j^H \mathcal{R}_j(\mathbf{q}^{(t-1)})\mathbf{w}_j}, \forall j \in \mathcal{U} \tag{3.7}$$

  where

$$\mathcal{R}_j(\mathbf{q}^{(t-1)}) = \sum_{k \in \mathcal{U}, k \neq j} q_k^{(t-1)} \widetilde{\mathcal{H}}_k + I\ . \tag{3.8}$$

  The solutions to the above generalized eigenproblems are given by the dominant generalized eigenvectors of the matrix pairs $\left[\widetilde{\mathcal{H}}_j, \mathcal{R}_j(\mathbf{q}^{(t-1)})\right]$ for all $j \in \mathcal{U}$.

- Find the largest eigenvalue $\lambda_{max}^{(t)}$ of $\Lambda(\mathbf{W}^{(t)}, P_{max})$ and the corresponding eigenvector $\mathbf{q}_{ext}^{(t)}$ of the form $\mathbf{q}_{ext}^{(t)} = [\mathbf{q}^{(t)};\ 1]$

- If $\lambda_{max}^{(t-1)} - \lambda_{max}^{(t)} \leq \epsilon$, break, endif

STEP 3: If $\lambda_{max}^{(t)} \leq 1$, *i.e.,* the set of users can be scheduled in the same timeslot, output 1. Otherwise, output 0.

It is shown in [39] that the sequence of eigenvalues $\{\lambda_{max}^{(t)}\}$ is monotonically decreasing and converges to the global minimum λ_{max}^{*}, which is related to the maximum common SINR ratio η_c^* by equation (3.4).

Once a set of users is determined to be feasible for scheduling in the same timeslot by the FEASIBILITY algorithm, it is beneficial to minimize the total transmission power $P_{sum} = ||\mathbf{p}||_1$ to the users. This is because minimization of P_{sum} reduces the interference to the users in neighboring cells and thus helps maximize the total system throughput. Towards this end, we adopt the following algorithm called MINIMIZE_POWER introduced in [39].

ALGORITHM II: MINIMIZE_POWER(P_{max}; $\mathcal{H}_j, n_j^2, \gamma_j, \forall j \in \mathcal{U}$)

STEP 1: Set $t = 0$, and $\mathbf{q}^{(0)} = [0, \cdots, 0]^T$.

STEP 2: While 1, do

- Set $t \leftarrow t + 1$. Solve a set of U generalized eigenproblems,

$$\mathbf{w}_j^{(t)} = \arg \max_{||\mathbf{w}_j||=1} \frac{\mathbf{w}_j^H \widetilde{\mathcal{H}}_j \mathbf{w}_j}{\mathbf{w}_j^H \mathcal{R}_j(\mathbf{q}^{(t-1)}) \mathbf{w}_j}, \forall j \in \mathcal{U}. \tag{3.9}$$

- Compute the uplink transmission power vector that achieves common relative SINR $\eta_c^U = 1$, which is given by

$$\mathbf{q}^{(t)} = (I - \mathbf{D}\mathbf{\Psi}^T(\mathbf{W}^{(t)}))^{-1}\mathbf{D}\underline{1}\,.$$

- If $||\mathbf{q}^{(t-1)}||_1 - ||\mathbf{q}^{(t)}||_1 \leq \epsilon$, break, endif

STEP 3: Compute the optimal downlink transmission power vector

$$\mathbf{p}^{(t)} = (I - \mathbf{D}\mathbf{\Psi}(\mathbf{W}^{(t)}))^{-1}\mathbf{D}\underline{1}$$

STEP 4: Output $\mathbf{p}^{(t)}, \mathbf{W}^{(t)}$.

It is shown in [39] that the sequence of total transmission powers $\{||\mathbf{q}^{(t)}||_1\}$ is monotonically decreasing and converges to the global power minimum.

3.4 Performance Degradation due to Inter-cell Interference

In this section we illustrate the effect of the inter-cell interference on the packet loss probability (PLP). In this thesis we select the PLP as the right physical layer QoS parameter. Adoption of PLP as a meaningful physical layer QoS parameter, instead of achieved SINR, is more natural as the performance of higher layer protocols does not depend directly on the achieved SINR, but *indirectly* through the achieved PLP. Moreover, as mentioned earlier, unlike in the case where the inter-cell interference can be accurately estimated/predicted, achieving certain target SINR as a physical layer QoS parameter is not possible in a multi-cell wireless network in the absence of centralized resource allocation.

For the numerical examples presented here we adopt the following scheduling algorithm at each base station [26]. Denote the sets of co-cell users and scheduled users as $\mathcal{N}$ and $\mathcal{U}$, respectively. The throughput of user j up to the beginning of timeslot t is given by $T_j(t)$ for all $j \in \mathcal{N}$. Each user j is assigned a credit c_j at the beginning of simulation.

ALGORITHM III: SCHEDULING_1($P_{max}; c_j, T_j(t), \mathcal{H}_j, n_j^2, \gamma_j, \forall j \in \mathcal{N}$)

STEP 1: Initialize $\mathcal{U} = \emptyset$ and $\mathcal{N}_1 = \mathcal{N}$.

STEP 2: While $\mathcal{N}_1 \neq \emptyset$,[2] do

- $j^* = \arg\min_{j \in \mathcal{N}_1} T_j(t)/c_j$
- $\mathcal{U} = \mathcal{U} \cup \{j^*\}$, $\mathcal{N}_1 = \mathcal{N}_1 \setminus \{j^*\}$
- if FEASIBILITY$(P_{max};\, \mathcal{H}_j, n_j^2, \gamma_j,\, \forall j \in \mathcal{U}) = 0$, then $\mathcal{U} = \mathcal{U} \setminus \{j^*\}$.

STEP 3: MINIMIZE_POWER$(P_{max}; \mathcal{H}_j, n_j^2, \gamma_j, \forall j \in \mathcal{U})$

Note that this scheduling algorithm attempts to equalize the normalized throughput $T_j(t)/c_j$ of the users. Hence, these credits c_j are used to provide differentiated classes of service in terms of the provided rates.

Suppose that the beamforming weights and transmission powers are calculated for the scheduled users at each base station using the above algorithm. Then, we can calculate the actual SINR value of each scheduled user that includes intra-cell and inter-cell interference. Once SINR values are computed, packet losses are determined by link curves. A link curve gives the probability of packet loss as a function of SINR for a given modulation and/or coding scheme. The target SINR value γ_j of each $j \in \mathcal{U}$ is determined by the target PLP and the link curves.

We conduct the experiment with randomly generated scenarios. Although several reuse patterns have been studied, here we only present the results for the reuse pattern of $(1, 1)$. The noise power is normalized to $n_j^2 = 1$ for every user j, and other parameters are scaled accordingly. The values of the parameters are listed in Table 3.1, where λ is the wavelength of the carrier electro-magnetic wave.

[2]Typically no more than M users can be scheduled simultaneously due to limited degrees of freedom created by M antenna elements, and the algorithm may terminate if M users are already scheduled.

(i,j)	$(1,1)$	P_{max}	10^{15}	PLP_{target}	2%
N	15	M	6	r	λ
R	1000	R'	100	L	6

Table 3.1: Simulation parameters.

For time varying channels, the variance of the channel gain $\{A_\ell(t); t = 1, 2, \ldots\}$ is a stochastic process. The rv $A_\ell(t) = (s_\ell(t) f_\ell(t))^2 / d_\ell^\kappa$, where $s_\ell(t)$ and $f_\ell(t)$ are sequences of log-normal and Rayleigh rvs, respectively, accounting for slow shadow fading and fast fading. The variable d_ℓ denotes the distance from the base station to the user along the ℓ-th path and κ is the path loss exponent and is set to be 3.5.[3] For studying the performance of the proposed algorithms, we consider the following four types of wireless channels in this thesis:

1. independent and identically distributed (i.i.d.) shadow fading channel model:

$$A_\ell(t) = s_\ell^2(t)/d_\ell^\kappa = e^{2r_\ell(t)}/d_\ell^\kappa$$

   where $\{r_\ell(t), t = 1, 2, \cdots\}$ is a sequence of i.i.d. Gaussian rvs with mean 0 and standard deviation 1.07. The corresponding values of $\mathbf{E}[s_\ell(t)]$ and $\sqrt{Var[s_\ell(t)]}$ are 1.78 and 2.61, respectively.

2. Time correlated shadow fading channel model:

$$A_\ell(t) = s_\ell^2(t)/d_\ell^\kappa = e^{2r_\ell(t)}/d_\ell^\kappa$$

   where $\{r_\ell(t)\}$ is a sequence of Gaussian rvs generated by $r_\ell(t+1) = (1-\rho) r_\ell(t) + \rho u_\ell(t)$ where $\{u_\ell(t)\}$ is a sequence of i.i.d. Gaussian rvs with mean

[3]Different path loss exponents yield similar qualitative results, although the numbers vary from one value to another.

0. The parameter ρ is set to 0.1, and the variance of $u_\ell(t)$ is selected so that $r_\ell(t)$ has standard deviation of 1.07 similarly as in the i.i.d. shadow fading channel model.

3. Rayleigh fading channel model:

$$A_\ell(t) = f_\ell(t)^2 / d_\ell^\kappa$$

where $\{f_\ell(t)\}$ is a sequence of i.i.d. Rayleigh rvs, and thus $\{f_\ell^2(t)\}$ is a sequence of i.i.d. exponential rvs. The parameters are selected so that $f_\ell^2(t)$ has mean 1 and standard deviation 1.

4. Time correlated shadow fading plus Rayleigh fading channel model:

$$A_\ell(t) = (s_\ell(t) f_\ell(t))^2 / d_\ell^\kappa$$

where the sequence $\{s_\ell(t)\}$ is generated in the same manner as in the time correlated shadow fading channel, $\{f_\ell(t)\}$ is a sequence of i.i.d. Rayleigh rvs with the same distribution as in the Rayleigh fading channel model.

The link curve we adopt is for the lowest transmission rate in a TDMA system, which is the left most curve in Fig. 3.2 [33]. The corresponding modulation scheme is binary offset quadrature amplitude modulation (B-O-QAM) [28]. The measured data points are shown as '*' and the solid curves are the fitted curves, which will be explained in Section 3.6.

The experimental results are shown in Fig. 3.3 for various channel models with a single class of service, *i.e.,* all users have the same credit. The experimental results are similar with multiple classes of service. Note that the realized PLPs are significantly higher than the target PLP, which is set to 2 percent in our experiment (shown as the solid horizontal line in the figure), because the inter-cell

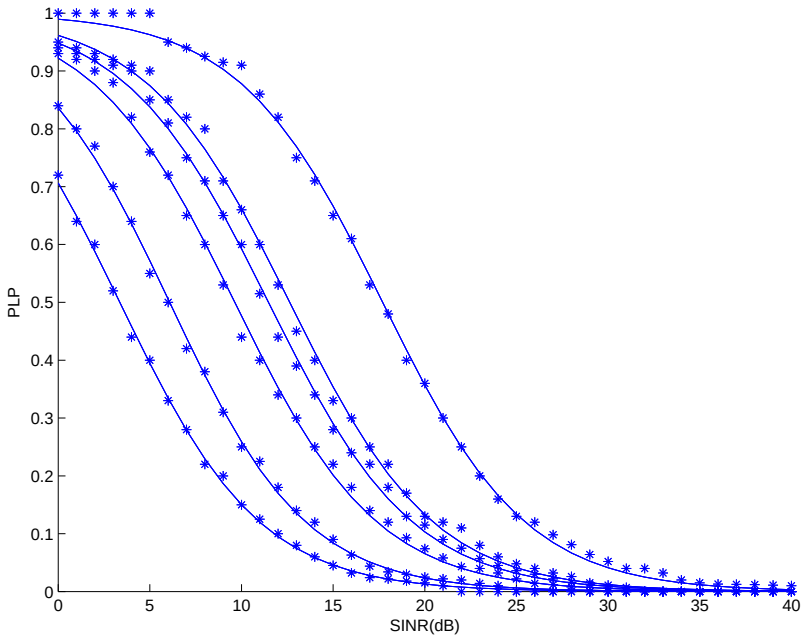


Figure 3.2: Link curves of a TDMA system.

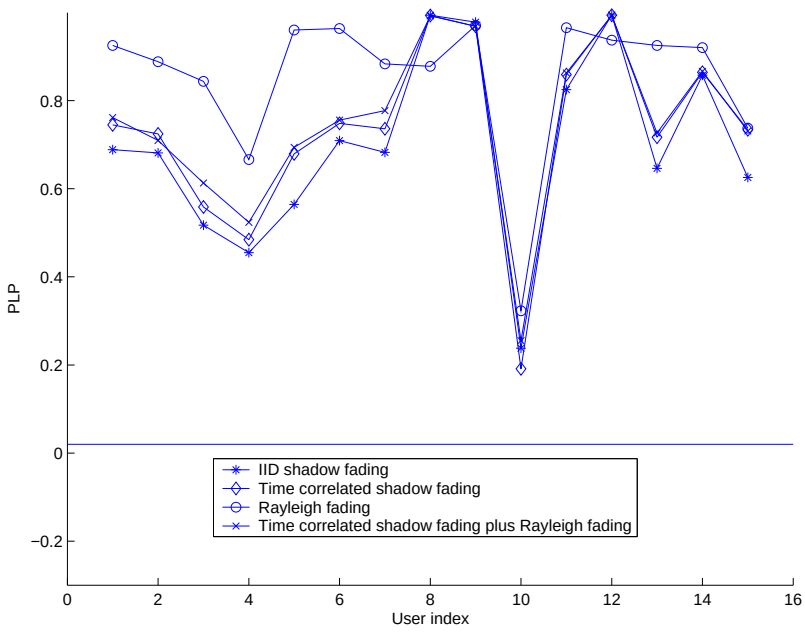


Figure 3.3: PLP for algorithm SCHEDULING_1 with single class of service.

interference is ignored in scheduling the users and calculating the beamforming weights and transmission powers. In the case of time correlated shadow fading plus Rayleigh fading channel, on the average, 4.273 users are scheduled in each timeslot. However, only 0.449 packet is successfully transmitted per timeslot. For data traffic, in order to maintain reasonable performance at higher layers the PLP at the MAC/physical layer needs to be kept fairly lower. For example, with Transmission Control Protocol (TCP), its performance degrades significantly if the PLP exceeds 5 percent [32]. Hence, the PLPs achieved by a beamforming algorithm that does not account for the inter-cell interference (Fig. 3.3) are not acceptable for data traffic. This calls for a design of a practical beamforming algorithm that can take into account the presence of random inter-cell interference and achieve (close to) target PLPs in multi-cell environments.

3.5 Average Packet Loss Probability and a Simple Heuristic Algorithm

In this section we first derive the expression for the average PLP. Then, using this expression, we demonstrate that a simple algorithm that replaces the noise term in the beamforming algorithms in Section 3.3 with the sum of average inter-cell interference and noise performs well for some special cases.

3.5.1 Average packet loss probability as a function of SINR

Most of the link curves, including the ones we used in the previous experiments (*e.g.,* [33, 36]), can be fitted using a function of the form

$$PLP(SINR) = \frac{1}{1 + e^{k(SINR_{dB} - z)}} \tag{3.10}$$

where $SINR_{dB}$ is the SINR in dB, *i.e.*, $SINR_{dB} = 10\log_{10}(SINR)$, and k and z are two fitting parameters that determine the slope and the position of a link curve, respectively.

The fitting curves are shown in Fig. 3.2 as solid curves for different modulation and/or coding schemes. One can clearly see that these fitting curves match the measured link data very closely. Link curves for systems other than the TDMA system we studied are similar in shape but with different parameters. Several example link curves are given in [36] for a CDMA system. The slope of a link curve reflects the sensitivity of PLP to SINR value and is determined by the modulation and/or coding scheme, packet length, characteristics of interference and noise, etc.

For a reasonable target PLP (less than 5-10 percent), we can approximate (3.10) as follows:

$$\begin{aligned} PLP(SINR) &\approx e^{-k(SINR_{dB}-z)} \\ &= e^{-k(10\log SINR - z)} \\ &= e^{kz}e^{-10k\log SINR} \\ &= e^{kz}e^{-10k\ln SINR/\ln 10} \\ &= e^{kz}SINR^{-10k/\ln 10} \\ &= e^{kz}SINR^{-\alpha} \end{aligned} \tag{3.11}$$

where $\alpha = 10k/\ln 10$. Since the realized SINR is a random variable (rv) due to random inter-cell interference, the (time) average packet loss probability $\overline{PLP}$ is given by[4]

$$\overline{PLP} = e^{kz}\overline{SINR^{-\alpha}} \tag{3.12}$$

where an overline is used to denote the (time) average of a rv.

[4]We replace the approximation in (3.11) with an equality.

3.5.2 A heuristic algorithm

Suppose that the inter-cell interference process is ergodic, and the time average of inter-cell interference converges to its expected value. Consider an algorithm that replaces the noise term n_j^2 in the algorithms in Section 3.3 with the sum of the average inter-cell interference and noise. We refer to this algorithm as SCHEDULING_2 algorithm. Here $\widehat{I}_j^{inter}(t)$ denotes exponentially averaged inter-cell intereference of user j.

ALGORITHM IV: SCHEDULING_2 (P_{max}; c_j, $T_j(t)$, $\mathcal{H}_j, n_j^2, \widehat{I}_j^{inter}(t), \gamma_j, \forall j \in \mathcal{N}$)

STEP 1: Initialize $\mathcal{U} = \emptyset$ and $\mathcal{N}_1 = \mathcal{N}$.

STEP 2: While $\mathcal{N}_1 \neq \emptyset$, do

- $j^* = \arg\min_{j \in \mathcal{N}_1} T_j(t)/c_j$
- $\mathcal{U} = \mathcal{U} \cup \{j^*\}$, $\mathcal{N}_1 = \mathcal{N}_1 \setminus \{j^*\}$
- If FEASIBILITY(P_{max}; $\mathcal{H}_j, n_j^2 + \widehat{I}_j^{inter}(t), \gamma_j, \forall j \in \mathcal{U}) = 0$, then $\mathcal{U} = \mathcal{U} \setminus \{j^*\}$.

STEP 3: MINIMIZE_POWER(P_{max}; $\mathcal{H}_j$, $n_j^2 + \widehat{I}_j^{inter}(t)$, γ_j, $\forall j \in \mathcal{U}$)

Assuming that the inter-cell interference is independent of the intra-cell interference, one can show that the expected value of the inverse of achieved SINR equals the inverse of $SINR_{target}$ as follows.

$$\begin{aligned}
\mathbf{E}\left[SINR^{-1}\right] &= \mathbf{E}\left[\mathbf{E}\left[\frac{I^{intra} + I^{inter} + n^2}{S}\middle| S, I^{intra}\right]\right] \\
&= \mathbf{E}\left[\frac{I^{intra} + \mathbf{E}\left[I^{inter}\right] + n^2}{S}\right] \\
&= SINR_{target}^{-1} \qquad (3.13)
\end{aligned}$$

where the last equality follows the fact that the beamforming weights and transmission powers are selected so as to achieve $SINR_{target}$ corresponding to the target PLP and the assumption that $\overline{I^{inter}} = \mathbf{E}\left[I^{inter}\right]$.

If the average inter-cell interference is used in calculating beamforming weights and transmission powers, from (3.11) and assumed ergodicity the algorithm effectively aims at a packet loss probability PLP_{target} given by

$$PLP_{target} = e^{kz} SINR_{target}^{-\alpha} = e^{kz} \mathbf{E}\left[SINR^{-1}\right]^{\alpha} = e^{kz} \overline{SINR^{-1}}^{\ \alpha} . \quad (3.14)$$

Note that if $\alpha \approx 1$, then

$$\overline{PLP} = e^{kz} \overline{SINR^{-\alpha}} \approx e^{kz} \overline{SINR^{-1}}^{\ \alpha} = PLP_{target} \quad (3.15)$$

Thus, if $\alpha \approx 1$, the realized $\overline{PLP}$ will be approximately equal to PLP_{target}, when the average inter-cell interference is added to the noise term in the beamforming algorithms. However, when α deviates considerably from one, the target PLP may not be achieved by simply adding the average inter-cell interference to the noise term.

The value of α for the link curves in Fig. 3.2 is approximately 1.1. Thus, the calculation of beamforming weights and transmission powers using the estimated average inter-cell interference performs satisfactorily. This is shown in Fig. 3.4. Note that if $\alpha > 1$, from Jensen's inequality we have

$$e^{kz} \overline{SINR^{-\alpha}} \geq e^{kz} \overline{SINR^{-1}}^{\alpha} \quad (3.16)$$

i.e., the average PLP is no smaller than the target PLP. This explains the PLPs in Fig. 3.4 being slightly larger than the target PLP of 2 percent on the average.

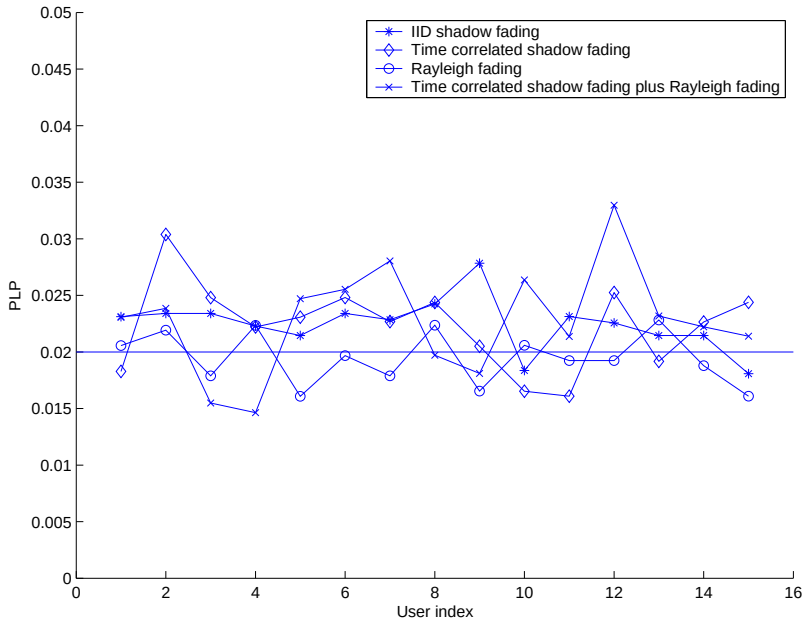
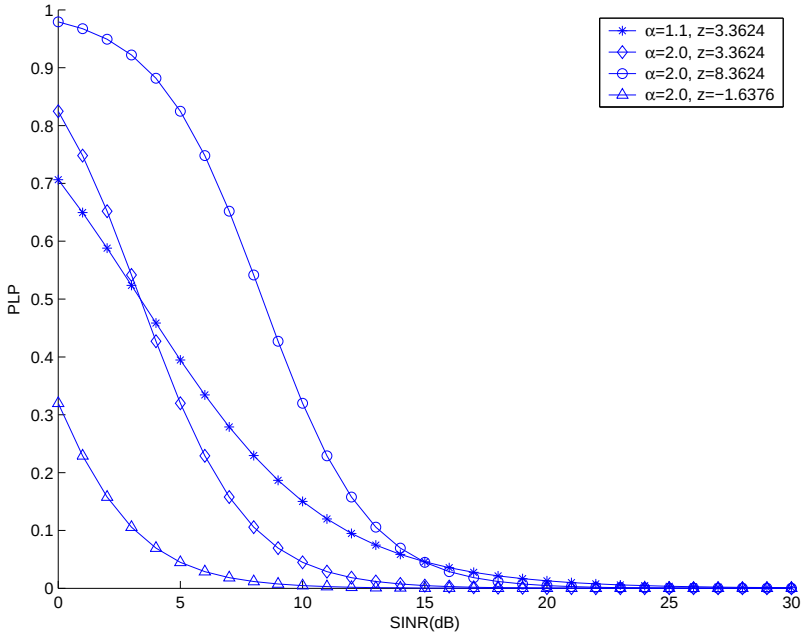


Figure 3.4: PLP for SCHEDULING_2 algorithm with single service class with link curve in Fig. 3.2.

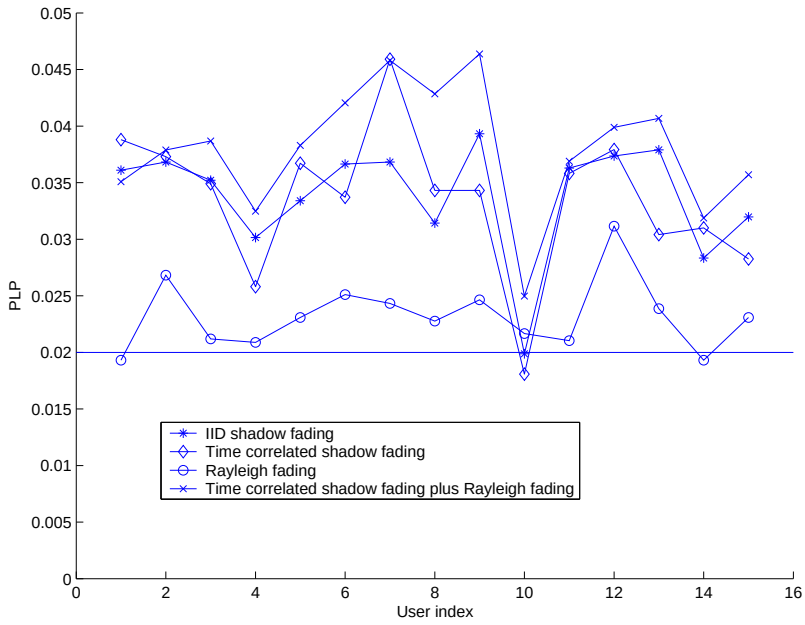
3.6 Proposed Algorithm for General Link Curves

As discussed in the previous section, SCHEDULING_2 algorithm works well only when the value of α is close to one. If the value of α deviates significantly from one (*e.g.,* link curves in [36]), its performance suffers. This is numerically demonstrated in Fig. 3.5. Fig. 3.5(a) shows the link curves with $\alpha = 2$. These link curves are obtained by increasing the slope of link curves in Fig. 3.2. Fig. 3.5(b) plots the PLPs for the middle link curve with $\alpha = 2$, and shows that the PLPs of some users are considerably higher than the target PLP of 2 percent. In this section we propose an algorithm that performs well with general link curves that can be approximated by a function in (3.10) regardless of the value of parameter α.

From (3.12) and (3.14) we observe that the discrepancy between a target PLP and a realized PLP is due to the fact that $\overline{SINR^{-1}}^{\alpha}$ and $\overline{SINR^{-\alpha}}$ differ. When



(a)



(b)

Figure 3.5: (a) Plot of link curve with $\alpha = 2$ and (b) PLP for SCHEDULING_2 algorithm with single service class.

the average inter-cell interference is used in the beamforming algorithms, when $\alpha > 1$, from the Jensen's inequality the beamforming algorithms need to aim at a smaller target PLP value, $\epsilon \cdot PLP_{target}$ for some $0 < \epsilon < 1$, than the desired target PLP in order to account for this difference. In other words, they should attempt to achieve

$$e^{kz} \cdot SINR_{target}^{-\alpha} = \epsilon \cdot PLP_{target} \ . \tag{3.17}$$

If we want to achieve a realized PLP of PLP_{target}, then from (3.12) and (3.17), it is easy to see that ϵ should be set to

$$\epsilon = \frac{\overline{SINR^{-1}}^{\ \alpha}}{\overline{SINR^{-\alpha}}} \tag{3.18}$$

We observe that ϵ depends on both the SINR distribution and the value of α. Qualitatively, a larger fluctuation in SINR leads to a smaller value of ϵ because SINR is more likely to degrade enough to cause large PLPs at times and this needs to be compensated for by setting a smaller target PLP. For a similar reason, a larger slope of the link curve, captured by α, leads to a smaller value of ϵ because the link curve is more sensitive to a fluctuation in SINR.

In order to calculate ϵ, both $\overline{SINR^{-1}}$ and $\overline{SINR^{-\alpha}}$ need to be estimated. These can be estimated by a user using an exponential averaging scheme. More precisely, when a user receives a packet, it carries out the following updates:

$$\begin{aligned} (1-\phi)\widehat{\overline{SINR^{-1}}} + \phi \cdot SINR_{new}^{-1} &\rightarrow \widehat{\overline{SINR^{-1}}} \\ (1-\phi)\widehat{\overline{SINR^{-\alpha}}} + \phi \cdot SINR_{new}^{-\alpha} &\rightarrow \widehat{\overline{SINR^{-\alpha}}} \end{aligned} \tag{3.19}$$

where $\widehat{\overline{SINR^{-1}}}$ and $\widehat{\overline{SINR^{-\alpha}}}$ are the estimates for $\overline{SINR^{-1}}$ and $\overline{SINR^{-\alpha}}$, respectively. The parameter ϕ is the exponential averaging weight and is set to 0.1 in our numerical studies. The variable $SINR_{new}$ is the SINR experienced by the user when it receives the packet.

Our new scheduling algorithm based on this observation, referred to as SCHEDULING_3, is described below.

ALGORITHM V: SCHEDULING_3(P_{max}; c_j, $T_j(t)$, $\mathcal{H}_j$, n_j^2, $\widehat{I}_j^{inter}(t)$, $\widehat{\overline{SINR_j^{-1}}}$, $\widehat{\overline{SINR_j^{-\alpha}}}$, $\forall j \in \mathcal{N}$)

STEP 1: Initialize $\mathcal{U} = \emptyset$ and $\mathcal{N}_1 = \mathcal{N}$.

STEP 2: While $\mathcal{N}_1 \neq \emptyset$, do

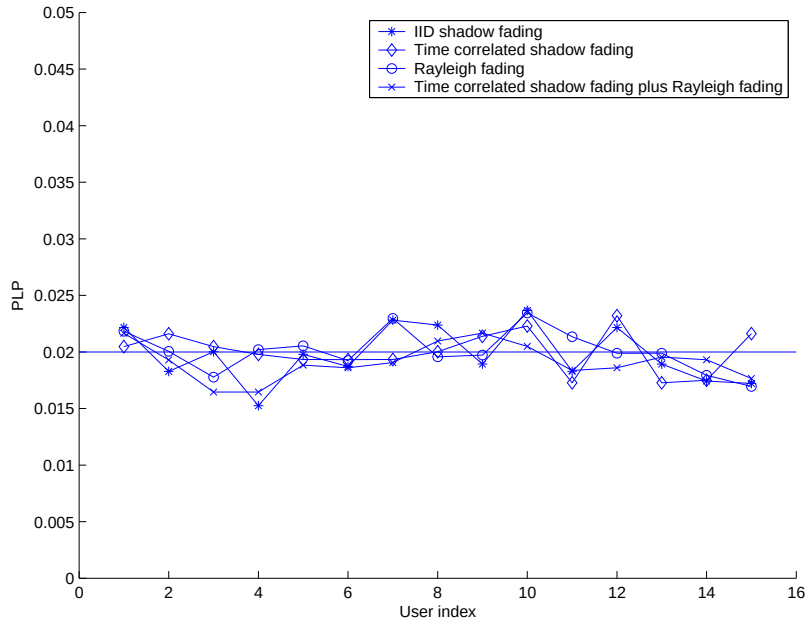
- $j^* = \arg\min_{j \in \mathcal{N}_1} T_j(t)/c_j$
- $\mathcal{U} = \mathcal{U} \cup \{j^*\}$, $\mathcal{N}_1 = \mathcal{N}_1 \setminus \{j^*\}$
- If FEASIBILITY(P_{max} ;$\mathcal{H}_j, n_j^2 + \widehat{I}_j^{inter}(t), \gamma_j, \forall j \in \mathcal{U}$) $= 0$, $\mathcal{U} = \mathcal{U} \setminus \{j^*\}$, where γ_j satisfies $e^{kz}\gamma_j^{-\alpha} = \epsilon_j \cdot PLP_{target}$.

STEP 3: MINIMIZE_POWER(P_{max}; $\mathcal{H}_j, n_j^2 + \widehat{I}_j^{inter}(t), \gamma_j, \forall j \in \mathcal{U}$)

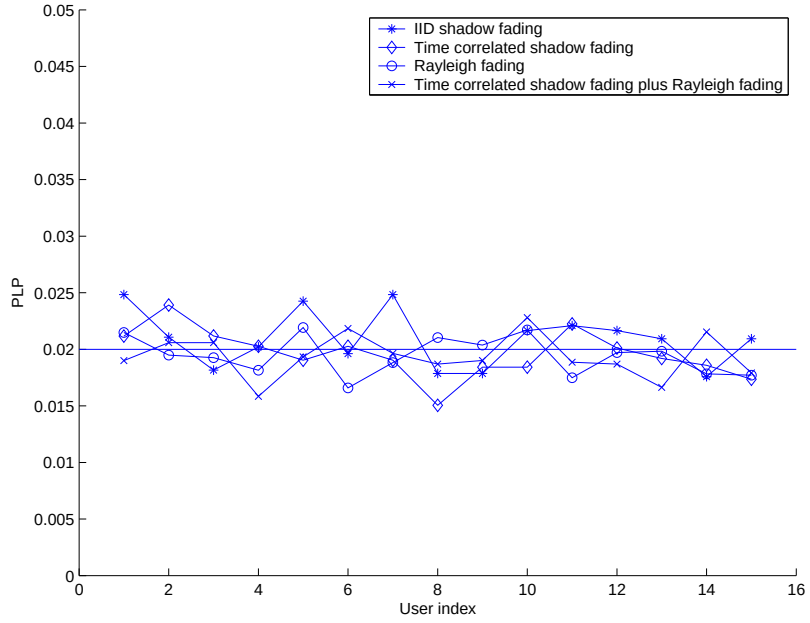
Note that in this new algorithm, the target PLP is replaced by $\epsilon_j \cdot PLP_{target}$ by selecting the target SINR satisfying $e^{kz}\gamma_j^{-\alpha} = \epsilon_j \cdot PLP_{target}$ in both STEP 2 and STEP 3.

We evaluate the performance of SCHEDULING_3 algorithm using the same setup used with the previous algorithms. The results are presented in Fig. 3.6 with both single service class and multiple service classes. For temporally correlated shadow fading plus Rayleigh fading channel model the average numbers of scheduled packets per timeslot are 2.072 and 2.045 for these two cases, respectively, and the numbers of successful transmissions are 2.032 and 2.006 correspondingly. For the multiple service classes case, there are 10 users with credit of 1, and 5 users with credit of 2. It is clear that all users achieve a PLP close to the target PLP of 2 percent under SCHEDULING_3 algorithm with all considered channel models.

Using the sum of average inter-cell interference and noise in the beamforming algorithm with reduced target PLP decreases the number of scheduled users in each



(a)



(b)

Figure 3.6: PLP for SCHEDULING_3 algorithm with a link curve of $\alpha = 2$. (a) single class, (b) multiple classes.

timeslot. The reason for this decrease in throughput is as follows. In the FEASIBILITY and MINIMIZE_POWER algorithms, the beamforming weights are computed to maximize the SINR of each user on the virtual uplink (see eq. (3.7)) [39]. The increase in noise reduces the elements in spatial covariance matrices $\widetilde{\mathcal{H}}_j$ and the relative contribution from the identity matrix I in $\mathcal{R}_j(\mathbf{q}^{(n-1)})$ becomes larger. Unlike spatial covariance matrices the identity matrix has no directional sensitivity, *i.e.,* for any beamforming vector $\mathbf{w}$ with $\mathbf{w}^H\mathbf{w} = 1$, $\mathbf{w}^H I\mathbf{w} = 1$. This is similar to having additional interference coming in from all directions (pp. 225-226 [35]). As a result, the matrix $\mathcal{R}$ is more spatially uniform with the average inter-cell interference term. This results in the beam pattern calculated by (3.7) becoming less focused. This is in contrast to the case where the beamforming vectors are calculated to minimize the interference in certain directions when noise is negligible. Because these beams are not as focused with inter-cell interference, users become less spatially separable and consequently fewer number of users can be scheduled simultaneously in each timeslot. The trade-off between the system throughput and target PLP is studied in Section 3.8.

3.7 Characterization of inter-cell interference

In this section we investigate the characteristics of inter-cell interference experienced by the users in the multi-cell environment. We study both the distribution and the temporal correlation of the inter-cell interference.

3.7.1 Log-normal distribution of inter-cell interference

We first study the distribution of the inter-cell interference experienced by the users. In Fig. 3.7 the histograms of the natural logarithm of the inter-cell interference experienced by a selected user are shown for various scheduling algorithms with i.i.d. shadow fading channel model. The x-axis is the natural logarithm of inter-cell interference and the y-axis is the normalized histogram. The solid lines represent the experimental data, and the dotted lines are the fitting curves using normal distributions. The inter-cell interference distributions are presented in Fig. 3.8 for SCHEDULING_3 with single service class for different channel models. We observe that the distribution can be well approximated by a normal distribution in all studied cases. That is, the inter-cell interference exhibits a log-normal distribution. We refer interested readers to [30] for an explanation for the emergence of a log-normal distribution. Clearly the distribution of the inter-cell interference in practice will depend on the adopted scheduling and beamforming algorithms. However, we expect the distribution to be approximately log-normal for a wide range of such algorithms and various channel models, as demonstrated by Fig. 3.7 and 3.8.

3.7.2 Temporal correlation of inter-cell interference

Leung investigated power control schemes for a TDMA system with a single antenna element in a multiple cell environment [34]. He assumes that each packet from the higher layer gets broken into multiple blocks, and a BS schedules a user continually until the transmission of a packet is completed before scheduling another user for transmission. These consecutive transmissions of the same user lead to strong temporal correlation in the inter-cell interference, and as a result the

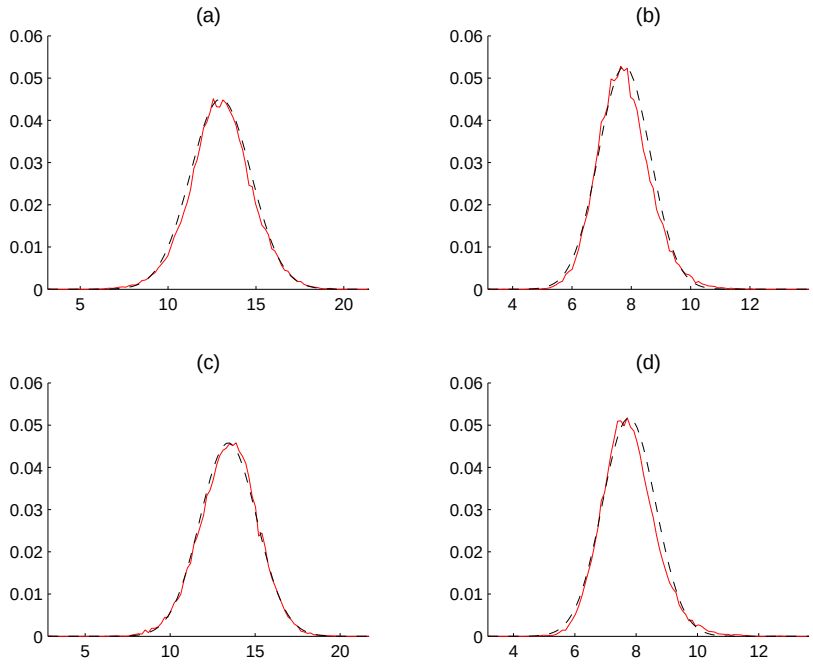


Figure 3.7: The distributions of inter-cell interference under i.i.d. shadow fading channel model. (a) SCHEDULING_1 with single class, (b) SCHEDULING_3 with single class, (c) SCHEDULING_1 with multiple classes, and (d) SCHEDULING_3 with multiple classes.

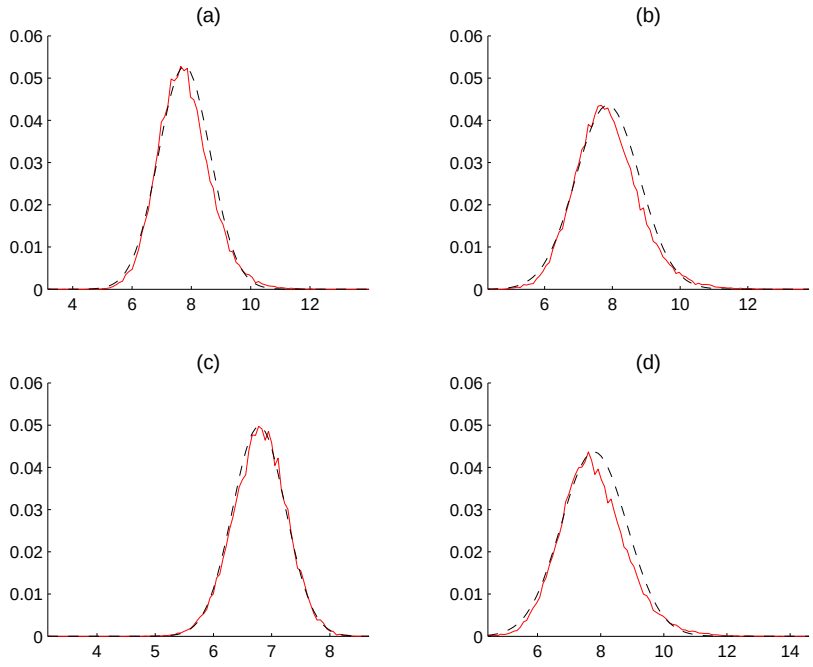


Figure 3.8: The distributions of inter-cell interference for SCHEDULING_3. (a) i.i.d. shadow fading channel model, (b) temporally correlated shadow fading channel model, (c) Rayleigh fading channel model, and (d) temporally correlated shadow fading plus Rayleigh fading channel model.

average inter-cell interference in the previous several timeslots gives a good prediction for the inter-cell interference in the next timeslot. His proposed scheme calculates the transmission power based on the predicted inter-cell interference, which is the average inter-cell interference over a sliding window of a fixed number of timeslots.

However, if the capacity of a wireless system is high enough, which is likely to be true in the future, a block or frame will be able to accommodate an entire packet, and it will take only one timeslot to complete the transmission of a packet, which is the model assumed in our study. We plot the measured autocorrelation function of the inter-cell interference of a user for various scheduling algorithms with temporally correlated shadow fading channel model in Fig. 3.9. It is clear that the inter-cell interference exhibits rather weak temporal correlation. This can be explained from the fact that each BS typically schedules a different (feasible) set of users for transmission in each timeslot, independently of other BSs. Therefore, the beamforming weights and transmission powers vary significantly from one timeslot to next. These characteristics of a packet switched cellular network lead to weak temporal correlation of inter-cell interference experienced by a user. The temporal correlation under a different channel model is in general *weaker* than this channel model.

3.8 Throughput vs. Target Packet Loss Probability Trade-off

As mentioned in Section 3.6, the number of users that can be scheduled together in a timeslot decreases when the beamforming algorithm needs to compensate for the

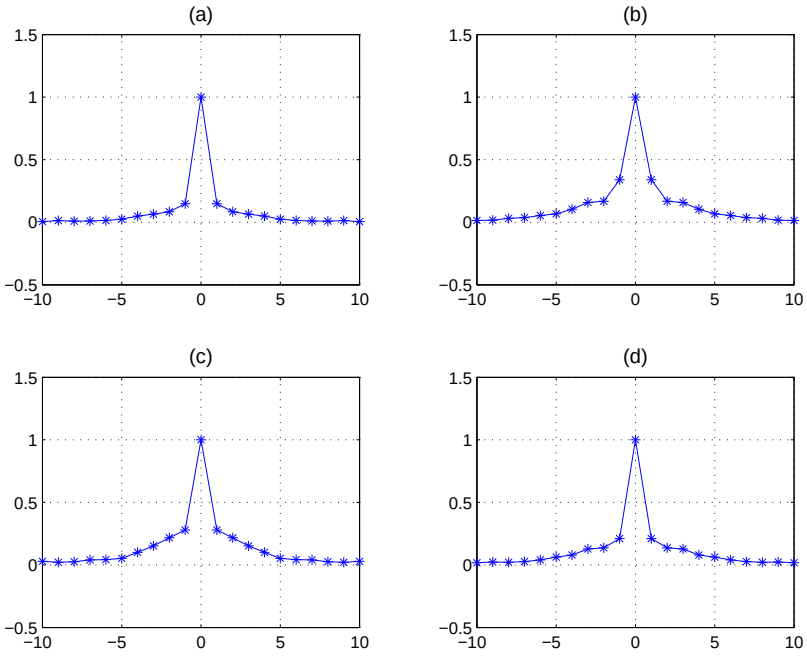


Figure 3.9: The autocorrelation functions of inter-cell interference. (a) SCHEDULING_1 with single class, (b) SCHEDULING_3 with single class, (c) SCHEDULING_1 with multiple classes, and (d) SCHEDULING_3 with multiple classes.

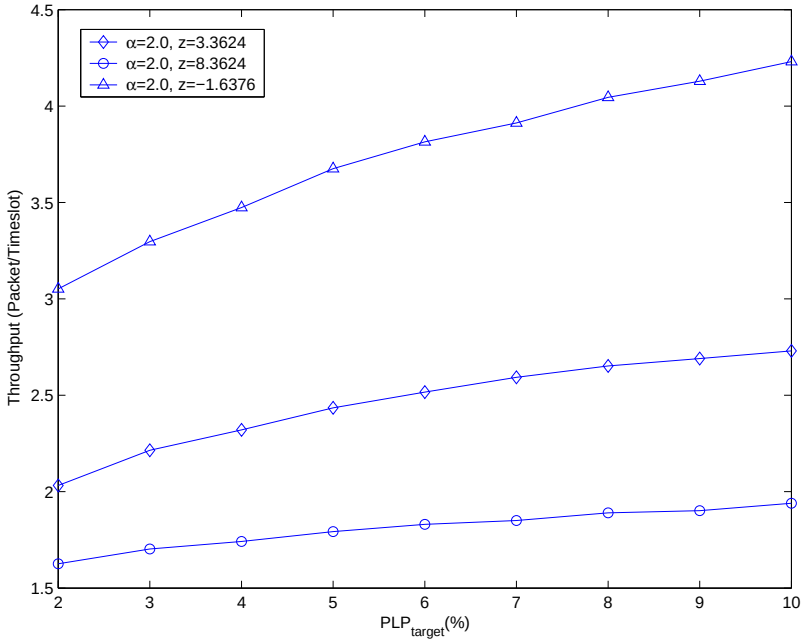


Figure 3.10: Plot of throughput vs. target PLP.

random inter-cell interference in multi-cell networks. This decrease in throughput depends on many factors, including the target PLP. In this section, using numerical examples, we study how the target PLP affects the number of scheduled users per timeslot and the resulting throughput of the system.

Fig. 3.10 shows the plot of the throughput in number of successfully transmitted packets per timeslot for different link curves with $\alpha = 2$. One can see that for reasonably small values of target PLP, the system throughput is a concave, increasing function of the target PLP. This increase in throughput in the target PLP is in fact rather significant. For example, raising the target PLP from 2 percent to 10 percent increases the system throughput by almost 20-40 percent. This suggests that in order to improve the system throughput by allowing larger target PLPs, a new transport layer protocol whose performance does not degrade even at 10 percent or higher PLP and/or large round-trip delay jitter may be desired.

Moreover, notice that the increase in the throughput with increasing target PLP tends to be larger for link curves with smaller SINR requirements.

3.9 Alternate Algorithm for Log-Normal Inter-Cell Interference

When the inter-cell interference exhibits weak temporal correlation as shown in the previous section, it is difficult to predict it accurately. However, if its distribution is known and the parameters of the distribution can be estimated, such information can be exploited to achieve the target PLPs. In this section we propose an alternate algorithm for achieving target PLP when the inter-cell interference is log-normally distributed. This is done by first estimating the parameters of the log-normal distribution and then using the estimated parameters in a beamforming algorithm to compute the PLP.

Recall from (3.10) that, for a fixed signal strength S, the achieved PLP, denoted by $PLP(S)$, can be expressed as

$$\begin{aligned} PLP(S) &= e^{kz}\mathbf{E}\left[\left(\frac{I^{intra}+I^{inter}+n_j^2}{S}\right)^{\alpha}\right] \\ &= e^{kz}\int_0^{\infty}\left(\frac{I^{intra}+n_j^2+y}{S}\right)^{\alpha} f(y)dy\ . \end{aligned}$$

We assume that the inter-cell interference I^{inter} is a log-normally distributed rv with probability density function (pdf) $f(y)$. A log-normally distributed rv I^{inter} can be written as $I^{inter} = e^X$ where X is a normally distributed rv with parameters (μ, σ^2), and the pdf of I^{inter} is given by

$$f(y) = e^{-\frac{1}{2\sigma^2}(\ln(y)-\mu)^2}/(y\sigma\sqrt{2\pi}) \tag{3.20}$$

. The j-th moment of I^{inter} can be computed from the pdf and is given by

$$\mathbf{E}\left[(I^{inter})^j\right] = e^{j\mu+\frac{1}{2}j^2\sigma^2} \ . \tag{3.21}$$

In order to characterize the inter-cell interference of a user, we need to estimate the parameters μ and σ^2. Denote the estimated values of μ_j and σ_j^2 at the start of timeslot t by $\widehat{\mu}_j(t)$ and $\widehat{\sigma}_j^2(t)$, respectively. Each user updates its estimates $\widehat{\mu}_j(t+1)$ and $\widehat{\sigma}_j^2(t+1)$ using the exponential update rule similar to (3.19) in each timeslot.

$$\widehat{\mu}_j(t) = (1-\phi)\widehat{\mu}_j(t-1) + \phi \ln(I_j^{inter}(t-1))$$
$$\widehat{\sigma}_j^2(t) = (1-\phi)\widehat{\sigma}_j^2(t-1) + \phi \cdot (\ln(I_j^{inter}(t-1)) - \widehat{\mu}_j(t))^2$$

Given the inter-cell interference distributions of the users, the objective of the beamforming algorithm is to calculate the beamforming weights and transmission powers such that the total transmission power is minimized subject to the constraint that the PLP of each user is no larger than a required target PLP. The PLP constraint can be written as

$$PLP(S) = e^{kz} \int_0^\infty \left(\frac{I^{intra} + n_j^2 + y}{S}\right)^\alpha f(y)dy \leq PLP_{target}$$

The above integration needs to be computed numerically for an arbitrary value of α. However, for an integer-valued α, we can obtain a closed form solution using (3.21). Here we assume $\alpha = 2$, and the PLP is given by

$$PLP(S) = e^{kz} \frac{(I^{intra} + n_j^2)^2 + 2(I^{intra} + n_j^2)e^{\mu+\frac{1}{2}\sigma^2} + e^{2\mu+2\sigma^2}}{S^2} \ .$$

It is clear that this constraint does not depend only on SINR, and thus changes the structure of the previous beamforming problem. As a consequence, the algorithms FEASIBILITY and MINIMIZE_POWER can no longer be used for computing beamforming weights and power control. Instead, we propose the following beamforming algorithm.

ALGORITHM VI: BEAMFORMING(P_{max};$\mathcal{H}_j, n_j^2, \widehat{\mu_j}, \widehat{\sigma}_j^2$,$\forall j \in \mathcal{U}$)

STEP 1: Set $n = 0$. Let $\mathbf{p}^{(0)} = [1, \cdots, 1]^T$.

STEP 2: While 1, do

- Set $n \leftarrow n + 1$, solve a set of U decoupled generalized eigenproblems.

$$\mathbf{w}_j^{(n)} = \arg \max_{\|\mathbf{w}_j\|=1} \frac{\mathbf{w}_j^H \mathcal{H}_j \mathbf{w}_j}{\mathbf{w}_j^H \mathcal{R}_j(\mathbf{p}^{(n-1)}) \mathbf{w}_j}, \quad \forall j \in \mathcal{U}.$$

  where $\mathcal{R}_j(\mathbf{p}^{(n-1)}) = \sum_{k \in \mathcal{U} \setminus \{j\}} p_k^{(n-1)} \mathcal{H}_k$.

- Calculate the gain $g_j^{(n)} = \mathbf{w}_j{}^{(n)\,H} \mathcal{H}_j \mathbf{w}_j{}^{(n)}$ and the intra-cell interference $I_j^{(n)} = \sum_{k \in \mathcal{U} \setminus \{j\}} p_k^{(n-1)} \left(\mathbf{w}_k^{(n)\,H} \mathcal{H}_j \mathbf{w}_k^{(n)} \right)$ for user $j \in \mathcal{U}$.

- Calculate the transmission power $p_j^{(n)}$ for user $j \in \mathcal{U}$

$$p_j^{(n)} = \left(\frac{\left(I_j^{(n)} + n_j^2\right)^2 + 2(I_j^{(n)} + n_j^2) e^{\widehat{\mu}_j + \frac{1}{2}\widehat{\sigma}_j^2} + e^{2\widehat{\mu}_j + 2\widehat{\sigma}_j^2}}{(g_j^{(n)})^2 \; PLP_{target} \; e^{-kz}} \right)^{\frac{1}{2}}$$

- If (i) $\max_{j \in \mathcal{U}} |p_j^{(n)} - p_j^{(n-1)}| \leq \epsilon$, (ii) $\sum_{j \in \mathcal{U}} p_j > \eta \cdot P_{max}$, or (iii) $n = thresh\_$, break.

STEP 3: If $\max_{j \in \mathcal{U}} |p_j^{(n)} - p_j^{(n-1)}| \leq \epsilon$ and $\sum_{j \in \mathcal{U}} p_j \leq P_{max}$, then set $flag = 1$. Otherwise, set $flag = 0$.

STEP 4: Output $flag$, $\mathbf{p}^{(n)}$, and $\mathbf{W}^{(n)}$.

Unfortunately, the convergence of this algorithm is not guaranteed and oscillation could occur. This is expected since we do not explicitly use the estimated probability distribution of inter-cell interference when computing the beamforming weights, and such information is applied only to the power control. Furthermore, the set of users being scheduled by the algorithm may not be feasible, which will prevent the algorithm from converging. For these reasons, the algorithm (STEP 2) is terminated when the number of iterations n reaches a preset threshold $thresh\_$ or the total transmission power exceeds η times of the maximum power constraint. However, we will show that the computed beamforming weights and transmission powers achieve PLP_{target} when they do converge.

The scheduling algorithm using BEAMFORMING is outlined below.

ALGORITHM VII: SCHEDULING_4(P_{max};c_j, $T_j(t), \mathcal{H}_j, n_j^2, \widehat{\mu}_j, \widehat{\sigma}_j^2, \forall j \in \mathcal{N}$)

STEP 1: Initialize $\mathcal{U} = \emptyset$, $\mathcal{N}_1 = \mathcal{N}$.

STEP 2: While $\mathcal{N}_1 \neq \emptyset$, do

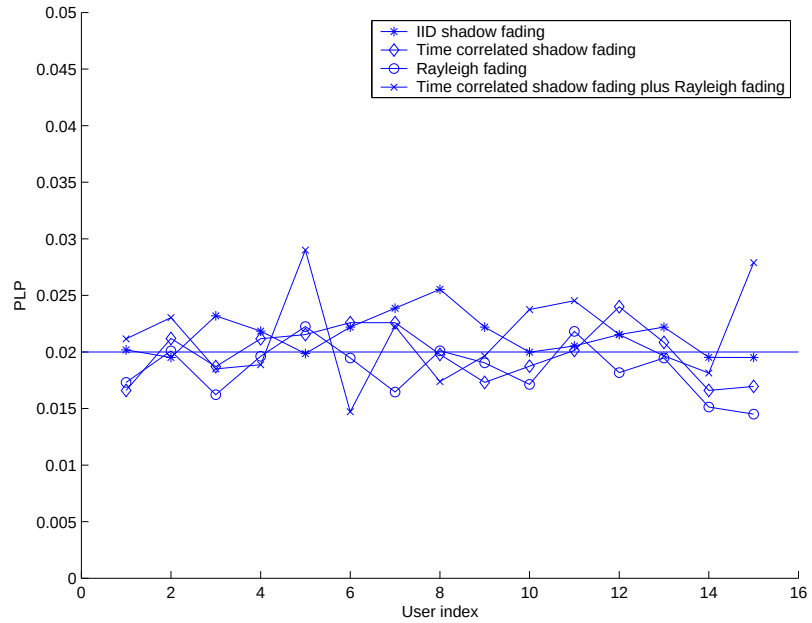
- $j^* = \arg\min_{j \in \mathcal{N}_1} T_j(t)/c_j$
- $\mathcal{U} = \mathcal{U} \cup \{j^*\}$, $\mathcal{N}_1 = \mathcal{N}_1 \setminus \{j^*\}$.
- BEAMFORMING(P_{max};$\mathcal{H}_j, n_j^2, \widehat{\mu}_j, \widehat{\sigma}_j^2, \forall j \in \mathcal{U}$).
- If $flag = 0$, $\mathcal{U} = \mathcal{U} \setminus \{j^*\}$.

STEP 3: BEAMFORMING(P_{max};$\mathcal{H}_j, n_j^2, \widehat{\mu}_j, \widehat{\sigma}_j^2, \forall j \in \mathcal{U}$).

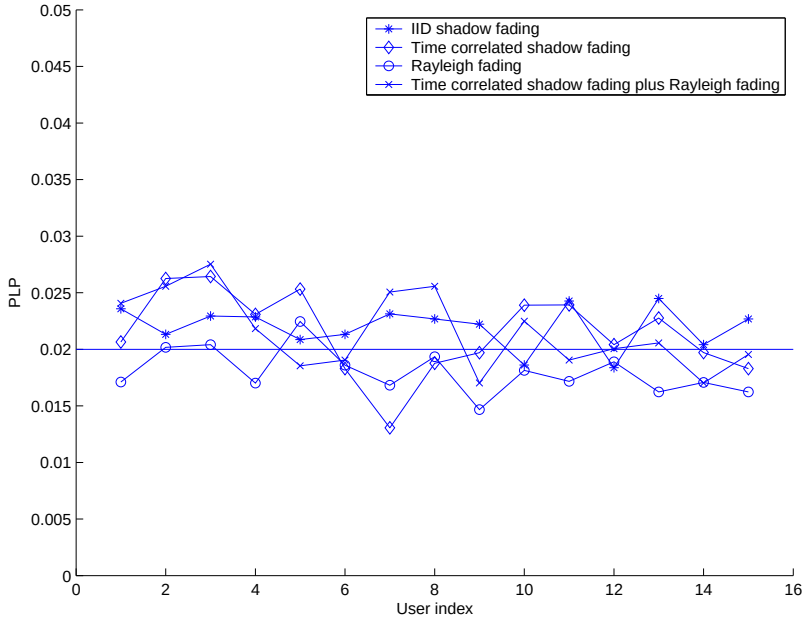
The performance of algorithm SCHEDULING_4 is illustrated in Fig. 3.11 with both single service class and multiple service classes. The figure shows that the achieved PLPs are close to PLP_{target} for all channel models. However, the achieved throughput of SCHEDULING_4 is considerably smaller than that of SCHEDULING_3. For the temporally correlated shadow fading plus Rayleigh fading channel model with single service class, 1.3988 packets are scheduled and 1.3687 packets are successfully received in each timeslot on the average.

3.9.1 Discussion

In this subsection we briefly compare the proposed algorithms SCHEDULING_3 and SCHEDULING_4. Algorithm SCHEDULING_3 utilizes the same optimal beamforming algorithms proposed in [39] by dynamically adjusting the target PLP and hence target SINR provided to the algorithms. This is done based on the derived expression for PLP in (3.12) for link curves of the form (3.10). The target PLP fed to the algorithms is computed from the desired target PLP and channel conditions summarized by the estimates of $\overline{SINR^{-1}}$ and $\overline{SINR^{-\alpha}}$. This allows us to keep the



(a)



(b)

Figure 3.11: PLP for algorithm SCHEDULING_4 with (a) single service class and (b) multiple service classes.

optimality of the previously proposed algorithms while satisfying the target PLPs of the users.

In SCHEDULING_4 on the other hand, the power control of the scheduled users is carried out *individually* while assuming that the transmission powers of other scheduled users remain the same. One can view this as a non-cooperative game among the scheduled users where each user iteratively attempts to minimize its own power based on the transmission powers of the other users subject to its own PLP constraint. An equilibrium (or a solution concept) of such a non-cooperative game is called a Nash equilibrium, and it is well known that a Nash equilibrium is in general inefficient [42]. This explains the decrease in system throughput compared to that of SCHEDULING_3.

3.10 Discussion

We proposed two beamforming algorithms for handling inter-cell interference in multi-cell networks. The first algorithm utilizes two input parameters based on the channel conditions and the adopted link curve. The second algorithm exploits the observations that the inter-cell interference exhibits log-normal distribution with weak temporal correlation, and allows sequential computation of the transmission powers of scheduled users. Both algorithms are shown to achieve target loss probabilities with a large family of link curves and scheduling algorithms under various wireless channel models.

Chapter 4

Optimal Transmission Scheduling with Base Station Antenna Array in Cellular Networks

4.1 Introduction

All of the previous studies on scheduling algorithms with beamforming model the users as infinite sources with packets waiting in the queues at all times. A major drawback of this assumption is that it focuses only on instant total throughput and does not consider the upper layer QoS requirements of individual users. Thus, the assignment of users on a channel only reflects the physical layer feasibility, but not the current buffer occupancy or traffic demand of each user. We will demonstrate that this separation of physical layer algorithms and upper layer QoS requirements leads to a degradation in overall system performance and user experience. This suggests that the joint design of a MAC layer scheduling policy and physical layer beamforming algorithms is beneficial.

In this thesis we study the problem of designing a scheduling algorithm with BSs that are equipped with an antenna array. We first consider the case where a central controller handles multiple BSs serving a set of users. In this case packets arrive at the central controller for transmission to mobile users. In addition to spatial separability of the users sharing the same channel, the scheduling policies consider the current buffer occupancy and thus reflect the QoS requirement of each user in terms of throughput. We model the system as a queueing system with multiple parallel servers, and the physical layer constraints are imposed on the selection of users that can be served in each timeslot. Instead of a policy that maximizes instant throughput, we look for an optimal scheduling policy that stabilizes the system if it is stable under some policy.

Similar queueing systems have been used to model other scenarios in [56–58], and were first proposed in [56] for a multi-hop radio network where the SINR requirement demands that two links can be active simultaneously only if they are separated by certain minimum required distance. The throughput region is defined as the set of arrival rate vectors for which the system is stable. An optimal scheduling policy that stabilizes the system whenever it is stable under some policy is identified. However, the complexity of the optimal scheduling policy increases exponentially with the number of users, and no practical sub-optimal scheduling policy is proposed in [56–58]. In this thesis, we follow a similar approach as in [58], and propose two scheduling policies with significantly lower complexity that achieve sub-optimal performance for our problem. In fact, the first proposed scheduling algorithm has linear complexity.

This chapter is organized as follows. In Section 4.2 we describe the problem of designing an efficient downlink scheduling algorithm with base station antenna

arrays, and derive an optimal scheduling policy based on feasible rate matrices. In Section 4.3 we describe our proposed heuristic algorithms that approximate the optimal scheduling policy with lower complexity. Simulation results of the proposed algorithms in a single cell network are given in Section 4.4. Section 4.5 provides simulation results of the proposed algorithms in a multi-cell environment. We conclude in Section 4.6.

4.2 Optimal Downlink Scheduling

In this section we consider the case where a centralized control agent carries out scheduling and beamforming for multiple BSs serving a set of users. We define an achievable rate vector and a throughput region, and present an optimal scheduling algorithm that can achieve any interior point in the throughput region.

4.2.1 System model

We consider a wireless network that consists of several BSs. Each BS is equipped with an antenna array so that several users can be served simultaneously. These BSs are coordinated by a single central controller. Mobile users in the network are able to receive data packets from any of these BSs. However, at any given time, a mobile user can receive data packet(s) from at most one BS. The central controller maintains a separate queue for incoming data packets destined for each mobile user. We assume a time slotted system where the transmission time of a packet equals the duration of a timeslot when the lowest transmission rate is selected. In each timeslot, the central controller collects the information regarding the wireless channel conditions of each user to different BSs. Based on this information and

the number of backlogged packets of each user, the central controller makes a scheduling decision for the timeslot. The scheduling decision made by the central controller includes assignment of BSs to the users and the transmission rate of each user, and the calculation of the beamforming weights for the selected transmission rates.

The block diagram of the system under study is depicted in Fig. 4.1. User packets enter the scheduling module at the central controller, which determines the assignments of BSs and transmission rates. Beamforming and power adaptation are subsequently carried out for scheduled users. Scheduling and beamforming are interdependent operations, and they also depend on network state (*i.e.,* queue sizes) and channel state information, which are assumed to be available at the central controller.

4.2.2 Problem statement

The network consists of I BSs shared by J mobile users. We denote the set of BSs by $\mathcal{I}$ and the set of users by $\mathcal{J} = \{1, \ldots, J\}$. There is a central controller that coordinates the operation of the I BSs. Each BS is equipped with an M-element antenna array.

Several transmission rates are available at the BSs based on the channel conditions. The set of available transmission rates is denoted by $\mathcal{V}$. We assume that each transmission rate is a positive integer number. If rate $v \in \mathcal{V}$ is chosen, up to v packets can be transmitted in one timeslot, depending on the number of packets waiting for transmission. We denote $|\mathcal{V}| = V$.

Packets arrive at the central controller for transmission, which maintains a separate queue for each user. Let $a_j(t), j = 1, 2, \cdots, J$ and $t = 0, 1, \cdots$, denote

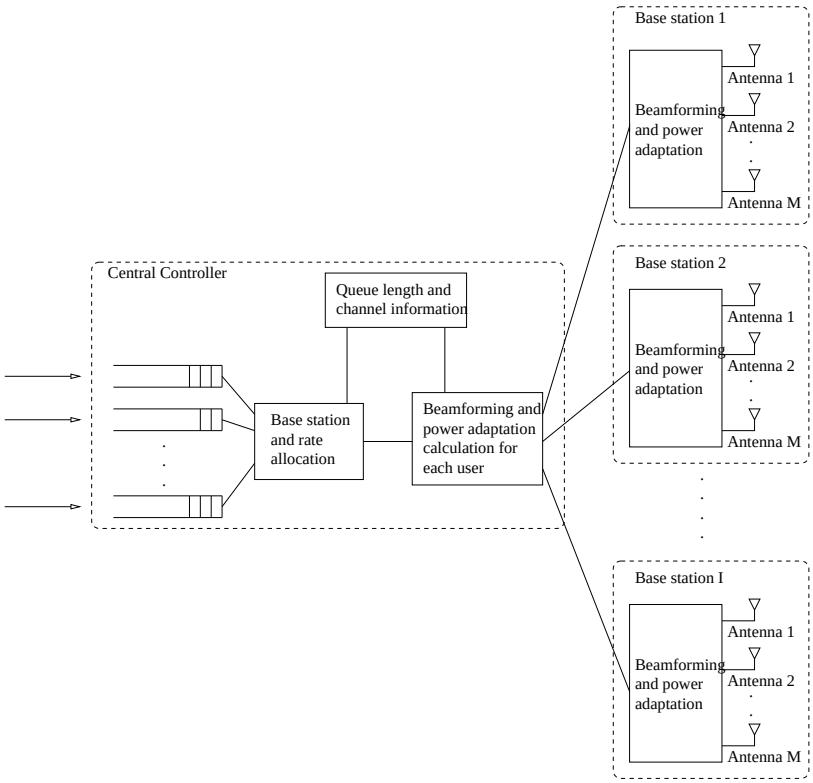


Figure 4.1: The multiple cellular communication system.

the number of packets that arrive at queue j in timeslot t. We assume that $a_j(t), t = 0, 1, 2, \ldots,$ are independent and identically distributed (i.i.d.) random variables (rvs) with a finite second moment, *i.e.,* $\mathbf{E}\left[a_j(t)^2\right] < \infty$. The average arrival rate of user j is denoted by $A_j = \mathbf{E}\left[a_j(t)\right]$. We call $\mathbf{A} = (A_1, A_2, \cdots, A_J)^T$ an arrival vector.

We assume that the central controller has perfect channel information of each user to the BSs. In each timeslot, the central controller (i) assigns the BSs to the users, (ii) computes transmission rates of scheduled users, and (iii) calculates the beamforming weights of the scheduled users. A scheduling decision by the central controller can be expressed as an $I \times J$ matrix $\mathbf{R} = [r_{ij}]$ where the element $r_{ij} \in \mathcal{V} \cup \{0\}, i = 1, \cdots, I,$ and $j = 1, \cdots, J,$ is the transmission rate of BS i to user j. However, a rate matrix $\mathbf{R}$ can be selected for transmission by the central controller only if it satisfies certain physical layer constraint described below.

The PLP requirement at higher layers demands that the SINR at each receiver be above some threshold value. A rate matrix is feasible if and only if each user receives packets from at most one BS and SINR requirement is satisfied for each user. Note that the feasibility of a rate matrix depends on the target PLP that determines the SINR requirement for each user.

We model the channel process for all users as a Markov chain (MC) with a stationary distribution π. Each channel state represents the set of all feasible rate matrices. In other words, a state of the MC is the set of all rate matrices that are feasible for transmission given the channel condition. The state space of the MC is denoted by $\mathcal{S}$. The problem we are interested in is to find an optimal scheduling policy that selects a feasible rate matrix in each timeslot given the queue sizes and channel state, so that the system achieves maximum throughput,

while maintaining a stable system whenever possible under some policy. In this thesis we only consider stationary policies, *i.e.,* the scheduling decisions do not depend on timeslot t, but only on the queue sizes $\mathbf{X}(t)$ and channel state $\mathbf{S}(t)$. A stationary scheduling policy can be viewed as a mapping that assigns to each pair $(\mathbf{X}, \mathbf{S}) \in \mathcal{Z}_+^J \times \mathcal{S}$ of queue sizes and channel state a distribution on the set of feasible rate matrices for the given state $\mathbf{S}$, where $\mathcal{Z}_+ := \{0, 1, \ldots\}$.

4.2.3 Throughput region

In this subsection we first define a stable arrival vector and then characterize the throughput region.

Definition 4.2.1 *An arrival vector* $\mathbf{A}$ *is said to be stable if there exists a scheduling policy such that*

$$\lim_{c\to\infty} \limsup_{t\to\infty} \frac{1}{t} \sum_{\tau=1}^{t} \mathbf{1}_{[x_j(\tau)>c]} = 0 \ , \qquad \textit{for all } j = 1, 2, \cdots, J \tag{4.1}$$

where $x_j(t)$ *is the number of backlogged packets in queue* j *at the beginning of timeslot* t*. If a scheduling algorithm satisfies (4.1), then we say that* $\mathbf{A}$ *is stable under the scheduling policy. The throughput region, denoted by* $\mathcal{A}$*, is defined to be the closure of the set of stable arrival vectors.*

The following proposition characterizes the throughput region $\mathcal{A}$.

Proposition 1 *A necessary and sufficient condition for an arrival vector* $\mathbf{A}$ *to belong to* $\mathcal{A}$*, is that there exists a scheduling policy that achieves*

$$\mathbf{A} \leq \mathbf{D} := \sum_{\mathbf{S}\in\mathcal{S}} \pi_{\mathbf{S}} \sum_{\mathbf{R}\in\mathbf{S}} c_{\mathbf{SR}} \mathbf{R}^T \mathbf{1}_{I\times 1}(1 - PLP) \tag{4.2}$$

where $c_{\mathbf{SR}}, \mathbf{S} \in \mathcal{S}, \mathbf{R} \in \mathbf{S}$*, are nonnegative numbers such that* $\sum_{\mathbf{R}\in\mathbf{S}} c_{\mathbf{SR}} = 1$ *for all* $\mathbf{S} \in \mathcal{S}$*.*

Proof A proof is provided in Appendix A.1 ■

4.2.4 Optimal scheduling policy

In this subsection we are interested in finding an optimal scheduling policy that satisfies (4.1) for each $\mathbf{A} \in int(\mathcal{A})$. In particular, we consider the following scheduling policy: Given backlog vector $\mathbf{X}(t) = (x_1(t), \ldots, x_J(t))^T$ and channel state $\mathbf{S}(t)$, the rate matrix selected by the scheduling algorithm is given by

$$\mathbf{R}(t) = \arg \max_{\mathbf{R} \in \mathbf{S}(t)} \mathbf{X}(t)^T (\mathbf{R}^{\mathbf{T}} \, \mathbf{1}_{I \times 1}) \ . \tag{4.3}$$

Ties are assumed to be broken arbitrarily.

The backlog process $\mathbf{X}(t)$ is a J-dimensional Markov process with countably infinite state space given that the scheduling policy is stationary.

Define the following Lyapunov function

$$L(\mathbf{X}(t)) = \sum_{j=1}^{J} (x_j(t))^2 \ .$$

In order to prove the existence of a stationary distribution of $\mathbf{X}(t)$ and hence the stability of the system, we use the following theorem.

Theorem 4.2.2 *([59, 60]) For a given Lyapunov function $L(\mathbf{X}(t))$, if there exists a compact region Σ of $\mathcal{R}^J$ and a constant $\alpha > 0$ such that*

1. $\mathbf{E}\left[L(\mathbf{X}(t+1))|\mathbf{X}(t)\right] < \infty$ *for all* $\mathbf{X}(t) \in \mathcal{R}^J$

2. $\mathbf{E}\left[L(\mathbf{X}(t+1)) - L(\mathbf{X}(t))|\mathbf{X}(t)\right] \leq -\alpha$ *whenever* $\mathbf{X}(t) \in \Sigma^C := \mathcal{R}^J \setminus \Sigma$,

then a steady state distribution of the vector $\mathbf{X}(t)$ exists and, thus, the system is stable.

Essentially the theorem states that it suffices to show that there is a negative drift in the Lyapunov function when the backlogs are sufficiently large.

Now we state a proposition that establishes the optimality of the scheduling policy given by (4.3).

Proposition 2 *Suppose that* $\mathbf{A} \in int(\mathcal{A})$*, where* $int(\mathcal{A})$ *is the interior of the throughput region* $\mathcal{A}$*. Then, the system is stable under the scheduling policy given by (4.3).*

Proof A proof is given in Appendix A.2 ■

In this section we have considered scenarios where a centralized controller carries out the scheduling and beamforming for multiple BSs serving a fixed set of users, and derived an optimal scheduling policy. The derived optimal scheduling policy, however, does not yield a practical implementation as it requires searching through *all* feasible rate matrices given the current channel state and identifying the one that maximizes the inner product given in (4.3).

A natural question that arises is how one can design more practical resource allocation algorithms based on the optimal scheduling policy. In the following sections, we assume that each user is associated with the closest BS, *i.e.,* static BS assignment, and investigate the issue of designing practical scheduling algorithms with an antenna array at the BS(s). We will first consider a simple case of single cell networks in Sections 4.4, and then discuss multiple-cell environments in Section 4.5. In the case of a multiple-cell network, a receiving user experiences inter-cell interference from co-channel cell BSs that share the same frequency spectrum. Therefore, a BS needs to compensate for this random inter-cell interference experienced by a receiver when computing beamforming weights and transmission power of a scheduled user. This issue will be discussed in Section 4.5.

4.3 Heuristic Algorithms

If the user channels are time invariant, *i.e.,* constant, the optimal policy described in Section 4.2.4 can be adopted. In other words, each BS may be able to exhaustively search through all possible transmission rate vectors *off-line*, and select the solution to (4.3) in each timeslot. However, when channels vary with time, this exhaustive search becomes too computationally expensive and impractical, if not impossible. This is because the number of possible rate vectors is given by

$$c_0 = \sum_{j=1}^{N} \binom{N}{j} V^j = (1+V)^N - 1 \ , \tag{4.4}$$

which increases exponentially with the number of co-cell users N. Hence, we turn to the problem of designing heuristic algorithms that will perform well and demand much lower computational requirement.

4.3.1 First Heuristic Algorithm

Although the optimal policy in (4.3) does not lead to a practical algorithm, it suggests that a good policy should attempt to give higher priority to users with larger queue sizes. This observation is intuitive in the sense that in order to maintain the stability of the system, there should be a balance between (i) maximizing the system throughput and (ii) keeping the queue sizes from growing without a bound. Therefore, the optimal policy considers the inner product of two vectors, namely $\mathbf{X}(t)$ and $\mathbf{R}^T\mathbf{1}_{I\times 1}$, where the first term is the queue size and the latter represents the transmission rate of each user.

A heuristic algorithm that attempts to mimic the behavior of the optimal policy can order the users based on either (i) the transmission rates they can achieve given the current channel state or (ii) queue sizes. The first approach is problematic as

the achievable transmission rates of the users depend on the set of scheduled users and it requires searching through all possible feasible rate vectors. Hence, in our first heuristic algorithm we attempt to order the users according to their queue sizes and give higher priority to users with a larger queue. More specifically, the algorithm starts with the user with the longest queue, and tries to schedule the users *sequentially* in the decreasing order of their queue lengths. Each new user is allocated the highest possible rate such that the SINR requirement is satisfied with the new rate vector. However, when we insert users into the channel sequentially according to their queue lengths, it is possible that a user already scheduled for transmission prevents a number of other users from accessing the channel because the necessary spatial separability cannot be provided. Therefore, in order to improve the performance of the system further and maintain linear complexity, we will consider several candidate rate vectors and select the one that maximizes (4.3). More specifically, we will consider P rate vectors out of all possible rate vectors. Clearly, this subset of candidate rate vectors should consist of the rate vectors that are more likely to maximize (4.3).

We explain how we generate this subset of candidate rate vectors to be considered. Suppose that we form an ordered list of users by decreasing queue size. In order to generate the p-th candidate rate vector, $p = 1, \ldots, P$, of the subset, we first move the p-th user in the list to the head of the list. Then, starting from the head of the list, go down the list sequentially and insert one user at a time using the largest rate that is allowed while maintaining the rates and required SINR values of the previously scheduled users. Note that in some cases, a user may need to be skipped because the user may not be compatible with other users already scheduled. Once the P candidate rate vectors are generated, out of these rate

vectors we select the one that maximizes (4.3). The pseudo-code of this algorithm is provided below.

ALGORITHM I: HEURISTIC_1($x_j(t), \forall j \in \mathcal{N}$)

STEP 1: Initialize $\mathcal{R} = \emptyset$.

STEP 2: For $p = 1$ to P, do

- Form a list $\mathcal{K}$ of users as follows: Insert the user with the p-th largest queue size at the head of the list, and insert the remaining users by decreasing queue size.
- Initialize the rate vector $\mathbf{r} = \underline{0}$ and the set of scheduled users $\mathcal{U} = \emptyset$.
- While $|\mathcal{K}| \neq 0$, do
  - $flag = 0$.
  - Schedule the user at the head of the list, denoted by j^*, $\mathcal{K} = \mathcal{K} \setminus \{j^*\}$, $\mathcal{U} = \mathcal{U} \cup \{j^*\}$ and $\mathcal{V}_1 = \mathcal{V}$
  - While $\mathcal{V}_1 \neq \emptyset$ do
    - $v_m = \max\{\mathcal{V}_1\}$, $r_{j^*} = v_m$, $\mathcal{V}_1 = \mathcal{V}_1 \setminus \{v_m\}$.
    - If FEASIBILITY(P_{max}; $\mathcal{H}_j, n_j^2, \gamma_j$, $\forall j \in \mathcal{U}$) $= 1$, $flag = 1$ where γ_j is the target SINR for $\forall j \in \mathcal{U}$, break;
  - If $flag = 0$, $r_{j^*} = 0$ and $\mathcal{U} = \mathcal{U} \setminus \{j^*\}$.
- $\mathcal{R} = \mathcal{R} \cup \mathbf{r}$.

STEP 3: Among the rate vectors in $\mathcal{R}$, select $\mathbf{r_o}$

$$\mathbf{r_o} = \arg\max_{\mathbf{r} \in \mathcal{R}} \sum_{j=1}^{J} r_j x_j(t) \ . \tag{4.5}$$

STEP 4: MINIMIZE_POWER($P_{max}; \mathcal{H}_j, n_j^2, \gamma_j, \forall j \in \mathcal{U}$)

The complexity of HEURISTIC_1 scheduling algorithm is

$$c_1 = O(PNV) \ ,$$

and hence it increases *linearly* with both the number of candidate rate vectors P and the number of users N.

4.3.2 Second Heuristic Algorithm

The optimal scheduling policy (4.3) is not suitable for implementation because its complexity increases exponentially with the number of users as mentioned earlier (see eq. (4.4)). However, in order to prevent the complexity from increasing with the number of users, in each timeslot an algorithm can first choose a subset of a *fixed* number of users to be considered for scheduling in the timeslot and then carry out the exhaustive search in (4.3) on the selected users. In other words, the algorithm will mimic the behavior of the optimal policy on a smaller set of users that are selected and do not consider the remaining users for scheduling in the timeslot. To be consistent with the observation that a higher priority should be given to users with a larger queue size, we select K users with largest queue lengths. Then, an exhaustive search is conducted for all the feasible rate vectors on this subset of users, and the rate vector that maximizes (4.3) is selected. A pseudo-code of the proposed algorithm is provided below.

ALGORITHM II: HEURISTIC_2($x_j(t)$, $\forall j \in \mathcal{N}$)

STEP 1: Initialize $\mathcal{R} = \emptyset$.

STEP 2: Select the K users with largest queue lengths. We denote this set of users as $\mathcal{K}$.

STEP 3: Let $\mathcal{S} = \{\mathbf{r} : r_j \in \mathcal{V} \cup \{0\}$ if $j \in \mathcal{K}, r_j = 0$ if $j \notin \mathcal{K}\}$.

STEP 4: For each $\mathbf{r} \in \mathcal{S}$ do

- $\mathcal{U} = \{j : j \in \mathcal{K}, r_j > 0\}$.
- If FEASIBILITY(P_{max}; $\mathcal{H}_j, n_j^2, \gamma_j$, $\forall j \in \mathcal{U}$) $= 1$, $\mathcal{R} = \mathcal{R} \cup \mathbf{r}$.

STEP 5: Among the rate vectors in $\mathcal{R}$, select $\mathbf{r_o}$

$$\mathbf{r_o} = \arg\max_{\mathbf{r} \in \mathcal{R}} \sum_{j=1}^{J} r_j x_j(t) \ . \tag{4.6}$$

STEP 6: MINIMIZE_POWER($P_{max}; \mathcal{H}_j, n_j^2, \gamma_j, \forall j \in \mathcal{U}$)

The complexity of HEURISTIC_2 is

$$c_2 = O(\sum_{k=1}^{K} \binom{K}{k} V^k) \ ,$$

and thus one can see that the complexity of the algorithm increases exponentially with the size of the subset K.

4.4 Performance evaluation

In this section, we evaluate the performance of the proposed heuristic scheduling algorithms in a single cell network, using simulations. We will first describe the simulation setup, and then present the numerical results.

4.4.1 Simulation setup

Wireless channel model

We adopt the time correlated shading fading plus Rayleigh fading for the wireless channel mode. The parameters used in the simulation to model the wireless channel are listed in Table 4.1, where λ is the wavelength of the carrier electro-magnetic wave.

		P_{max}	10^{15}	PLP_{target}	2%
N	10	M	4	r	λ
R	1000	R'	100	L	6
κ	3.5	ρ	0.1	$\mathbf{E}\left[u_\ell(t)\right]$	0
$\mathbf{E}\left[r_\ell(t)^2\right] - \mathbf{E}\left[r_\ell(t)\right]^2$	1.07	$\mathbf{E}\left[f_\ell(t)^2\right]$	1	$\mathbf{E}\left[f_\ell(t)^2\right] - \mathbf{E}\left[f_\ell(t)\right]^2$	1

Table 4.1: Parameters used in performance evaluation of scheduling algorithms

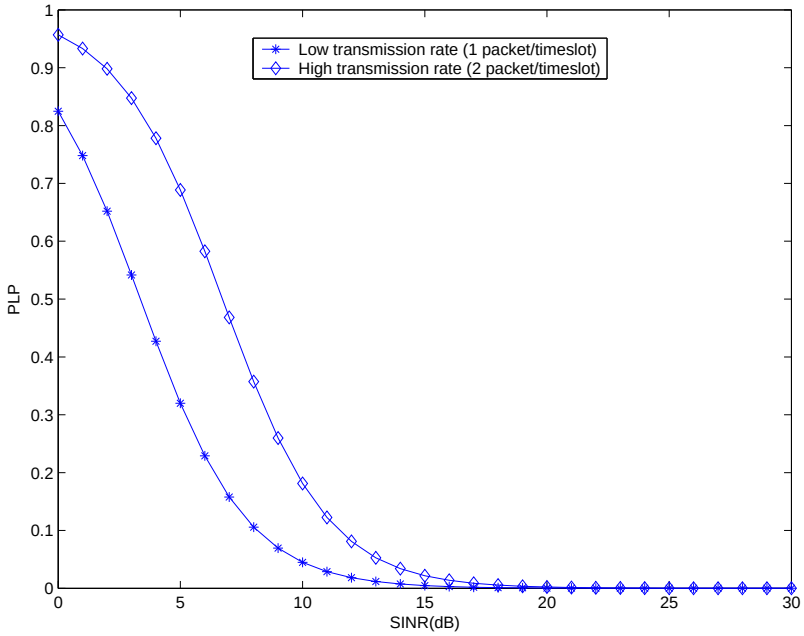


Figure 4.2: The linkcurves for low and high transmission rates.

Transmission rate

We assume that the BS can transmit packets to a user using either a low transmission rate or a high transmission rate. When a low (resp. high) transmission rate is used, one (resp. two) packet is transmitted in a timeslot. The adopted link curves for both the low and high transmission rates are shown in Fig. 4.2.

Traffic load

In our experiment, we generate an $N \times 1$ random vector $\widehat{a}$. The arrival rate vector $\mathbf{A} = s \cdot \widehat{a}$ where s is a parameter used to scale the arrival rate. Traffic load is defined as $||\mathbf{A}||_1$. Since it is difficult to characterize the throughput region using simulations, instead we observe the average delay experienced by the packets with increasing traffic load. Typically when the system loses stability there is a sharp increase in the average delay at some point (called the *knee*) due to the instability.

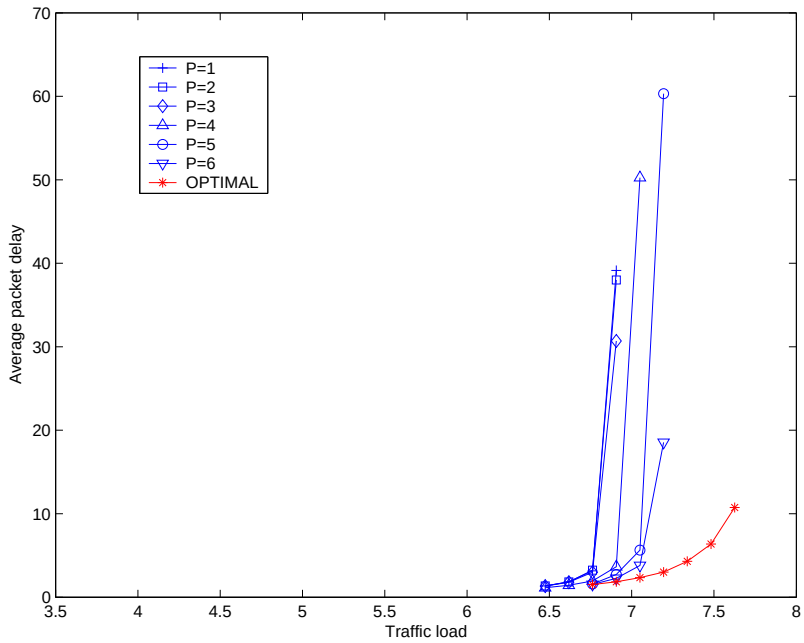


Figure 4.3: Average packet delay vs. traffic load for HEURISTIC_1 algorithm for a single cell.

4.4.2 Numerical results

In Fig. 4.3 we show the performance of HEURISTIC_1 algorithm for different values of parameter P. We can observe that the average delay experienced by packets is similar for $P = 1, 2$ and 3. This is because users are more spatially separable in a single cell network when there is no inter-cell interference, and most of the times all three users with the largest queues can be scheduled with the high transmission rate (2 packets). This can be seen from Fig. 4.3 that the knee lies to the right of traffic load of 6 (3 users times 2 packets per user). Hence, changing the order of the first three users with the largest queue sizes in the scheduling algorithm leads to similar rate vectors and does little to increase the number of distinct candidate rate vectors. As a result increasing the number of candidate rate vectors from 1 to 3 does little to reduce the average delay. However, when P is further increased to 4,

the maximum stable traffic load (or the knee) increases. This is due to the fact that the subset of candidate rate vectors expands when we first schedule the user with the 4-th longest queue, which is otherwise seldom scheduled or scheduled with the low transmission rate (1 packet) when the users are considered in the order of their queue sizes. This can be partially observed from the numerical results that the knee lies between 6.5 and 7 in Fig. 4.3. Hence, this introduces an opportunity to generate candidate rate vector quite different from those generated by considering only the first three users. For a similar reason, the maximum stable traffic load increases with $P = 5$ (compared to $P = 4$). When the value of P is increased to 6, the average delay (and hence the maximum traffic load that can be handled without losing stability) does not change much. We suspect that this is due to the fact that the additional candidate rate vectors generated schedule a user with a much smaller queue size and hence the inner products in (4.3) of these additional candidate rate vectors are smaller than those of previously generated candidate rate vectors that schedule users with larger queue sizes. The maximum stable traffic load of HEURISTIC_1 scheduling algorithm with $P = 6$ is about 95 percent of that of the optimal scheduling algorithm as shown in the figure.[1] However, remarkably our HEURISTIC_1 algorithm can achieve approximately 90 percent of the optimal policy even with $P = 1$ or $P = 2$.

We also evaluate the performance of HEURISTIC_2 algorithm, which is plotted in Fig. 4.4. As expected, the maximum stable traffic load increases with K, the size of the subset of users considered for scheduling in each timeslot. Since there are 10 users in the system, the algorithm with $K = 10$ is the optimal policy because

[1] Although the plotted average delay of the optimal policy does not increase significantly after the load of 7.5, this is due to the limited simulation run, and increasing the simulation duration results in much larger delays.

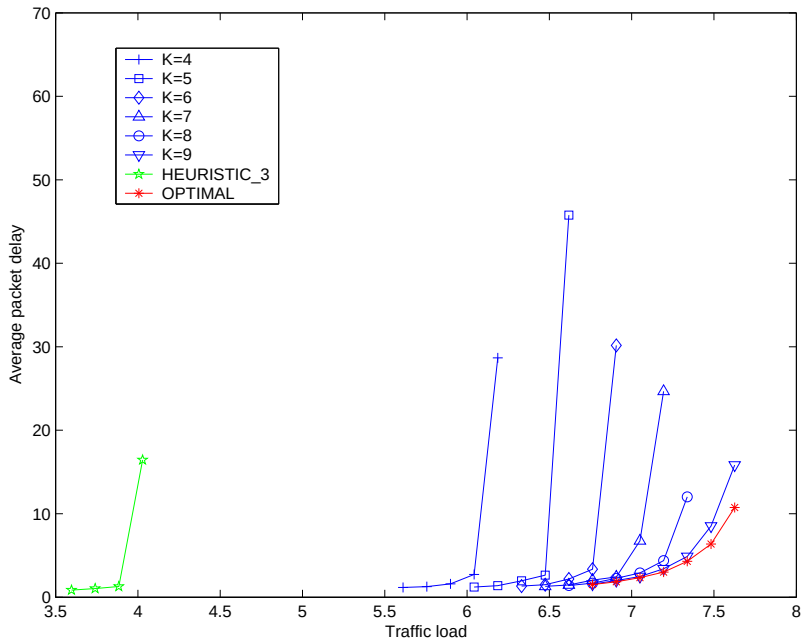


Figure 4.4: Average packet delay vs. traffic load for HEURISTC_2 and HEURISTIC_3 algorithms for a single cell.

all the possible feasible rate vectors formed by all users will be considered. The maximum stable traffic load for $K = 4$ is about 81 percent of that of the optimal scheduling policy in this example.

One thing to keep in mind is that since HEURISTIC_2 algorithm considers all possible rate vectors with a subset of K users, the complexity of the algorithm even for $K = 5$ is higher than that of HEURISTIC_1 algorithm with $P = 6$. Therefore, these simulation results of a single cell network indicate that HEURISTIC_1 algorithm may be preferable to HEURISTIC_2 algorithm as it outperforms HEURISTIC_2 algorithm with much lower complexity. In this example, the complexity of the optimal policy in (4.3) increases proportionally to $(1 + V)^{10} - 1 = 59,048$, while that of the HEURISTIC_1 algorithm with $P = 6$ increases proportionally to $PNV = 120$. Hence, the complexity of the HEURISTIC_1 algorithm is orders of

magnitude lower even with $N = 10$. Similarly, the complexity of the HEURISTIC_2 algorithm with $K = 5$ increases proportionally to $(1 + V)^K - 1 = 242$.

For the purpose of comparison, we also evaluate the performance of an algorithm that does not use the queue length information. This algorithm, called HEURISTIC_3, assigns a credit c_j to each user, for example, based on the class of service requested by the user, and keeps track of the throughput $T_j(t)$ of the users. Then, the users are scheduled in the increasing order of their ratio of throughput to credit $T_j(t)/c_j$.

ALGORITHM III: HEURISTIC_3(c_j, $T_j(t)$, $\forall j \in \mathcal{N}$)

STEP 1: Initialize $\mathcal{U} = \emptyset$, $\mathbf{r} = \underline{0}$ and $\mathcal{K} = \mathcal{N}$.

STEP 2 While $|\mathcal{K}| \neq 0$, do

- $j^* = \arg\min_{j \in \mathcal{K}} T_j(t)/c_j$
- $\mathcal{U} = \mathcal{U} \cup \{j^*\}$, $\mathcal{K} = \mathcal{K} \setminus \{j^*\}$, $\mathcal{V}_1 = \mathcal{V}$
- While $\mathcal{V}_1 \neq \emptyset$ do
  - $flag = 0$.
  - $v_m = \max\{\mathcal{V}_1\}$, $r_{j^*} = v_m$, $\mathcal{V}_1 = \mathcal{V}_1 \setminus \{v_m\}$.
  - If FEASIBILITY(P_{max}; $\mathcal{H}_j, n_j^2, \gamma_j$, $\forall j \in \mathcal{U}$) $= 1$, set $flag = 1$ and break;
- If $flag = 0$, $r_{j^*} = 0$ and $\mathcal{U} = \mathcal{U} \setminus \{j^*\}$.

STEP 3: MINIMIZE_POWER($P_{max}; \mathcal{H}_j, n_j^2, \gamma_j, \forall j \in \mathcal{U}$) .

It is clear from the pseudo-code of the algorithm that since the user with the smallest normalized throughput $T_j(t)/c_j$ is selected for scheduling, the algorithm does not take the network state, *i.e.,* queue sizes, into account.

In the simulation the credits of the users are set to their arrival rates. The average packet delay under this algorithm is plotted in Fig. 4.4. As one can see from the plot, the maximum traffic load that can be accommodated under HEURISTIC_3

algorithm is only about 51 percent of that of the optimal scheduling policy, which is considerably smaller than that of HEURISTIC_1 algorithm. This poor performance is due to the fact that the current queue lengths of the users are not considered for scheduling.

4.5 Multiple cells

In the previous section we have considered a simple case where there is only one cell in a network. In practice, in order to improve the spectral efficiency of the system, the available spectrum is reused in multiple cells, which are called *co-channel* cells. Since the same frequency band is used in more than one cell, these co-channel cells cause inter-cell interference at the users in these cells. Therefore, for a thorough evaluation of the performance of our proposed algorithms, we need to evaluate their performance in the presence of inter-cell interference that reduces the spatial separability of the users as will be shown shortly.

4.5.1 Performance Evaluation

For our simulation with multiple cells, we adopt the time correlated shading fading plus Rayleigh fading channel model in the multiple cell network described in Chapter 3, and utilize the proposed beamforming algorithms in Chapter 3. As with the single cell scenarios, we study the average packet delay with varying traffic load only in the center cell under the proposed scheduling algorithms. The BSs in the surrounding cells utilize HEURISTIC_1 algorithm with $P = 1$, and the traffic load is assumed to be constant.

The performance of HEURISTIC_1 algorithm is shown in Fig. 4.5. We observe

that the maximum stable traffic load achieved by the optimal scheduling policy in (4.3) for this multi-cell network is about 72 percent of that of a single cell network due to the presence of inter-cell interference, which reduces the spatial separability of the users.

It is worth noting that unlike in the single cell network, the maximum stable traffic load increases from $P = 1$ to $P = 2$. This is due to the inter-cell interference that makes the users less spatially separable. Therefore, switching the order of the two users with largest queue lengths typically yields different rate vectors because the two users with the largest queue sizes may not be scheduled together any more due to the lack of spatial separability. This increases the subset of candidate rate vectors, and leads to better performance and larger throughput region. Algorithm HEURISTIC_1 with $P = 2$ and $P = 3$ has similar maximum stable traffic load that is about 87 percent of that achieved by the optimal scheduling policy (4.3). Recall that HEURISTIC_1 algorithm achieves 95 percent of the maximum stable traffic load of the optimal scheduling policy in the single cell scenario. The relative performance degradation of HEURISTIC_1 scheduling algorithm in a multi-cell network is due to the fact that the users are less spatially separable in a multi-cell network in the presence of inter-cell interference. This often times prevents the users with large queue sizes from being spatially separable, and as a result they cannot be scheduled together and sequential scheduling of the users based on their queue lengths may result in a rate vector not close to the optimal one selected by (4.3).

The performance of HEURISTIC_2 algorithm with different values of parameter K is displayed in Fig. 4.6. The maximum stable traffic load increases with K as expected. With $K = 4$, the maximum stable traffic load is about 73 percent of

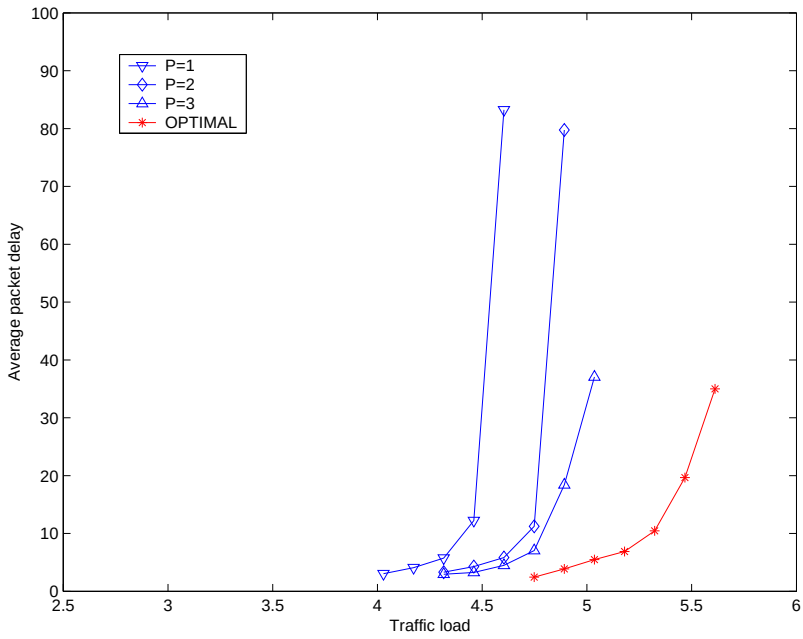


Figure 4.5: Average packet delay vs. traffic load for HEURISTIC_1 algorithm for multiple cells.

that achieved by optimal scheduling policy, compared to 81 percent for the single cell network. This is again due to the fact that users are less spatially separable with inter-cell interference. Therefore, considering the feasible rate vectors only for a small number of users often selects a rate vector that is not close to the optimal one given by (4.3).

Comparing the figures for single cell and multi-cell networks, we observe that the average packet delay increases more smoothly for multi-cell networks. This is because in a single cell scenario things are more deterministic due to the absence of inter-cell interference, while in a multi-cell scenario the presence of inter-cell interference introduces much more stochastic disturbance or randomness to system dynamics.

The performance of HEURISTIC_3 algorithm is also shown in Fig. 4.6 for com-

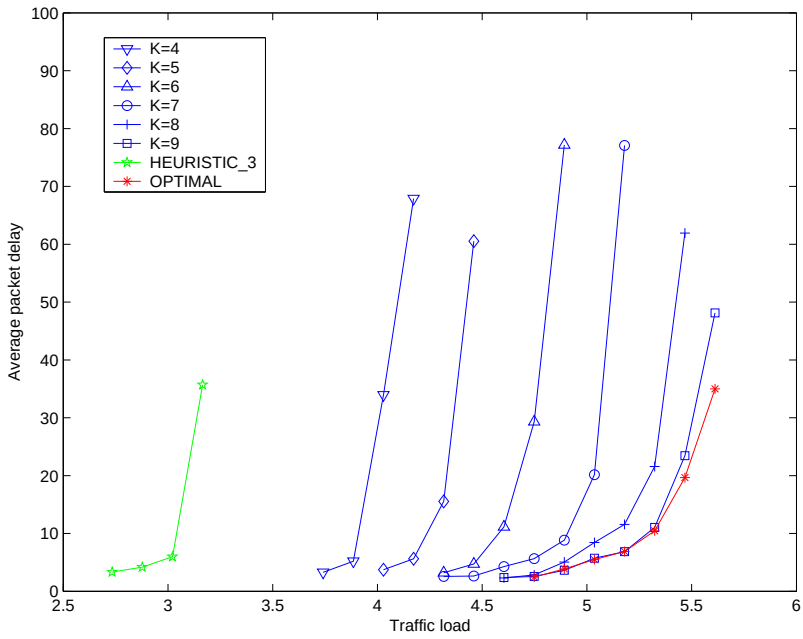


Figure 4.6: Average packet delay vs. traffic load for HEURISTIC_2 and HEURISTIC_3 algorithms for multiple cells.

parison. The maximum stable throughput of HEURISTIC_3 is about 55 percent of that of the optimal scheduling policy (4.3) in this multi-cell scenario, and is considerably smaller than that of the proposed algorithms that take the network state into account for scheduling.

4.6 Discussion

The use of antenna arrays at the base stations has been proposed to improve the system throughput and to provide quality-of-service (QoS) guarantees to mobile users in wireless networks. In this chapter we studied the problem of wireless scheduling with base station antenna arrays with a physical layer constraint of providing certain packet loss probability and higher layer QoS guarantees in the

form of throughput. An optimal scheduling policy that achieves the throughput region is derived.

We have proposed two heuristic algorithms that attempt to mimic the behavior of the optimal policy with much lower complexity. Simulation results suggest that these algorithms yield significant performance improvement over another algorithm that does not consider queue state for scheduling decisions. Furthermore, the first proposed algorithm is shown to achieve the schedulable region close to the throughput region with linear complexity in the number of candidate rate vectors and the number of users, whereas the complexity of the optimal policy increases exponentially with the number of users. Furthermore, simulation results indicate that the number of candidate rate vectors required to enjoy most of the benefits is close to the number of antenna elements at the base stations, which could be orders of magnitude smaller than the number of all feasible rate vectors.

Chapter 5

Power Control with Distributed Scheduling in Ad-Hoc Networks

5.1 Introduction

Unlike wireline or even cellular networks with a fixed infrastructure, multi-hop wireless networks can be deployed without any centralized agents and be self organized through neighbor discovery and link establishment. Because of their versatility multi-hop wireless networks offer much potential for a variety of military, scientific, and commercial applications. One of the fundamental differences between a cabled network and a (multi-hop) wireless network is the characteristics of the communication medium. In a wireline network, links are dedicated to point-to-point communication between two end nodes and do not change frequently. On the other hand, in a wireless network links are fictitious as connectivity (*i.e.,* ability to communicate) is determined by achievable signal-to-interference-and-noise ratio (SINR). Hence, the connectivity of the nodes (*i.e.,* topology of the network) is determined not only by the distance between nodes, but also by the density of the

communicating nodes as well as the performance of the underlying physical layer algorithms and availability of resources such as energy.

Lack of a fixed infrastructure and/or a time-varying topology due to mobility in a multi-hop wireless network renders a centralized packet scheduling difficult because of prohibitive required communication overhead and delays. Moreover, coordinated scheduling based on a *pre-determined* sequence of packet scheduling vectors agreed to by *all* nodes, is difficult to realize in practice when topology varies with time. Therefore, nodes must rely on distributed packet scheduling and necessary physical layer resource allocations (*e.g.,* power control) to support the packet scheduling, possibly with some local coordination. Distributed packet scheduling and power control results in *random* interference at the receivers as the interference cannot be predicted accurately without the *full knowledge* of the set of transmitters and their transmission powers during packet reception. A similar problem arises even in a CDMA system where Pseudo-Noise (PN) sequences are used for different links to reduce the interference when nodes operate in an asynchronous manner. This is because the interference experienced at a receiver depends on the set of links that are accessing the channel simultaneously and their transmission powers. This potentially *widely varying, unknown* interference at a receiver causes uncertainty in achieved SINR value during packet reception even when accurate channel gains are available.

We show that, unfortunately, most of previously proposed physical layer models adopted for performance evaluation and algorithm design do not accurately capture the effects of stochastic nature of interference at the receivers. Thus, the simulation results obtained using these inaccurate physical layer models can be misleading and give misguiding intuition. Here we only focus on the aspect of a

physical layer model that decides whether a packet transmission is successful or not. Consequently, algorithms designed based on such premises will not perform satisfactorily in practice (see subsection 5.3.1 for a numerical example). Furthermore, these models do not reveal some of important, intrinsic trade-offs in wireless network operation.

In this chapter we first propose the use of a more accurate physical layer model based on link curves for performance evaluation and algorithm design. Based on this more accurate physical layer model, we develop a new power control algorithm that can provide a physical layer quality-of-service (QoS) in terms of packet loss probability (PLP). For algorithm design and performance evaluation, PLP is the suitable physical layer parameter to consider, for the performance of higher layer protocols depends on the achieved PLPs at the physical layer and the SINR affects their performance only *indirectly* through achieved PLP. We show that our novel and yet *simple* approach leads to a new paradigm for *robust* algorithm design that does not require unrealistic assumptions on the interference estimation or physical layer behavior, with minimal communication overhead. Simplicity and robustness of an algorithm is of paramount importance, as multi-hop wireless networks are envisioned to operate in widely varying and sometimes unexpected environments, supporting a variety of applications.

This chapter is organized as follows: Section 5.2 summarizes some of previously proposed physical layer models for simulation and analysis. Section 5.3 explains the nature of random interference at the receivers and the shortcomings of previous models, which is followed by our proposed power control algorithm in Section 5.4, based on a more accurate physical layer model using link curves. We study the problem of minimizing the average total transmission power as an optimization

problem and establish the convergence results in Section 5.5. We conclude in Section 5.6.

5.2 Background

In this section we will summarize some of previously proposed physical layer models that have been used for performance analyses and algorithm designs in the past. Some of the most widely used models include (i) the disk model [44–46], (ii) a model based on received signal strength [47–49], and (iii) a model based on SINR [44,50–52]. Researchers have also used the simple *node exclusive interference model* in evaluating the performance of congestion control and scheduling algorithms [53, 54]. However, since this model does not reflect the true characteristics of wireless medium very accurately, we do not consider it in this thesis. .

Under the disk model two nodes can communicate reliably if the distance between them is smaller than some threshold value R (*e.g.,* transmission range of nodes) and there is no other node transmitting within another threshold value $\widetilde{R}$ from the receiver. Oftentimes the transmission power is assumed to be fixed, *i.e.,* there is no power control. The second model is similar to the disk model in that two nodes can communicate reliably only if they lie within some threshold value from each other and the received signal strength constraint can be satisfied at the maximum transmission power, but now transmitters exercise power control so that, given the channel gain, the received signal strength at a receiver equals some target signal strength, in order to reduce the interference to other receivers. The third model computes the realized SINR value at the receivers and assumes that the transmission is successful if and only if the achieved SINR exceeds certain target threshold SINR value. In the last two models the target received signal

strength (model (ii)) or the target SINR (model (iii)) can be viewed as the physical layer QoS parameters.

5.3 Stochastic nature of interference & its implications on network performance

This section describes one of major shortcomings of the physical layer models in Section 5.2 that does not permit simulation of realistic scenarios, and presents a numerical example that demonstrates its implications.

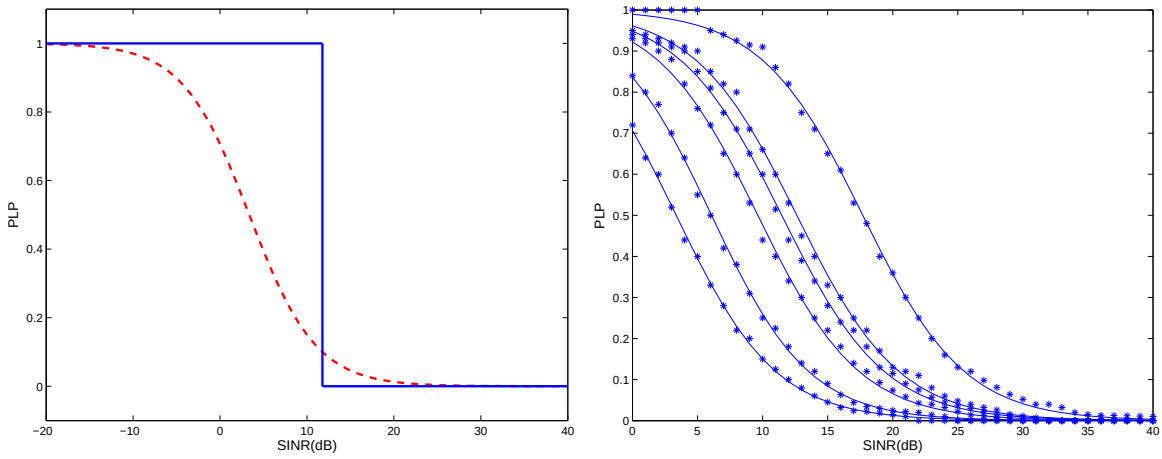
In a multi-hop wireless network it is unlikely that there will be a centralized controller that carries out packet scheduling, power control, and other physical layer resource allocation. If no such centralized agent is available, the nodes must rely on distributed packet scheduling and power control. When packet scheduling is carried out in a distributed manner, since the set of transmitter-receiver pairs is time-varying and is not known in advance, neither the transmitters nor receivers can accurately predict the interference during the packet reception.

In the first two models of Section 5.2, the issue of (random) interference is not relevant at all as only the distance between a pair of nodes or achieved signal strength as a function of distance and transmission power is used to determine whether a packet transmission is successful or not. The third model is used often under the assumption that all transmitter-receiver pairs are known as done under a *centralized* scheme so that the transmission powers that satisfy the target SINR can be computed for *all* transmitter-receiver pairs together [51, 55]. When the interference can be computed accurately as under a centralized approach, this type of model and the threshold policy for determining successful packet transmission

provide a reasonable approximation as the physical layer QoS requirement, such as target PLP, can be translated to a corresponding target SINR. However, when the interference is random, this model no longer provides a good approximation of the physical layer behavior because even the slightest drop below the target SINR leads to *unsuccessful* packet transmissions. Hence, for a fixed transmission power the packet transmission is successful if and only if the realized interference is less than or equal to certain threshold value that depends on the transmission power and target SINR threshold value because of the *discontinuous* threshold rule used to determine successful packet transmissions.

In practice the probability of successful transmission is given by a link curve. A link curve gives the PLP as a function of the achieved SINR, and is typically a *continuous* function of the SINR. Hence, in the presence of random interference the achieved PLP depends on the distribution of interference and the sensitivity of the link curve to the SINR, which are *not* sufficiently captured by any of the above models. As a result these models do not allow accurate estimation of the achieved PLP and evaluation of network performance under stress. For instance, consider the link curves in Fig. 5.1(a). The discontinuous one represents the threshold rule used by the third model with discontinuity taking place at the target SINR, whereas the continuous function is the same fitting curve to collected data in a TDMA system we utilized in Chapter 3. [28,33] (left most curve in Fig. 5.1(b)). In the example if the distribution of realized SINR in dB is continuous and symmetric (*e.g.,* Gaussian distribution) with respect to the target SINR value, the third model will tell us that the achieved PLP is 50 percent as the probability of SINR lying to the left of the target SINR is 1/2. However, if the distribution is concentrated around the target value (*i.e.,* small variance), then in reality the achieved PLP

obtained from the continuous link curve will be close to the target PLP. Thus, the PLP predicted by the third model may be considerably higher than the actual PLP.



(a) (b)

Figure 5.1: (a) An example of a link curve and a discontinuous threshold policy, and (b) link curves of a TDMA system [33].

The above observations argue for the need of a faithful physical layer model that captures the impacts of the randomness in interference and allows network engineers to design physical layer algorithms that will deliver consistent and predictable performance both in simulation and in practice. For accurate modeling of physical layer, the events of successful packet transmission must be modeled in a realistic manner rather than relying on simple threshold policies as done in the previous models. To this end we use a physical layer model where the event of a successful transmission is determined by the achieved SINR and link curves obtained from field measurements (which are available for various MCS schemes) [5, 33, 36]. This seemingly simple change in the physical layer model (compared to the third model based on SINR and threshold policy) has rather

profound implications on network performance evaluation (subsection 5.3.1) and algorithm design (section 5.4). First, this allows us to accurately model the PLPs experienced at the receivers and the impact of (random) interference on the network performance. Second, we will show in Section 5.4 that, when link curves are used for determining successful packet transmissions, although achieving a fixed target SINR is not possible, achieving a target PLP can be accomplished through a rather *simple* and *robust* mechanism with minimal communication overhead even in face of random interference without having to introduce unrealistic assumptions on the nature of interference or physical layer behavior, allowing more faithful simulation of multi-hop wireless networks.

5.3.1 Numerical Example

In this subsection, using a numerical example, we will illustrate the discrepancy in performance between the simulation results obtained under previous models and the more realistic physical layer model using link curves described in the previous subsection. Here we are interested in studying the achieved PLPs and the impact of distributed scheduling on the overall network performance including higher layer protocols. We will show that the algorithms based on disk models do not perform satisfactorily under distributed scheduling even though there is no contention or interference within the transmission range. The provided example is not intended to be a comprehensive study with the previously proposed models. Instead, its goal is to highlight some of the problems that might arise when these models are used for analyses and algorithm design.

In our example 100 nodes are randomly placed in a 1 km $\times$ 1 km rectangular region. For the simplicity of demonstration we assume a discrete-time system

throughout and the time is slotted into contiguous timeslots. The transmission range of the nodes is denoted by R, and the set of nodes is given by $\mathcal{I}$. We assume $R = \widetilde{R}$.

The scheduling algorithm we use for simulation with the disk-based models is simple. In each timeslot we find a set of links to be transmitted in a sequential manner as follows. In each iteration we randomly select a potential transmitter from a set of possible transmitters that can transmit without violating the physical layer constraint of the disk model. Then, if there exists a valid receiver within the transmission range of the transmitter, a packet transmission is scheduled to a receiver randomly selected among such nodes within the transmission range. We repeat this until no more transmitter-receiver pair can be scheduled without violating the physical layer constraint.

I. Scheduling policy with the disk model:

STEP 1: Set $\mathcal{T} = \mathcal{I}$ and $\mathcal{R} = \mathcal{I}$.

STEP 2: While 1, do

- Select a node $j \in \mathcal{T}$.
- If there is a node in $\mathcal{R}$ that lies within the transmission range of node j, then
  - randomly select a node $i \in \mathcal{R}$ within the transmission range of node j;
  - let $R(j) = i$; % node i is the receiver of transmitter j
  - $\mathcal{R} \leftarrow \mathcal{R} \setminus \{i, j\}$ and $\mathcal{T} \leftarrow \mathcal{T} \setminus \{i, j\}$;
  - remove all nodes in $\mathcal{R}$ within the transmission range of node j from $\mathcal{R}$;
  - remove all nodes in $\mathcal{T}$ within the transmission range of node i from $\mathcal{T}$;

  else
  - $\mathcal{T} \leftarrow \mathcal{T} \setminus \{j\}$;

  end
- If $\mathcal{T} = \emptyset$ or $\mathcal{R} = \emptyset$,

```
        – break;
    end
end
```

Note that the physical layer constraint of the disk model is enforced explicitly during the link scheduling.

Clearly, this scheduling policy is not designed to support any flow rates between source and destination pairs. Instead this scheduling policy typically selects very different scheduling vectors (a set of scheduled links) from one timeslot to next, hence ensuring sufficient randomness in interference at the receivers. However, we suspect that this is a reasonable approximation to the network behavior when the network is congested and many queues are not empty. When the network (or a neighborhood) is congested, queues begin to build up and nodes will choose different links to transmit on in consecutive timeslots (if possible) rather than transmitting to the same neighbor for many consecutive timeslots, in an attempt to prevent other queues from overflowing and experiencing high packet drop probabilities and to produce more smooth flow of packets throughout the network and reduce the delay jitter of packets. Therefore, the set of scheduled links will change dynamically from one timeslot to next as done in our simulation. This will also result in *weak* temporal correlation in the interference experienced at the receivers and make it difficult to accurately predict the interference to be experienced during a packet reception from the current estimate.

The same issue exists in an asynchronous system as well, whether it is a TDMA, CDMA, or OFDM system. In these asynchronous systems, including a CDMA system where pseudo-noise sequences are assigned to different links, the experienced

interference during a packet reception depends on the set of other simultaneous packet transmissions and the amounts of overlap in time.

Under the disk model the transmission power is fixed at one, and we assume that the system is interference limited and ignore the receiver noise. Under this assumption the realized SINR at a receiver i is given by

$$SINR_i = \frac{P_{T(i)} G_{T(i)i}}{\sum_{j \in \mathcal{R}^\star \setminus \{i\}} P_{T(j)} G_{T(j)i}} \ , \tag{5.1}$$

where $\mathcal{R}^\star$ is the set of receivers, G_{ij} is the path gain from node i to node j, P_j is the transmission power of node j, and $T(j)$ denotes the intended transmitter of receiver j. Here the path gain G_{ij} is given by $d_{ij}^{-\beta}$, where d_{ij} is the distance between i and j, and β is the path-loss exponent and is set to 3 in the simulation. One can easily see from (5.1) that if the transmission power of the transmitters is the same, the realized SINR does not depend on the selected transmission power due to cancellation. In this case the SINR value is simply given by $SINR_i = \frac{G_{T(i)i}}{\sum_{j \in \mathcal{R}^\star \setminus \{i\}} G_{T(j)i}}$.

The scheduling algorithm used with the second model is identical to that used with the disk model except for the transmission power selected for the transmitters. Here we assume that the transmission power of a transmitter is chosen so that the received signal strength at the intended receiver equals 10^{-5}. Note that the received signal strength is linear in the transmission power. Hence, similarly as with the disk model, the realized SINR at the receivers does not depend on the selected target received signal strength from (5.1).

In the numerical example, although disk models are used in link scheduling for imposing certain physical layer constraint, for determining successful packet transmissions we use the realized SINR in (5.1) and the continuous link curve in Fig. 5.1(a) to model a more realistic physical layer and to show the discrepancy

in performance of the scheduling algorithm using the two different physical layer models.

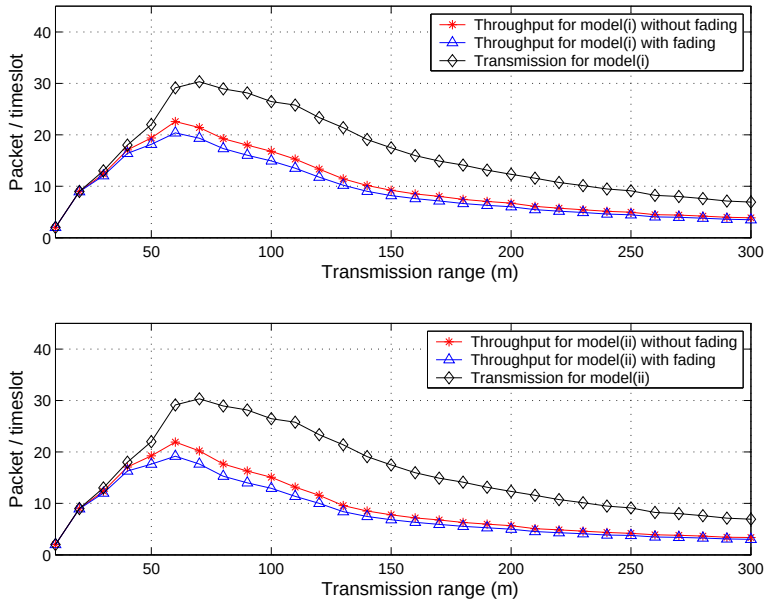


Figure 5.2: Plot of the number of transmissions per timeslot and throughput.

We plot the average number of scheduled packet transmissions per timeslot and the average number of successful transmissions per timeslot (throughput) both with and without shadow fading as a function of selected transmission range R under these two models in Fig. 5.2. The probability of successful transmission is given by the ratio of the throughput to the number of scheduled transmissions. The shadow fading is modeled using independent and identically distributed lognormal random variables (rvs). The mean and variance of the Gaussian rvs are 0 and 1, respectively. In the case of model (ii) with shadow fading we assume that the channel gain with fading is known at the transmitter and the power control algorithm selects the transmission power so that the received signal strength equals the target received strength.

When the transmission range is small, the number of transmissions is low and

the probability of successful transmission is high. This is due to the fact that when the transmission range is sufficiently small, network connectivity is poor and the set of nodes that can communicate is small and, as a result, the interference experienced at the receivers remains low. However, as the transmission range increases, the connectivity of the network improves and the interference at the receivers increases at the same time as more nodes transmit in each timeslot. This can be seen from the sharp drop in the probability of successful transmission in Fig. 5.2. In fact, at the transmission range that maximizes the system throughput the probability of successful transmission is well below 80 percent. Hence, simulation results that ignore unsuccessful packet transmissions will incorrectly overestimate the performance of the network, yielding misleading results. In fact, the impact of ignoring unsuccessful packet transmissions goes beyond a reduction in network throughput, for high PLP tends to have adverse effects on packet loss probability due to buffer overflow and the performance of end-to-end congestion control mechanism. For these reasons, in order to avoid unanticipated adverse effects on the higher layer protocols, the physical layer algorithm should attempt to achieve certain maximum target PLP. In addition, note that when a power control algorithm is used to reduce the interference (model (ii)), this results in larger PLPs and smaller throughput. Therefore, simulation results that do not consider this issue will only emphasize the benefits of power control without revealing the downside.

Another issue with the disk-based models that has not been addressed satisfactorily in the past is the selection of suitable transmission range. As shown in Fig. 5.2 the performance of the system, both in terms of the throughput and achieved probability of successful transmission, depends very much on the selected

transmission range. However, it is not clear how to select a transmission range R suitable for performance evaluation in advance.

5.4 Proposed Power Control Algorithm

In this section we approximate the PLP from a link curve as described in Chapter 3, and then, using the approximation, describe the proposed power control algorithm that can handle the issue of random interference and provide PLP guarantees. Using the same set-up used in the previous example, we demonstrate that the proposed algorithm does achieve the target PLPs.

From the viewpoint of a transmitter-receiver pair, the goal of a power control algorithm is to find the transmission power that will satisfy the target PLP. As mentioned in the previous section, this problem is complicated by the fact that the interference experienced at the receiver is difficult to predict.

5.4.1 Approximation of Packet Error Rate

Our proposed approach to power control does not make any assumptions regarding the nature/distribution of the interference, and is simple and robust. As in Chapter 3, it is based on the observation that the link curves can be well approximated by the following family of functions

$$PLP(SINR) = \frac{1}{1 + e^{k(SINR_{dB} - z)}} \tag{5.2}$$

where the fitting parameters can be determined from the given link curves *off-line*.

For the simplicity of illustration we assume that accurate channel gains are available at the transmitter, which can be estimated from control packets. For a

reasonably small target PLP, we can approximate (5.2) as follows:

$$PLP(SINR) \approx e^{-k(SINR_{dB}-z)} = e^{kz} SINR^{-\alpha} \tag{5.3}$$

where $\alpha = 10k/\ln 10$ and determines the sensitivity of PLP to SINR. Since the realized SINR is a rv due to random interference, assuming necessary ergodicity, the realized average PLP is given by

$$PLP_{avg} = e^{kz} \mathbf{E}\left[SINR^{-\alpha}\right] \ . \tag{5.4}$$

In the rest we replace the approximation in (5.3) with an equality.

5.4.2 Proposed Power Control Algorithm

In order to achieve the average PLP close to the target PLP, the transmitter of link l must select the transmission power so that

$$\begin{aligned} e^{kz} \mathbf{E}\left[SINR_l^{-\alpha}\right] &= e^{kz} \frac{\mathbf{E}\left[(\text{Interference}_l)^{\alpha}\right]}{(P_l \cdot G_l)^{\alpha}} \\ &= PLP_{target} \ . \end{aligned} \tag{5.5}$$

Here Interference_l includes both the interference and noise at the receiver of link l. Note that $\mathbf{E}\left[(\text{Interference}_l)^{\alpha}\right]$ in (5.5) is the mean of $(\text{Interference}_l)^{\alpha}$ at the receiver of link l during packet receptions. Thus, the transmission power should be set to

$$P_l = \left(e^{kz} \frac{\mathbf{E}\left[(\text{Interference}_l)^{\alpha}\right]}{PLP_{target} \cdot G_l^{\alpha}} \right)^{1/\alpha} \ . \tag{5.6}$$

Note from (5.6) that when accurate channel gains are available, the transmitter requires only one parameter $\mathbf{E}\left[(\text{Interference}_l)^{\alpha}\right]$ to compute the transmission power. This parameter can be estimated using exponential averaging. In other words, the estimate for link l is updated after each packet transmission over link l

according to

$$\mathbf{E}\left[(\text{Interference}_l)^\alpha\right]_{new} = (1-\omega)\cdot\mathbf{E}\left[(\text{Interference}_l)^\alpha\right]_{old} + \omega\cdot(\text{Interference}_{l,cur})^\alpha \quad (5.7)$$

where $\text{Interference}_{l,cur}$ is the new experienced interference. This estimate can be either fed to the transmitter by the receiver when it experiences a significant change in its value or piggybacked in the acknowledgment after each transmission.

In practice, in order for exponential averaging in (5.7) to be effective, the averaging constant w must be selected large enough so that the estimate can be updated in a timely manner with time-varying channel conditions due to mobility and (slow) fading. However, if a link is not used often enough, the receiver may not be able to update the estimates often enough and these estimates may not be accurate.

In order to solve this problem we can maintain only one estimate at each receiver rather than per link. Hence, after every packet reception the node updates the estimate according to (5.7). This reduces the number of parameters each node needs to maintain to one, leading to a more scalable algorithm regardless of the density of the network, and faster convergence of the estimates. In the numerical example in the following subsection, we adopt this simpler version of the algorithm and show that it achieves realized PLPs very close to the target PLPs.

5.4.3 Numerical Example

We have simulated our power control algorithm with various target PLPs. The scheduling policy used in the simulation is similar to that used with the disk models, and the same link curve is used for determining successful packet transmissions. However, a possible receiver is not limited to nodes within a transmission range of

the transmitter, and any node to which the selected (potential) transmitter j can communicate while satisfying the target PLP requirement with a maximum power budget of 10 can be selected as a receiver. In addition, we have introduced a noise of 10^{-14} at the receivers to study the system throughput and energy consumption as a function of target PLP.

Numerical results are shown in Fig. 5.3. Here we only show the results with lognormal fading. In all cases the realized PLPs of all transmitter-receiver pairs were close to the target PLP, as shown in Fig. 5.4. One important thing to note from Figs. 5.3(a) and 5.3(b) is that both achieved throughput and average transmission power normalized by the probability of successful transmission increase with the target PLP over the region of interest (PLP $\leq$ 15 percent). In fact, the gain in throughput is more than 30 percent (from 16.8 to 22) when the target PLP is increased from 2 percent to 10 percent. Furthermore, the achieved throughput with target PLP of 10 percent is comparable to the maximum throughput achieved in Fig. 5.2 (approximately 22). As the target PLP is raised from 2 percent to 10 percent, the average transmission power per successful transmission increases by 15 percent at the same time (from 3.67 to 4.2 in Fig. 5.3(b)). We have also run the same simulation with a different link curve. Fig. 5.5 shows the same plots using a link curve with $\alpha = 2$. One can see from Fig. 5.5 that the system throughput tends to be larger for most values of target PLP and the change in both throughput and energy consumption is more pronounced with varying target PLP than in the previous case with $\alpha = 1.1$ (60 percent increase in throughput and 50 percent increase in energy consumption when target PLP is increased from 2 percent to 10 percent), due to higher sensitivity of PLP to SINR. Therefore, Figs. 5.3 and 5.5 suggest a clear trade-off between the network throughput and energy consumption

with target PLP as the control parameter.

Our simulation results also reveal that the transmission power among the nodes varies widely, which depends on the characteristics of the interference experienced at the receivers. Fig. 5.3(c) indicates that the interference experienced at the receivers varies considerably both in its mean value and variance. Large variance of interference implies larger $\mathbf{E}\left[(\text{Interference}_l)^{\alpha}\right]$, resulting in more restrictive physical layer constraints, and higher transmission power. Consequently, fewer number of simultaneous transmissions are possible in a neighborhood.

5.5 Optimal Power Control & Convergence

The previous section tells us that one can provide physical layer QoS in the form of PLP under distributed scheduling even when the exact value of interference is not known at a receiver during packet reception. In a wireless network many nodes are expected to operate on batteries and, hence, are energy constrained. Therefore, the power control algorithm should not only satisfy the physical layer QoS, but should also minimize the energy consumption at the nodes at the same time. This problem can be studied in an optimization framework.

5.5.1 Optimization Formulation

Let $\mathcal{I} = \{1, \ldots, I\}$ denote the set of nodes and $\mathcal{L} = \{1, \ldots, L\}$ the set of unidirectional links. Here a link is a pair of nodes (i, j) such that node i can communicate to node j. We are given a set of source-destination pairs $\mathcal{K} = \{1, \ldots, K\}$. Each source-destination pair has certain flow rate demand/requirement, and the demand (in bits per timeslot) of the k-th pair is denoted by x_k. The routes of the source-

destination pairs are fixed, and the routing matrix is given by a $K \times L$ matrix A, *i.e.,* $A_{kl} = 1$ if the route of the k-th source-destination pair traverses link l, and $A_{kl} = 0$ otherwise. Let $\widetilde{\mathcal{L}} = \{l \in \mathcal{L} | A_{kl} = 1 \text{ for some } k \in \mathcal{K}\}$, and $\widetilde{L} = |\widetilde{\mathcal{L}}|$. In the rest of this section we focus on the links in $\widetilde{\mathcal{L}}$ as other links are not being used. For simplicity assume that the transmission rates (in the unit of bits per timeslot) of the links are constant, which are given by a diagonal transmission rate matrix R. In other words, $R_{ll}, l \in \widetilde{\mathcal{L}}$, is the transmission rate of link l.

Let $\underline{s}$ be an $\widetilde{L} \times 1$ scheduling vector where $s_l = 1$ if link $l \in \widetilde{\mathcal{L}}$ is on, *i.e.,* transmitter of link l sends a packet to the receiver of link l. We only consider scheduling vectors that satisfy the following: no node (1) receives and transmits simultaneously, (2) receives from more than one node, or (3) transmits to more than one receiver. Obviously, some, if not all, assumptions can be relaxed, depending on the capabilities of devices. We denote the set of scheduling vectors satisfying these conditions by $\mathcal{S}$.

We assume that the state of the system can be modeled by an ergodic discrete-time Markov chain. The scheduling policy, whose function is to select a feasible scheduling vector $\underline{s} \in \mathcal{S}$ in each timeslot, is stationary, *i.e.,* scheduling decisions depend on the state of the system, but not on time. For example, under a random access scheduling policy, each node will attempt to schedule a link with a non-empty queue with certain probability that may depend on the queue size. A pair of links that have the same destination node will collide and a subset of the attempted links that do not collide with others will be scheduled successfully. Under this assumption one can compute the probability a particular scheduling vector will result given the system state. We assume that the system is at steady state, and the stationary distribution is given by π. We denote the resulting steady-

state distribution over $\mathcal{S}$ by $\mathbf{d}$. In other words, $\mathbf{d}_{\underline{s}}, \underline{s} \in \mathcal{S}$, is the probability that scheduling policy selects scheduling vector $\underline{s}$ at steady state. This probability is given by

$$\mathbf{d}_{\underline{s}} = \sum_{\theta \in \Theta} \pi_\theta \cdot \mathbf{p}_{\underline{s}|\theta} \ ,$$

where Θ is the state space of the system, and $\mathbf{p}_{\underline{s}|\theta}$ denotes the conditional probability that scheduling vector $\underline{s}$ will be selected given that the system is at state θ.

We assume that the resulting distribution $\mathbf{d}$ satisfies

$$(I_{\widetilde{L}\times\widetilde{L}} - \Gamma) \cdot R \sum_{\underline{s}\in\mathcal{S}} \mathbf{d}_{\underline{s}} \cdot \underline{s} \geq \widetilde{A}^T \underline{x} \ , \tag{5.8}$$

where $\Gamma = \mathrm{diag}(\lambda_l; l \in \widetilde{\mathcal{L}})$, λ_l is the target PLP of link l, $I_{\widetilde{L}\times\widetilde{L}}$ is an $\widetilde{L} \times \widetilde{L}$ identity matrix, and $\widetilde{A}$ is a submatrix of the routing matrix A only with the columns corresponding to the links in $\widetilde{\mathcal{L}}$. The left hand side of (5.8) is the vector of the average goodput over the links, and the right hand side is the link demands determined by the rate demand vector $\underline{x}$ and the routing matrix $\widetilde{A}$. If the transmission power levels are fixed and we ignore channel fading, the distribution of the interference experienced at the receivers is completely determined by the distribution $\mathbf{d}$ and transmission power vector $\mathbf{p}$.

We formulate the problem of power control as the following optimization problem:

$$\begin{aligned} \text{minimize}_{\mathbf{p}\in\mathcal{P}} \quad & \mathbf{p}^T \Big(\sum_{\underline{s}\in\mathcal{S}} \mathbf{d}_{\underline{s}} \cdot \underline{s} \Big) \\ \text{subject to} \quad & PLP_l(\mathbf{p}, \mathbf{d}) \leq \gamma_l \ , \end{aligned} \tag{5.9}$$

where $\mathcal{P} = \prod_{l\in\widetilde{\mathcal{L}}} [\mathbf{p}_{l,\min}, \mathbf{p}_{l,\max}]$, and $\mathbf{p}_{l,\min}$ and $\mathbf{p}_{l,\max}$ are the minimum and maximum power constraints of link l, respectively. The minimum power constraint

exists because the transmission power of a radio device cannot be arbitrarily small. We assume that the solution of (5.9) is an interior point of $\mathcal{P}$, *i.e.,* the constraints are not active at the solution. Note that the PLPs depend both on the transmission powers and the distribution $\mathbf{d}$ because the interference experienced at the receivers depends on both.

We assume that the channel gains are fixed, and they are denoted by $G_l, l \in \widetilde{\mathcal{L}}$. We define $\mathbf{G} = \text{diag}(G_l, l \in \widetilde{\mathcal{L}})$ and, for each $l \in \widetilde{\mathcal{L}}$,

$$\mathbf{d}^l_{\underline{s}} = \mathbf{P}\left[\text{scheduling vector } \underline{s} \text{ selected} \mid \text{link } l \text{ scheduled}\right] .$$

Note that $\mathbf{d}^l_{\underline{s}}$ is the conditional probability that the scheduling vector $\underline{s}$ is selected given that link l is scheduled. Let $\mathbf{G}^l = \text{diag}(G_{Tx(l')Rx(l)}; l' \in \widetilde{\mathcal{L}})$, where $Tx(l)$ and $Rx(l)$ are the transmitter and receiver of link l, respectively. Under this assumption, it is plain to see that

$$\begin{aligned}\mathbf{E}\left[SINR_l^{-\alpha}\right] &= \sum_{\underline{s}\in\mathcal{S}:\underline{s}_l=1} \mathbf{d}^l_{\underline{s}} \left(\frac{\mathbf{p}^T\mathbf{G}^l\underline{s} - G_l\cdot\mathbf{p}_l + n_l}{G_l\cdot\mathbf{p}_l}\right)^{\alpha} \\ &= \frac{\mathbf{E}\left[(\text{Interference}_l)^{\alpha}\right]}{(G_l\cdot\mathbf{p}_l)^{\alpha}}\end{aligned} \tag{5.10}$$

where $n_l > 0$ is the variance of the noise at the receiver of link l.

One can easily see from (5.10) that $\mathbf{E}\left[SINR_l^{-\alpha}\right]$ is convex in each $\mathbf{p}_{l'}, l' \neq l$ if $\alpha \geq 1$ and is strictly convex if there exists $\underline{s}'$ such that $\mathbf{d}^l_{\underline{s}'} > 0$, $\underline{s}'_l = \underline{s}'_{l'} = 1$ and $\alpha > 1$. As most of the link curves, if not all, that we have seen have α larger than one, we assume that $\alpha > 1$ [33, 36].

5.5.2 Uniqueness of Solution

Define a multi-dimensional mapping $F(\mathbf{p})$, where

$$F_l(\mathbf{p}) = \min\left\{\mathbf{p}_{l,\max}, \max\left\{\mathbf{p}_{l,\min}, \left(e^{kz}\frac{\mathbf{E}\left[(\text{Interference}_l(\mathbf{p}))^{\alpha}\right]}{\gamma_l\cdot G_l^{\alpha}}\right)^{1/\alpha}\right\}\right\} . \tag{5.11}$$

It is easy to see that the solution to the optimization problem in (5.9) must be a fixed point of the mapping F. This is because at the solution the transmission power of each link must be the smallest transmission power that satisfies the PLP constraint given the transmission powers of other links, which is obtained from the mapping F.

The following lemma tells us that there exists a unique fixed point of the mapping F.

Lemma 5.5.1 *There exists a unique fixed point of the mapping $F(\cdot)$.*

Proof A proof is given in Appendix A.3 ■

Combined with the previous observation that the solution to (5.9) is a fixed point of the mapping F, Lemma 5.5.1 tells us that the unique fixed point of the mapping is the solution to (5.9).

We now investigate the problem of convergence of the distributed power control algorithm to the solution.

5.5.3 Synchronous Update

In this subsection we first consider the simpler case where the updates of the transmission powers are synchronized and the updates are based on the latest values. Consider the following updating rule. We model the updates with a discrete-time model. For each $n = 0, 1, 2, \ldots$, let $\mathbf{p}(n) = (\mathbf{p}_l(n); l \in \widetilde{\mathcal{L}})$, and each link updates its transmission power according to

$$\mathbf{p}_l(n+1) = F_l(\mathbf{p}(n)) \ . \tag{5.12}$$

Once all links update their transmission powers, they wait long enough so that they can estimate $\mathbf{E}\left[(\text{Interference}_l(\mathbf{p}))^{\alpha}\right]$. Once this estimate is available at all

links, they repeat the above update procedure, based on the new estimates. This is called *Jacobi update scheme.*

We assume that $\mathbf{p}(0) \in \mathcal{P}$. The following lemma tells us that the link transmission powers $\mathbf{p}(n)$ converge to the solution.

Lemma 5.5.2 *Under the update rule (5.12) we have* $\lim_{n\to\infty}\mathbf{p}(n) = \mathbf{p}^\star$*, where* $\mathbf{p}^\star$ *is the unique fixed point of the mapping* F*.*

Proof A proof is given in Appendix A.4 ■

5.5.4 Asynchronous update

The convergence results in the previous subsection assume that users are synchronized and the latest information is available for every link. However, in practice it is unlikely that such updates will take place simultaneously or even at the same update frequency, and in many cases only delayed information may be available depending on the update frequency and so on. Hence, it is important to show the convergence of the update algorithm under an asynchronous update scheme with possibly delayed information.

Let T_l be the set of periods at which the transmission power of link l is updated, and

$$\mathbf{p}_l(n+1) = F_l(\mathbf{p}(\tau_l(n))) \qquad \text{for all } n \in T_l \ , \tag{5.13}$$

where $0 \leq \tau_l(n) \leq n$. We assume that the sets $T_l, l \in \widetilde{\mathcal{L}}$, are infinite and if $\{n_k\}$ is a sequence of elements in T_l that tends to infinity, then

$$\lim_{k\to\infty} \tau_l(n_k) = \infty \ .$$

This update scheme is called a *totally asynchronous update scheme.*

The following lemma tells us that the link transmission powers converge to the solution under totally asynchronous updates, starting from any initial vector $\mathbf{p}(0) \in \mathcal{P}$.

Lemma 5.5.3 *Under the update rule (5.13) we have* $\lim_{n\to\infty}\mathbf{p}(n) = \mathbf{p}^\star$ *for all* $\mathbf{p}(0) \in \mathcal{P}$.

Proof A proof is given in Appendix A.5 ■

5.6 Discussion

We studied the problem of distributed power control in the presence of unknown interference at the receivers. Using a more accurate physical layer model based on link curves, we developed a new power control algorithm that is simple and robust and can provide guaranteed packet error rate (PLP). We then formulated the problem of minimizing the average aggregate transmission power as an optimization problem, and showed that the proposed power control algorithm converges to a solution of the optimization problem.

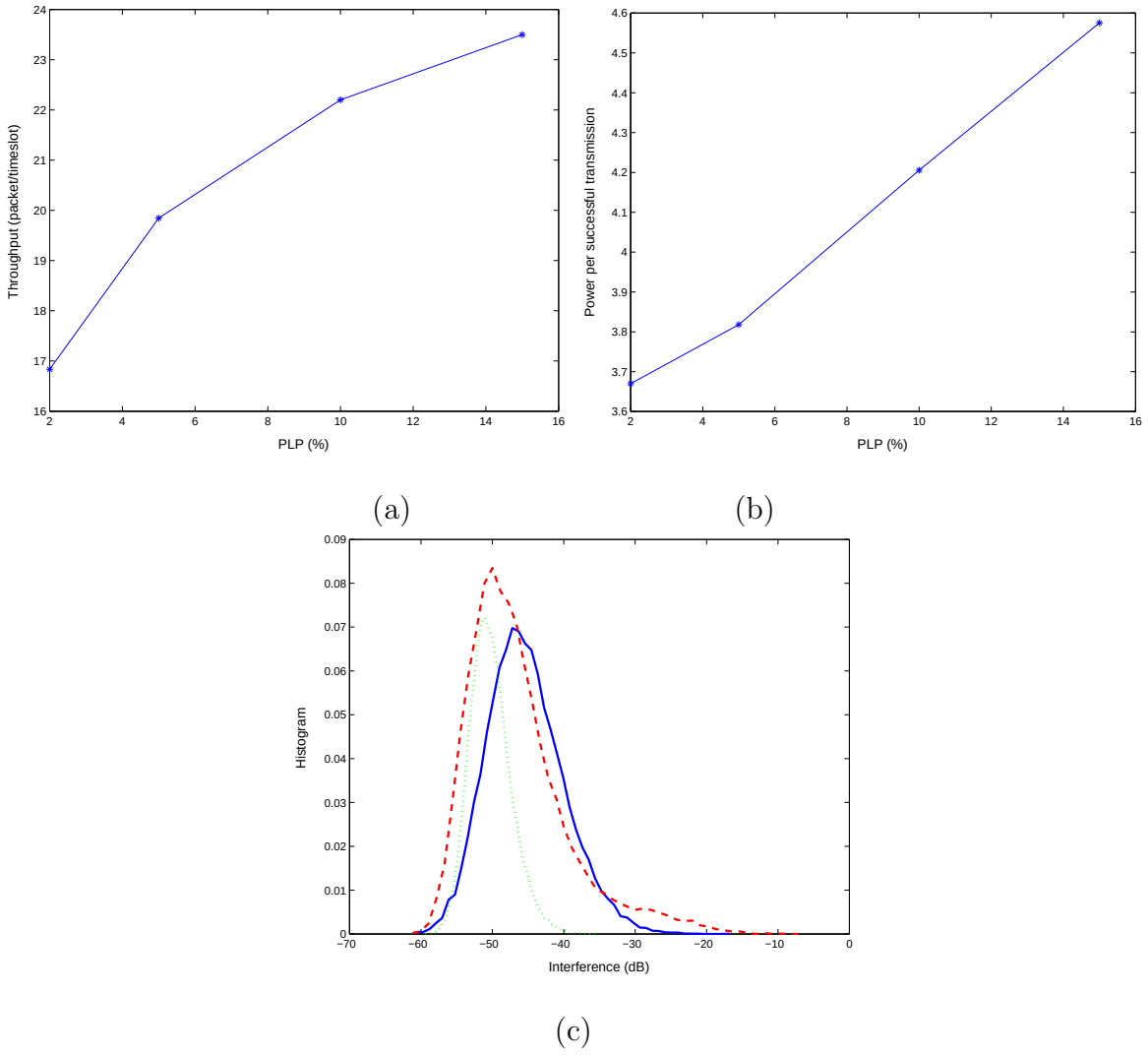


Figure 5.3: Plot of (a) network throughput vs. target PLP, (b) average transmission power per successful transmission vs. target PLP, and (c) histogram of interference at three different nodes.

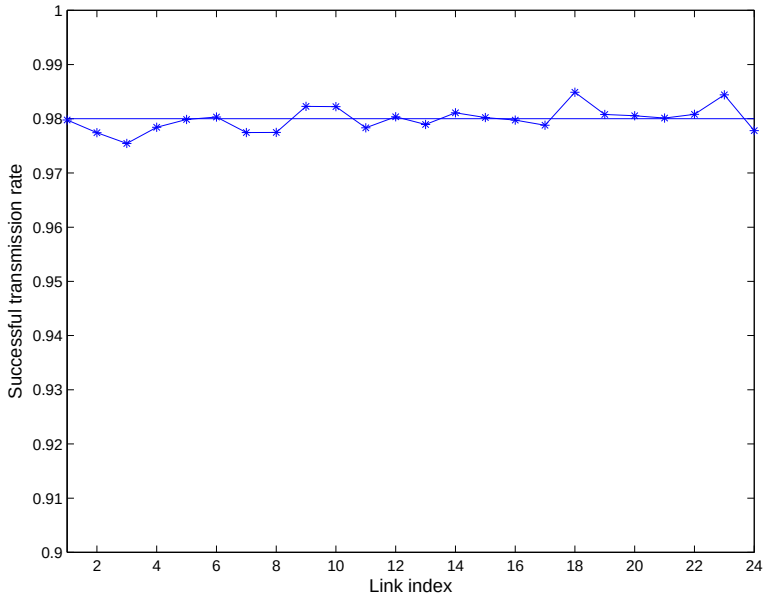
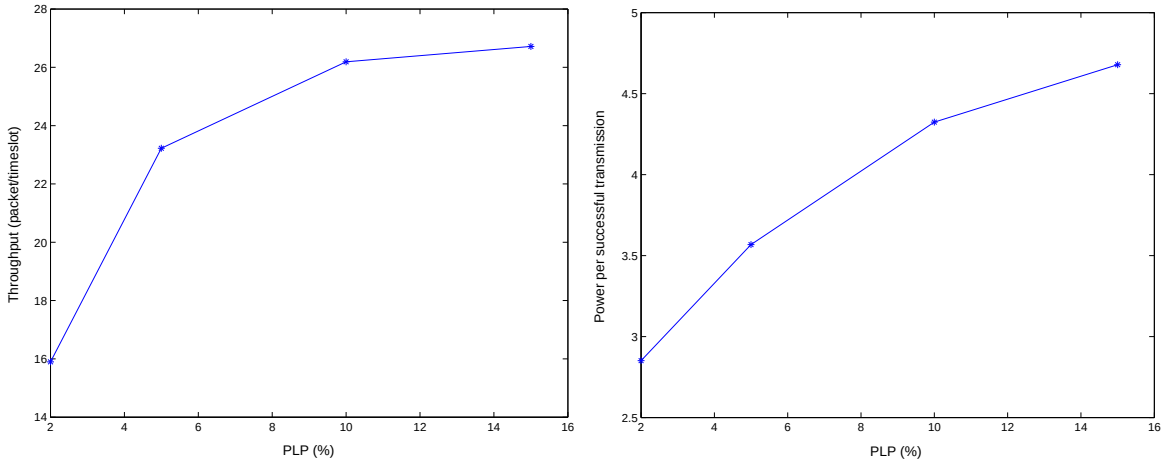


Figure 5.4: Plot of PLP.



(a) (b)

Figure 5.5: Plot of (a) network throughput vs. target PLP, (b) average transmission power per successful transmission vs. target PLP, using a link curve with $\alpha = 2$.

Chapter 6

Scheduling of Real Time Traffic in a Cellular Network

6.1 Introduction

Scheduling in wireline or wireless systems concerns allocation of the shared resource to the users on a per packet basis. Scheduling is challenging in wireless systems, since the volatile link results in error bursts, during which packets cannot be reliably transmitted. Furthermore, channel errors and capacity are location dependent, due to different fading characteristics of the users, while the channel quality varies randomly and asynchronously for the users. Hence, the scheduling decision relies on the channel states as well as the packet flows of all users [65].

Two general approaches for scheduling can be identified in the literature. The first one focuses on fair resource allocation to the users over a link. For wireline networks, weighted fair queuing (WFQ) was proposed in [9] as a packet-by-packet approximation to generalized processor sharing (GPS) [10] for worst-case performance guarantees on throughput and delay. Modified versions of WFQ for wireless

links are presented in [11] [12], where the impact of the wireless link is reflected in lagging and leading flows. A flow is said to be lagging (leading) if its queue length is greater (smaller) than the length of a virtual queue that corresponds to error-free channel. The idea is to allow lagging flows to make up their lag by causing leading flows to give up their lead. In [63], the authors studied opportunistic scheduling algorithms under certain resource allocation constraint.

The second approach deals with optimization of scheduling policies in a wider sense. In [13], the authors investigate the tradeoff between scheduling policies that are optimal in the sense of minimizing buffer or delay requirements. In [14], optimal scheduling without deadline constraints is studied for a wireless system with a number of queues and a single server, where packet arrivals and user channels are both modeled as i.i.d Bernoulli processes. It is shown that the policy that minimizes the total number of packets and delay in the system is the one which serves the longest connected queue (LCQ).

For real-time traffic, each packet has a deadline, beyond which the packet is not useful to the user. The objective of a scheduling policy is to transmit maximum number of packets before their deadlines, or equivalently minimize packet loss due to deadline expirations. In [15] [17], the authors prove that earliest deadline first (EDF) policy is optimal for wireline networks and in [16], a modified version of EDF, the feasible earliest due date (FEDD) policy is proposed for scheduling in wireless systems with deadlines. FEDD policy schedules packets based on EDF over channels that are perceived to be in good state. The authors showed that FEDD is optimal for symmetric systems and a class of deterministic arrival processes, but it is not optimal in general. Thus, to the best of our knowledge, the issue of optimal scheduling for real-time traffic with deadlines over wireless links

has not been hitherto addressed in the literature. Furthermore, the relative impact of user link qualities and packet deadline constraints on the performance of the scheduling strategy has not been precisely defined.

In this thesis, we consider a specific traffic model, constant bit rate (CBR) traffic. We cast the scheduling problem over wireless links as a Markov decision process (MDP) and derive the minimum average long-term packet loss due to deadline expirations. We will identify the tradeoff between scheduling packets of users with better link quality and scheduling packets with smallest residual time.

Even though we study the problem of real-time packet scheduling using CBR traffic, it is obvious that the approaches and discussions apply to other traffic models as well.

This chapter is organized as follows. In Section 6.2 we provide the network model and assumptions and in Section 6.3 we formulate the problem and describe our solution. Numerical results are shown in Section 6.4. Finally, Section 6.5 concludes this chapter.

6.2 System Model

We consider the downlink transmission from the base station to N users. The base station scheduler consists of N queues, one for each user. An underlying slotted scheme is assumed. Equal-length packets arrive at the queues and need to be transmitted over the wireless channel to the users. The duration of each timeslot equals the transmission time of one packet.

Packet arrivals for each queue correspond to a constant bit rate (CBR) traffic. The arrival process for queue i is thus a deterministic periodic process and packets arrive every D_i timeslots. A packet for user i has to be transmitted by the end of

the $\bar{d}_i$th timeslot after arrival. The residual life time of a packet is the difference between its deadline and the current time, and specifies the time until which the packet is useful for the receiver. A packet is dropped from the queue if its residual life time reaches 0. Therefore, the residual life time of a queueing packet for user i is between 1 and $\bar{d}_i$. If the first (head-of-line, HOL) packet of queue i is $d^{(i)}$, then the residual life times of the successive packets in queue i are $d^{(i)} + D_i, d^{(i)} + 2D_i, \ldots \leq \bar{d}_i$.

The queue states of the users at the beginning of timeslot t are represented by the residual life times of the HOL packets, $\mathbf{d}_t = [d_t^{(i)} : i = 1, \ldots, N]$.

Wireless link quality is captured by packet loss probability (PLP) and varies for each user and timeslot, as a result of location-dependency and time-variance of errors. PLP takes values in the L-element set $\mathcal{P} = \{p_1, \ldots, p_L\}$. At timeslot t, user i has channel state $s_t^{(i)} = \ell$, if $p_t^{(i)} = p_\ell$, where $p_t^{(i)}$ is the PLP at user i in timeslot t. Channel conditions in timeslot t are independent for each user and are known to the scheduler. They are described by vector $\mathbf{s}_t = [s_t^{(i)} : i = 1, \ldots, N]$. For each user i, the time-varying channel condition is described by an L-state Markov chain, with transition probabilities $P(s_{t+1}^{(i)} = m | s_t^{(i)} = \ell) = p_{\ell m}$.

Feedback for a transmitted HOL packet is assumed to be available at the end of the corresponding timeslot. If a packet is correctly received, it is removed from the queue. If the packet is not correctly received, it stays in the queue as HOL packet and can be re-transmitted at a future time, provided that the deadline of the packet is not exceeded. In the event of deadline expiration, the packet is discarded from the queue and is considered to be lost.

6.3 Scheduling of CBR traffic with Deadline Constraint in Wireless Networks

The system state is described by a discrete-time Markov chain $\{\mathbf{Y}_t\}_{t=0}^{\infty}$, where $Y_t = (\mathbf{d}_t, \mathbf{s}_t)$ is the system state at the beginning of timeslot t.

The scheduler is informed about HOL packet residual life time and channel condition for each user at the beginning of each timeslot and makes the scheduling decision. Let $u_t \in \{1, \ldots, N\}$ denote the control (decision) variable, indicating the served queue at timeslot t and assume that the scheduler is work conserving. A scheduling policy π is a process $U^{\pi} = u_1^{\pi}, u_2^{\pi}, \ldots$, that includes the decision variables at consecutive timeslots. In this thesis, we focus on the class of stationary scheduling policies Π, for which scheduling decisions are independent of t and depend only on $\mathbf{d}_t$ and $\mathbf{s}_t$.

A first-in-first-out service order is applied, so that the scheduler allocates the HOL packet from a queue to each timeslot for transmission. After each timeslot, the residual life time of each packet in the queue is decremented by one. If the HOL packet with residual life time d_i from the selected queue i is received correctly, the second packet in the queue becomes HOL packet and HOL packet residual life time becomes $d_i + D_i - 1$. If the HOL packet is not received correctly, it remains as HOL packet in the queue and its residual life time is simply decremented by one. Clearly, the residual life times of the HOL packets of unselected queues are also decremented by one at each timeslot. When the residual life time of a HOL packet reaches zero, the packet leaves the queue, regardless of the scheduling decision or successful transmission. It is counted as lost when it is not transmitted or when it is not received correctly. Then, the second packet in queue becomes HOL packet

with residual life time D_i.

After the successful transmission of the HOL packet from queue i, the new HOL packet has residual life time in range $[D_i, \bar{d}_i + D_i - 1]$. Thus, the number of states is $L^N \prod_{k=1}^{N}(\bar{d}_k + D_k - 1)$, and the state space is $\mathcal{Y} = \prod_{k=1}^{N}(\bar{d}_k + D_k - 1) \times \mathcal{P}^N$.

Note that a packet can not be scheduled before its arrival. Therefore $d_i \leq \bar{d}_i$, for $i = 1, \ldots, N$ for a packet of user i to be eligible for transmission.

Let $\mathbf{x}_t^{(k)}$ be the $N \times 1$ vector with k-th component equal to D_k and all other components zero and let $\mathbf{1}$ denote the $N \times 1$ vector of all ones. Furthermore, let $Z_t = \{k : d_t^{(k)} = 1\}$ be the subset of queues with HOL packet residual life times equal to 1 at time t and let its cardinality be $|Z_t|$. The state transitions $\mathbf{Y}_t \to \mathbf{Y}_{t+1}$ depend on current state $\mathbf{Y}_t = (\mathbf{d}_t, \mathbf{s}_t)$ and the decision rule u_t. The channel state transitions $\mathbf{s}_t \to \mathbf{s}_{t+1}$ are determined by the Markov model for the channel. The HOL packet residual life time transitions $\mathbf{d}_t \to \mathbf{d}_{t+1}$ when $u_t = k$ and $k \notin Z_t$ can be succinctly given as follows:

$$\mathbf{d}_{t+1} = \begin{cases} \mathbf{d}_t - \mathbf{1} + \sum_{i \in Z_t} \mathbf{x}_t^{(i)} + \mathbf{x}_t^{(k)}, & \text{w.p. } 1 - p_t^{(k)} \\ \mathbf{d}_t - \mathbf{1} + \sum_{i \in Z_t} \mathbf{x}_t^{(i)}, & \text{w.p. } p_t^{(k)}, \end{cases} \tag{6.1}$$

where $p_t^{(k)}$ depends on $\mathbf{s}_t$. When $k \in Z_t$, we have

$$\mathbf{d}_{t+1} = \mathbf{d}_t - \mathbf{1} + \sum_{i \in Z_t} \mathbf{x}_t^{(i)}, \text{ w.p. } 1. \tag{6.2}$$

The instantaneous cost C_t at timeslot t is determined by the number of discarded packets due to deadline expirations and can be expressed as

$$C_t = \begin{cases} |Z_t|, & \text{if } u_t \in Z_t, \quad \text{w.p. } p_t^{(u_t)} \\ |Z_t| - 1, & \text{if } u_t \in Z_t, \quad \text{w.p. } 1 - p_t^{(u_t)} \\ |Z_t|, & \text{if } u_t \notin Z_t. \end{cases} \tag{6.3}$$

The long-term average cost per timeslot due to deadline expirations for policy

$\pi \in \Pi$ is,

$$C^{\pi} = \lim_{t \to \infty} \frac{1}{t} \mathbf{E}^{\pi} \left[\sum_{\tau=0}^{t-1} C_{\tau}^{\pi} \right] \tag{6.4}$$

where $\mathbf{E}^{\pi}[\cdot]$ denotes expectation with respect to policy π. Therefore, our problem can be rigorously stated as follows:

$$\begin{aligned} &\text{minimize } C^{\pi} \\ &\text{over all stationary scheduling policies } \pi \in \Pi. \end{aligned} \tag{6.5}$$

A policy $\pi^* \in \Pi$ is optimal in the sense of minimizing long-term average cost, if $C^{\pi^*} \leq C^{\pi}$ for any $\pi \in \Pi$.

6.3.1 Solution to the formulated MDP problem

The infinite-horizon MDP problem is solved by using the policy iteration algorithm [6]. Since $\mathbf{Y_t}$ is a unichain Markov chain with finite state space and decision space, and the cost is bounded, an optimal solution is guaranteed to exist.

6.4 Simulation Results

We consider the scheduling problem for $N = 2$ queues, so as to keep complexity at a reasonable level and demonstrate our arguments. Packet inter-arrival time at queue i is D_i timeslots, for $i = 1, 2$. The classical 2-state Gilbert model with a good (g) and a bad (b) state and transition probabilities P_{bg} and P_{gb} is adopted for the wireless channel. The good and bad states are characterized by PLP p_L and p_H respectively, with $p_H > p_L$. Unless otherwise stated, $P_{gb} = 0.01$ and P_{bg} is a variable quantity. By computing the stationary distribution for the Markov chain, we find that the channel is in good and bad state for $P_{bg}/(P_{bg} + P_{gb})$ and

$P_{gb}/(P_{bg} + P_{gb})$ of time on average. We evaluate and compare the performance of the following scheduling policies:

- Markov Decision Process (MDP). Results for this policy are generated by solving the MDP problem (6.5) with the policy iteration algorithm.

- Earliest Deadline First (EDF). This policy selects the queue with the smallest HOL packet residual life time at each timeslot. If HOL packets of both queues have the same residual life time, the user with the best channel (lowest PLP) is selected.

- Best Channel First (BCF). This policy schedules the user with the best channel (lowest PLP) at each timeslot. If users have the same PLP, the queue with the smallest HOL packet residual life time is selected. The BCF policy thus resembles the FEDD policy, which is studied in [16].

The performance metric is the average long-term packet loss ratio (PLR) due to deadline expiration. Results were averaged over 1000 experiments and each experiment included measurements for $n = 10^4$ timeslots. The policy iteration algorithm for MDP converged in 5-6 iterations. For long-term average cost C as in (6.4), PLR $= CD_1D_2/(D_1 + D_2)$, since n/D_i packets arrive at queue i for transmission. First, we consider a system with $p_H = 0.5$, $p_L = 0.05$ and $\bar{d} = 20$. In Figure 6.1, PLR is shown as a function of transition probability P_{bg}, for inter-arrival times denoted by vectors $\mathbf{D} = (2, 3)$ and $(3, 5)$ respectively. MDP approach always provides the lower bound in PLR. The BCF policy performs better than EDF for $\mathbf{D} = (2, 3)$, which corresponds to a scenario of small packet inter-arrival times in each queue and "dense" arrival events between the two queues. According to BCF policy, priority should be given to good channel conditions, rather than deadlines.

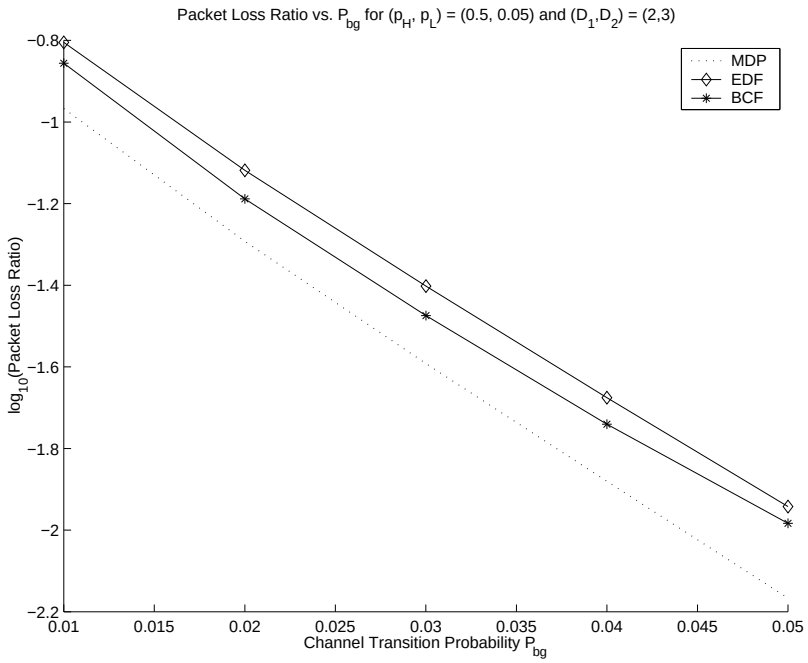
On the other hand, EDF policy performs better than BCF for $\mathbf{D} = (3, 5)$, i.e, for larger inter-arrival times and sparser arrivals between queues. In that case, the scheduler can handle better HOL packet deadlines. It can be seen from Figure 6.1 that EDF performs gradually better than BCF for $D = (3, 5)$ as P_{bg} increases, which implies a channel in good state for more time.

In Figure 6.2, we consider $P_{bg}/P_{gb} = 3$, so that the channel is in bad state for 25% of the time and we study the impact of channel state switching rate on performance. The lower PLR bound is again provided by the MDP policy. However, the relative performance of BCF and EDF policies changes for different ranges of P_{bg}. For $P_{bg} < 0.022$, i.e, for low channel switching rates, EDF policy yields lower PLR. A possible explanation is that deadline expirations are more likely in BCF due to longer periods when the channel is in bad state. On the other hand, BCF yields significantly lower PLR for $P_{bg} > 0.022$. Indeed, when channel switching rate is higher, a queue is more likely to experience good channel state before its HOL packet deadline expires, so that packet will be successfully transmitted, if that queue is selected.

Significant insight can be drawn from these graphs. The MDP policy establishes the lower bound on PLR, since it stems from the solution to problem (6.5). The relative performance of practical EDF and BCF policies depends on traffic load, channel model and channel switching rate. EDF policy performs better for light traffic load and low channel switching rates, whereas BCF is better when traffic load increases and channel state changes rapidly.

6.5 Discussion

In this chapter, we addressed the problem of scheduling CBR traffic subject to deadline constraints, with the objective to reduce packet loss due to deadline expirations. The problem was studied in the context of MDP. Our primary goal is to quantify the relative impact of deadline constraints and channel conditions on the scheduling policy, and draw the guidelines for the design of practical scheduling algorithms.



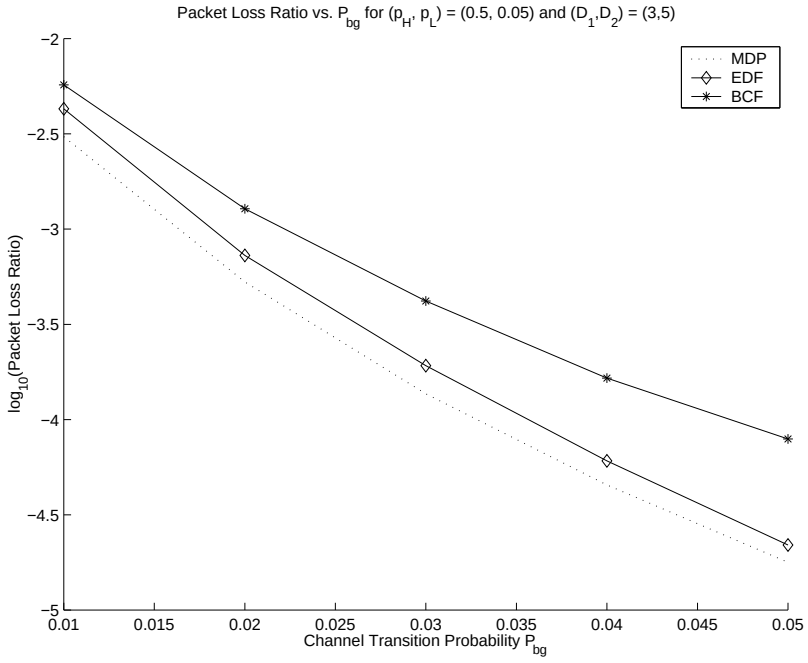


Figure 6.1: PLR vs. P_{bg} for $(p_H, p_L) = (0.5, 0.05)$ and $\bar{d} = 20$, $(D_1, D_2) = (2, 3)$and $(3, 5)$ respectively

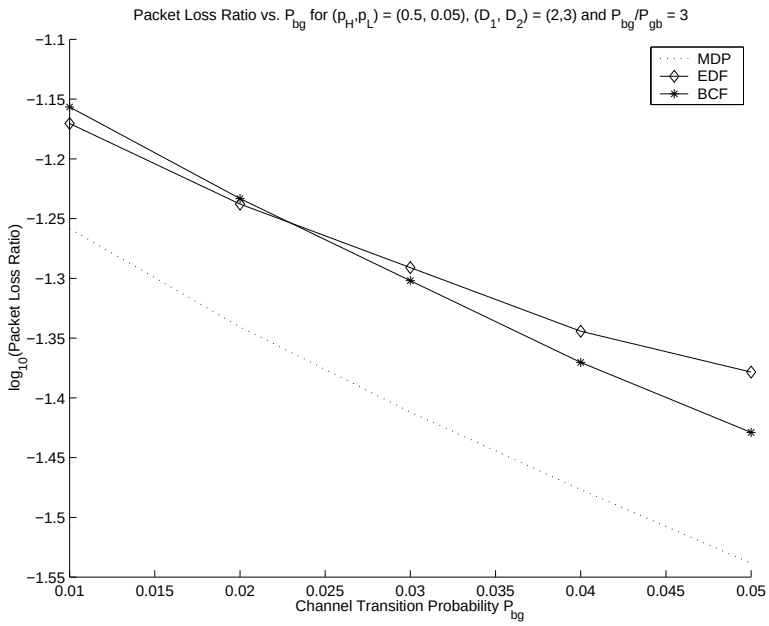


Figure 6.2: PLR vs. P_{bg} for $(p_H, p_L) = (0.5, 0.05)$, $(D_1, D_2) = (2, 3)$ and $\bar{d} = 12$. The ratio $P_{bg}/P_{gb} = 3$ is constant.

Chapter 7

Conclusion and Future Work

In this thesis, several cross-layer resource allocation problems in wireless networks were considered.

Antenna array is extensively studied and considered the last frontier of capacity enhancement. In this thesis, three problems related to the application of antenna arrays at the base stations are considered.

In Chapter 2, we proposed protocols for spatial signature acquisition and evaluated their performance. We assumed that the base station is able to form a single beam at any given time.

An interesting situation arises if the base station is equipped with several transceivers. Then, multiple beams can be formed to scan the space simultaneously towards different directions, so that the time required to locate the users is reduced. A synergy between beams could further improve protocol performance. Furthermore, the possibility of adapting beam width in different stages of the algorithm, depending on the outcome of the contention resolution process is another issue that deserves further investigation.

In Chapter 3, we proposed beamforming algorithms that take random inter-

cell interference into account and achieve target packet loss probability (PLP). It was shown that the amplitude distribution of inter-cell interference can be closely approximated by a log-normal random variable. And the temporal correlation of inter-cell interference is weak.

In this thesis, we assumed each base station knows the channel conditions of the users in its own cell. However, in practice, each base station may only have inaccurate channel information. This inaccuracy in channel condition introduces another source of randomness and needs to be modeled carefully. Robust beamforming algorithms that are able to accommodate this randomness are highly desirable.

In Chapter 4, we considered the joint scheduling and beamforming algorithms with the goal to maintain the stability of the system. We proved an optimal scheduling policy which has exponential complexity in the number of users. Moreover, we proposed two heuristic scheduling policies that achieve sub-optimal performance with significantly lower complexity.

When the system is stable, the traffic demand of each user is satisfied. However, when it is impossible to maintain the stability of the system. The resources (*e.g.,* timeslot, power) need to be allocated in a fair manner. Different notions of fairness are proposed in the literature. We can address the problem in a utility maximization framework that enables the network operator to achieve any working point between fair queueing scheduling and throughput maximizing scheduling.

In Chapter 5, the idea of achieving physical layer QoS in terms of PLP in the presence of random interference is applied to study the ad-hoc networks. We proposed a power control algorithm to achieve target PLP and proved this algorithm minimizes the aggregate transmission power subject to PLP constraint.

The performance of this algorithm and its convergence rate depend on the

estimation frequency. High estimate frequency enables each node to respond to the congestion condition change in a timely manner. However, high update frequency introduces fluctuations. This tradeoff should be investigated thoroughly in the future.

In Chapter 6, we considered the delivery of real-time packets to a number of users by a base station and studied the tradeoff between scheduling the users with good channel condition and scheduling the users with long delayed packets. The problem was modeled as a Markov decision process and the performance of different heuristic scheduling algorithms was studied in different conditions.

It is the ideal model that channel conditions and their transition probabilities are perfectly known. In an environment where channel conditions are hard to estimate accurately or the channel state transitions are unpredictable, the scheduling policy has to rely on inaccurate or out-of-date channel information and thus the system performance is degraded. The study of optimal scheduling policy in such environment is of practical value.

Devising practical scheduling policies with near-optimal performance is another issue that warrants further investigation. To maximize throughput, the base station should serve the user with best channel condition. To match deadline requirement, the packet with the earliest deadline should be served. A simple heuristic solution could be: set a threshold, among all the users whose deadlines are before this threshold, the one with best channel condition is served.

Appendix A

Proofs

A.1 Proof of Proposition 1

In order to prove the proposition we first show that the throughput region achieved by stationary scheduling policies that consider only the channel state $\mathbf{S}$ is given by the throughput region stated in the proposition. Then, we prove that restricting the scheduling policies to such stationary scheduling policies does not reduce the throughput region.

Define

$$\underline{D}_j := \liminf_{t\to\infty} \frac{1}{t}\sum_{\tau=1}^{t} r_j(\tau)(1 - I_j^e(\tau)) = \liminf_{t\to\infty} \frac{1}{t}\sum_{\tau=1}^{t} d_j(\tau) \tag{A.1}$$

where $r_j(t)$ is the transmission rate for user j in timeslot t, $d_j(t) = r_j(t)(1 - I_j^e(t))$, and $I_j^e(t)$ is an indicator function

$$I_j^e(t) = \begin{cases} 1, & \text{if user } j\text{'s transmission in timeslot } t \text{ is unsuccessful} \\ 0, & \text{otherwise} \end{cases} .$$

These indicator functions $\{I_j^e(t); t = 1, 2, \ldots\}$ are given by a sequence of i.i.d. Bernoulli rvs with $\mathcal{E}I_j^e(t) = PLP$. The indicator functions $\{I_j^e(t); t = 1, 2, \ldots\}$,

$j = 1, 2, \ldots, J$, for different users are assumed to be mutually independent.

Lemma A.1.1 *Consider a single queue j, $j = 1, 2, \ldots, J$. A necessary condition for queue j to be stable is $A_j \leq \underline{D}_j$. Moreover, if the departure process $\{d_j(t); t = 1, 2, \ldots\}$ is given by a finite state, ergodic Markov chain, then a sufficient condition for stability is $A_j < \underline{D}_j$.*

Proof It is well known that with Markovian arrival and departure processes a sufficient condition for the queues to be stable is $A_j < \underline{D}_j$ [58], [61]. Hence, here we only prove that $A_j \leq \underline{D}_j$ is a necessary condition.

Suppose that $A_j > \underline{D}_j$. Select $\epsilon > 0$ such that $A_j - \underline{D}_j - 2\epsilon > 0$. We can find a subsequence $\{t_i\}$, where $t_i \to \infty$, such that for all t_i

$$\frac{\sum_{\tau=1}^{t_i} a_j(\tau)}{t_i} \;\geq\; A_j - \epsilon \;\text{ and }\; \frac{\sum_{\tau=1}^{t_i} d_j(\tau)}{t_i} \leq \underline{D}_j + \epsilon \;.$$

Then, it is easy to see that the queue size $x_j(t_i)$ satisfies

$$x_j(t_i) = \sum_{\tau=1}^{t_i} a_j(\tau) - \sum_{\tau=1}^{t_i} d_j(\tau) \geq (A_j - \underline{D}_j - 2\epsilon)t_i \;\text{ for all } t_i \;.$$

Define $\alpha := A_j - \underline{D}_j - 2\epsilon$, and let T_i denote the additional time it takes for the queue size $x_j(t)$ to drop below a threshold value c, starting at the value $x_j(t_i)$ in timeslot t_i. Clearly $T_i \geq (\alpha t_i - c)/v_{max}$, where $v_{max} = \max\ \mathcal{V}$ is the largest transmission rate available. Thus, at time $t_i + T_i$ the fraction of time the queue size exceeds c is lower bounded by $T_i/(t_i + T_i)$, which is greater than or equal to $(\alpha t_i - c)/(\alpha t_i - c + v_{max} t_i)$. Therefore,

$$\begin{aligned} \limsup_{t\to\infty} \frac{1}{t} \sum_{\tau=1}^{t} \mathbf{1}_{[x_j(\tau)>c]} &\geq \lim_{t_i\to\infty} (\alpha t_i - c)/(\alpha t_i - c + v_{max} t_i) \\ &= \alpha/(\alpha + v_{max}) \;. \end{aligned} \tag{A.2}$$

Since (A.2) is true for all $c > 0$,

$$\lim_{c\to\infty} \limsup_{t\to\infty} \frac{1}{t} \sum_{\tau=1}^{t} \mathbf{1}_{[x_j(\tau)>c]} \geq \alpha/(\alpha + v_{max}) > 0 \;,$$

and the system is not stable. ■

Using a stationary scheduling policy that utilizes only the channel state information leads to a departure process $\{d_j(t); t = 1, 2, \ldots\}$ produced by a Markov chain for all queues j, with an average rate given by the right hand side of the condition in (4.2). Therefore, Lemma A.1.1 ensures stability when the arrival vector $\mathbf{A}$ is an interior point of the throughput region in Proposition 1.

We now proceed to prove that the condition in (4.2) is a necessary condition even when the above restriction on the scheduling policy is removed. Suppose that all queues can be stabilized with some scheduling policy. Then, from the proof of Lemma A.1.1 a necessary condition for stability is that $A_j \leq \underline{D}_j$ for all users $j \in \{1, 2, \cdots, J\}$, where $\underline{D}_j$ is defined in (A.1).

Define

$$1(\mathbf{S}, t) = \begin{cases} 1, & \text{if channel is in state } \mathbf{S} \text{ in timeslot } t \\ 0, & \text{otherwise} \end{cases}$$

and

$$1(\mathbf{SR}, t) = \begin{cases} 1, & \text{if channel is in state } \mathbf{S} \text{ and a rate matrix } \mathbf{R} \text{ is selected in timeslot } t \\ 0, & \text{otherwise} \end{cases} .$$

Fix $\epsilon > 0$. There exists sufficiently large $\widetilde{t}$ such that, for all $\mathbf{S} \in \mathcal{S}$ and $\mathbf{R} \in \mathbf{S}$,

$$\frac{\sum_{\tau=1}^{\widetilde{t}} 1(\mathbf{S},\tau)}{\widetilde{t}} \leq \pi_{\mathbf{S}} + \epsilon \ , \ \underline{D}_j \leq \frac{1}{\widetilde{t}} \sum_{\tau=1}^{\widetilde{t}} d_j(\tau) + \epsilon \ , \quad \text{and}$$
$$\frac{\sum_{\tau=1}^{\widetilde{t}} 1(\mathbf{SR},\tau)(1-I_j^e(\tau))}{\sum_{\tau=1}^{\widetilde{t}} 1(\mathbf{SR},\tau)} \leq 1 - PLP + \epsilon \ \text{ for all } j = 1, 2, \ldots, J \quad .$$

Therefore, for each $j \in \{1, 2, \ldots, J\}$ we have

$$A_j \leq \underline{D}_j = \frac{1}{\widetilde{t}} \sum_{\tau=1}^{\widetilde{t}} r_j(\mathbf{R}(\tau))(1 - I_j^e(\tau)) + \epsilon$$

$$
\begin{aligned}
&= \sum_{\mathbf{S}\in\mathcal{S}} \frac{\sum_{\tau=1}^{\widetilde{t}} 1(\mathbf{S},\tau)}{\widetilde{t}} \frac{1}{\sum_{\tau=1}^{\widetilde{t}} 1(\mathbf{S},\tau)} \sum_{\tau=1}^{\widetilde{t}} 1(\mathbf{S},\tau) r_j(\mathbf{R}(\tau))(1-I_j^e(\tau)) + \epsilon \\
&\leq \sum_{\mathbf{S}\in\mathcal{S}} (\pi_{\mathbf{S}}+\epsilon) \frac{1}{\sum_{\tau=1}^{\widetilde{t}} 1(\mathbf{S},\tau)} \sum_{\mathbf{R}\in\mathbf{S}} \sum_{\tau=1}^{\widetilde{t}} 1(\mathbf{SR},\tau) r_j(\mathbf{R})(1-I_j^e(\tau)) + \epsilon \\
&= \sum_{\mathbf{S}\in\mathcal{S}} (\pi_{\mathbf{S}}+\epsilon) \sum_{\mathbf{R}\in\mathbf{S}} \frac{\sum_{\tau=1}^{\widetilde{t}} 1(\mathbf{SR},\tau) r_j(\mathbf{R})(1-I_j^e(\tau))}{\sum_{\tau=1}^{\widetilde{t}} 1(\mathbf{S},\tau)} + \epsilon \\
&= \sum_{\mathbf{S}\in\mathcal{S}} (\pi_{\mathbf{S}}+\epsilon) \sum_{\mathbf{R}\in\mathbf{S}} \frac{\sum_{\tau=1}^{\widetilde{t}} 1(\mathbf{SR},\tau)}{\sum_{\tau=1}^{\widetilde{t}} 1(\mathbf{S},\tau)} \frac{\sum_{\tau=1}^{\widetilde{t}} 1(\mathbf{SR},\tau) r_j(\mathbf{R})(1-I_j^e(\tau))}{\sum_{\tau=1}^{\widetilde{t}} 1(\mathbf{SR},\tau)} + \epsilon \ ,
\end{aligned}
$$

where $r_j(\mathbf{R})$ is the j-th element of the vector $\mathbf{R}^T\mathbf{1}_{I\times 1}$.

Define

$$
C_{\mathbf{SR}}(\widetilde{t}) = \frac{\sum_{\tau=1}^{\widetilde{t}} 1(\mathbf{SR},\tau)}{\sum_{\tau=1}^{\widetilde{t}} 1(\mathbf{S},\tau)} \ .
$$

Then,

$$
\begin{aligned}
A_j &\leq \sum_{\mathbf{S}\in\mathcal{S}} (\pi_{\mathbf{S}}+\epsilon) \sum_{\mathbf{R}\in\mathbf{S}} C_{\mathbf{SR}}(\widetilde{t}) r_j(\mathbf{R}) \frac{\sum_{\tau=1}^{\widetilde{t}} 1(\mathbf{SR},\tau)(1-I_j^e(\tau))}{\sum_{\tau=1}^{\widetilde{t}} 1(\mathbf{SR},\tau)} + \epsilon \\
&\leq \sum_{\mathbf{S}\in\mathcal{S}} (\pi_{\mathbf{S}}+\epsilon) \sum_{\mathbf{R}\in\mathbf{S}} C_{\mathbf{SR}}(\widetilde{t}) r_j(\mathbf{R})(1-PLP+\epsilon) + \epsilon \\
&\leq \sum_{\mathbf{S}\in\mathcal{S}} \pi_{\mathbf{S}} \sum_{\mathbf{R}\in\mathbf{S}} C_{\mathbf{SR}}(\widetilde{t}) r_j(\mathbf{R})(1-PLP) + \epsilon + \epsilon v_{max} + \epsilon|\mathcal{S}|v_{max}(1-PLP+\epsilon) \\
&= \sum_{\mathbf{S}\in\mathcal{S}} \pi_{\mathbf{S}} \sum_{\mathbf{R}\in\mathbf{S}} C_{\mathbf{SR}}(\widetilde{t}) r_j(\mathbf{R})(1-PLP) + \epsilon(1+v_{max}+|\mathcal{S}|v_{max}(1-PLP+\epsilon)) \ .
\end{aligned}
$$

Since ϵ can be arbitrarily close to zero, this completes the proof.

A.2 Proof of Proposition 2

The evolution of the backlog vector $\mathbf{X}(t)$ is given by the following recursive equation:

$$
x_j(t+1) = \max\{x_j(t) + a_j(t) - r_j(\mathbf{R}(t))(1-I_j^e(t)),\ 0\}
$$

Define $d_j(t) = r_j(\mathbf{R}(t))(1 - I_j^e(t))$. It is clear that Property 1 in Theorem 1 holds. Now we prove Property 2 of the theorem.

$$\begin{aligned} x_j^2(t+1) &\leq (x_j(t) + a_j(t) - d_j(t))^2 \\ &\leq x_j(t)^2 - 2x_j(t)d_j(t) + 2x_j(t)a_j(t) + d_j(t)^2 + a_j(t)^2 \end{aligned}$$

Using the above inequality

$$\begin{aligned} &\mathbf{E}[L(\mathbf{X}(t+1)) - L(\mathbf{X}(t))|\mathbf{X}(t)] \\ &\leq \sum_{j=1}^{J} \mathbf{E}[a_j(t)^2|\mathbf{X}(t)] + \sum_{j=1}^{J} \mathbf{E}[d_j(t)^2|\mathbf{X}(t)] - 2\sum_{j=1}^{J} x_j(t)\mathbf{E}[d_j(t) - a_j(t)|\mathbf{X}(t)] \\ &\leq B - 2\sum_{j=1}^{J} x_j(t)(\mathbf{E}[d_j(t)|\mathbf{X}(t)] - A_j) \end{aligned}$$

where $B := \sum_{j=1}^{J} \mathbf{E}[a_j(t)^2] + J \cdot v_{max}^2$ because $\sum_{j=1}^{J} \mathbf{E}[d_j(t)^2|\mathbf{X}(\mathbf{t})] \leq J \cdot v_{max}^2$.

Since $\mathbf{A}$ lies in $int(\mathcal{A})$, we have

$$\begin{aligned} \sum_{j=1}^{J} x_j(t)A_j &\leq \sum_{j=1}^{J} x_j(t)D_j \\ &= \sum_{\mathbf{S}\in\mathcal{S}} \pi_{\mathbf{S}} \sum_{\mathbf{R}\in\mathbf{S}} c_{\mathbf{SR}} \sum_{j=1}^{J} x_j(t)r_j(\mathbf{R})(1 - PLP) \\ &\leq \sum_{\mathbf{S}\in\mathcal{S}} \pi_{\mathbf{S}} \max_{\mathbf{R}\in\mathbf{S}} \sum_{j=1}^{J} x_j(t)r_j(\mathbf{R})(1 - PLP) \\ &= \sum_{j=1}^{J} x_j(t)\mathbf{E}[d_j(t)|\mathbf{X}(t)] \end{aligned}$$

where D_j is the j-th element of $\mathbf{D}$ with a scheduling policy that satisfies the inequality in (4.2).

Since $\mathbf{A} \in int(\mathcal{A})$, we can find a $J \times 1$ vector $\varepsilon = (\epsilon, \cdots, \epsilon)^T$ such that $\mathbf{A} + \varepsilon$ belongs to $int(\mathcal{A})$ and satisfies

$$\sum_{j=1}^{J} x_j(t)(A_j + \epsilon) \leq \sum_{j=1}^{J} x_j(t)\mathbf{E}[d_j(t)|\mathbf{X}(t)] \ .$$

Therefore

$$\sum_{j=1}^{J} x_j(t)(\mathbf{E}[d_j(t)|\mathbf{X}(t)] - A_j) = \sum_{j=1}^{J} x_j(t)(\mathbf{E}[d_j(t)|\mathbf{X}(t)] - (A_j + \varepsilon) + \varepsilon)$$
$$\geq \epsilon \sum_{j=1}^{J} x_j(t)$$

and

$$E[L(\mathbf{X}(t+1)) - L(\mathbf{X}(t))|\mathbf{X}(t)] \leq B - 2\epsilon \sum_{j=1}^{J} x_j(t) \ .$$

For any positive α, we can define a compact region

$$\Sigma_\alpha = \left\{ \mathbf{X}(t) \in \mathcal{R}^J \;\middle|\; \sum_{j=1}^{J} x_j(t) \leq \frac{B+\alpha}{2\epsilon} \right\} \ .$$

It is an easy exercise to show that whenever $\mathbf{X}(t) \in \mathcal{R}^J \setminus \Sigma_\alpha$, we have $\mathbf{E}[L(\mathbf{X}(t+1)) - L(\mathbf{X}(t))|\mathbf{X}(t)] \leq -\alpha$, hence satisfying the second condition in Theorem 4.2.2. Therefore, by Theorem 4.2.2 the system is stable under the scheduling policy given by (4.3).

A.3 Proof of Lemma 5.5.1

First, recall from (5.11) that at a fixed point $\mathbf{p}$, we have

$$\mathbf{p}_l = \left(e^{kz} \frac{\mathbf{E}\left[(\text{Interference}_l(\mathbf{p}))^\alpha\right]}{\gamma_l \cdot G_l^\alpha} \right)^{1/\alpha} \ .$$

Hence, we have

$$G_l = \left(\frac{e^{kz}}{\gamma_l} \right)^{1/\alpha} \frac{\left(\mathbf{E}\left[(\text{Interference}_l(\mathbf{p}))^\alpha\right]\right)^{1/\alpha}}{\mathbf{p}_l} \ . \tag{A.3}$$

We prove the lemma by contradiction: Assume that there are more than one fixed point of $F(\cdot)$. Let $\mathbf{p}^1$ and $\mathbf{p}^2$, $\mathbf{p} \neq \mathbf{p}^2$, be two fixed points. Recall that, given

a transmission power vector,

$$\begin{aligned}&\mathbf{E}\left[(\text{Interference}_l(\mathbf{p}))^\alpha\right]\\&=\sum_{\underline{s}\in\mathcal{S}:\underline{s}_l=1}\mathbf{d}^l_{\underline{s}}\left(\mathbf{p}^T\mathbf{G}^l\underline{s}-G_l\cdot\mathbf{p}_l+n_l\right)^\alpha\end{aligned}$$

Suppose that the links are ordered by decreasing ratio $\mathbf{p}^1_l/\mathbf{p}^2_l$. Without loss of generality we assume that $\eta_1=\mathbf{p}^1_1/\mathbf{p}^2_1>1$. Then, from (A.3) we know that

$$\begin{aligned}&\frac{(\mathbf{E}\left[(\text{Interference}_1(\mathbf{p}^1))^\alpha\right])^{1/\alpha}}{\mathbf{p}^1_1}\\&=\frac{(\mathbf{E}\left[(\text{Interference}_1(\mathbf{p}^2))^\alpha\right])^{1/\alpha}}{\mathbf{p}^2_1}\ .\end{aligned}\tag{A.4}$$

However, since $\mathbf{p}^1_l\le\eta_1\cdot\mathbf{p}^2_l$ for all $l\in\widetilde{\mathcal{L}}$,

$$\begin{aligned}\mathbf{E}\left[(\text{Interference}_1(\mathbf{p}^1))^\alpha\right]&\le\mathbf{E}\left[(\text{Interference}_1(\eta_1\cdot\mathbf{p}^2))^\alpha\right]\\&<\eta_1^\alpha\cdot\mathbf{E}\left[(\text{Interference}_1(\mathbf{p}^2))^\alpha\right]\ .\end{aligned}$$

Therefore, this implies that

$$\begin{aligned}&\frac{(\mathbf{E}\left[(\text{Interference}_1(\mathbf{p}^1))^\alpha\right])^{1/\alpha}}{\mathbf{p}^1_1}\\&<\frac{(\eta_1^\alpha\cdot\mathbf{E}\left[(\text{Interference}_1(\mathbf{p}^2))^\alpha\right])^{1/\alpha}}{\mathbf{p}^1_1}\\&=\frac{\eta_1(\mathbf{E}\left[(\text{Interference}_1(\mathbf{p}^2))^\alpha\right])^{1/\alpha}}{\eta_1\cdot\mathbf{p}^2_1}\\&=\frac{(\mathbf{E}\left[(\text{Interference}_1(\mathbf{p}^2))^\alpha\right])^{1/\alpha}}{\mathbf{p}^2_1}\ ,\end{aligned}$$

which contradicts (A.4).

A.4 Proof of Lemma 5.5.2

Let

$$V(n)=\max_{l\in\widetilde{\mathcal{L}}}\frac{|\mathbf{p}_l(n)-\mathbf{p}^\star_l|}{\mathbf{p}^\star_l}\ ,$$

where $\mathbf{p}^\star$ is the unique fixed point of the mapping F. In order to prove the convergence, it suffices to show that $\lim_{n\to\infty} V(n) = 0$.

To show that $V(n) \downarrow 0$ as $n \to \infty$, we first show that $V(n+1) < V(n)$ if $V(n) > 0$, and then $\lim_{n\to\infty} V(n) = V^*$ necessarily implies that $V^* = 0$.

First, we show that

$$\frac{|\mathbf{p}_l(n+1) - \mathbf{p}_l^\star|}{\mathbf{p}_l^\star} < V(n) \qquad \text{for all } l \in \widetilde{\mathcal{L}}$$

as follows. Here we do not explicitly consider the minimum or maximum power constraint as it does not affect the convergence as will be obvious in the proof.

$$\begin{aligned}
&\frac{|\mathbf{p}_l(n+1) - \mathbf{p}_l^\star|}{\mathbf{p}_l^\star} \\
&= \frac{\left| \left(\sum_{\underline{s}\in\mathcal{S}:\underline{s}_l=1} \mathbf{d}_{\underline{s}}^l \left(\mathbf{p}(n)^T \mathbf{G}^l \underline{s} - G_l \cdot \mathbf{p}_l(n) + n_l\right)^\alpha\right)^{1/\alpha} - \left(\sum_{\underline{s}\in\mathcal{S}:\underline{s}_l=1} \mathbf{d}_{\underline{s}}^l \left(\mathbf{p}^{\star T} \mathbf{G}^l \underline{s} - G_l \cdot \mathbf{p}_l^\star + n_l\right)^\alpha\right)^{1/\alpha} \right|}{\left(\sum_{\underline{s}\in\mathcal{S}:\underline{s}_l=1} \mathbf{d}_{\underline{s}}^l \left(\mathbf{p}^{\star T} \mathbf{G}^l \underline{s} - G_l \cdot \mathbf{p}_l^\star + n_l\right)^\alpha\right)^{1/\alpha}}
\end{aligned}$$

Let

$$\Lambda_l = \left(\sum_{\underline{s}\in\mathcal{S}:\underline{s}_l=1} \mathbf{d}_{\underline{s}}^l \left(\mathbf{p}^{\star T} \mathbf{G}^l \underline{s} - G_l \cdot \mathbf{p}_l^\star + n_l\right)^\alpha\right)^{1/\alpha} .$$

First, we upper bound the numerator as follows.

case (i): $\mathbf{p}_l(n+1) \geq \mathbf{p}_l^\star$.

$$\begin{aligned}
&\Big(\sum_{\underline{s}\in\mathcal{S}:\underline{s}_l=1} \mathbf{d}_{\underline{s}}^l \left(\mathbf{p}(n)^T \mathbf{G}^l \underline{s} - G_l \cdot \mathbf{p}_l(n) + n_l\right)^\alpha\Big)^{1/\alpha} \\
&\leq \Big(\sum_{\underline{s}\in\mathcal{S}:\underline{s}_l=1} \mathbf{d}_{\underline{s}}^l ((1+V(n)) \cdot (\mathbf{p}^{\star T} \mathbf{G}^l \underline{s} - G_l \cdot \mathbf{p}_l^\star) \\
&\qquad\qquad + n_l)^\alpha\Big)^{1/\alpha} \\
&< \Big((1+V(n))^\alpha \times \\
&\qquad \sum_{\underline{s}\in\mathcal{S}:\underline{s}_l=1} \mathbf{d}_{\underline{s}}^l ((\mathbf{p}^{\star T} \mathbf{G}^l \underline{s} - G_l \cdot \mathbf{p}_l^\star) + n_l)^\alpha\Big)^{1/\alpha} \\
&= (1+V(n)) \cdot \Lambda_l \ .
\end{aligned}$$

Hence, the numerator is upper bounded by

$$(1 + V(n)) \cdot \Lambda_l - \Lambda_l = V(n) \cdot \Lambda_l \ .$$

case (ii): $\mathbf{p}_l(n+1) \leq \mathbf{p}_l^\star$.

$$\begin{aligned}
&\Big(\sum_{\underline{s} \in \mathcal{S}: \underline{s}_l = 1} \mathbf{d}_{\underline{s}}^l \left(\mathbf{p}(n)^T \mathbf{G}^l \underline{s} - G_l \cdot \mathbf{p}_l(n) + n_l \right)^\alpha \Big)^{1/\alpha} \\
&\geq \Big(\sum_{\underline{s} \in \mathcal{S}: \underline{s}_l = 1} \mathbf{d}_{\underline{s}}^l ((1 - V(n)) \cdot (\mathbf{p}^{\star T} \mathbf{G}^l \underline{s} - G_l \cdot \mathbf{p}_l^\star) \\
&\qquad\qquad + n_l)^\alpha \Big)^{1/\alpha} \\
&> \Big((1 - V(n))^\alpha \times \\
&\qquad \sum_{\underline{s} \in \mathcal{S}: \underline{s}_l = 1} \mathbf{d}_{\underline{s}}^l ((\mathbf{p}^{\star T} \mathbf{G}^l \underline{s} - G_l \cdot \mathbf{p}_l^\star) + n_l)^\alpha \Big)^{1/\alpha} \\
&= (1 - V(n)) \cdot \Lambda_l \ .
\end{aligned}$$

Therefore, the numerator is upper bounded by

$$\Lambda_l - (1 - V(n)) \cdot \Lambda_l = V(n) \cdot \Lambda_l \ .$$

From the above two cases we have

$$\frac{|\mathbf{p}_l(n+1) - \mathbf{p}_l^\star|}{\mathbf{p}_l^\star} < \frac{V(n) \cdot \Lambda_l}{\Lambda_l} = V(n) \ \text{ for all } l \in \widetilde{\mathcal{L}} \ .$$

The second part of the proof that $V^* = 0$ follows directly from the continuity of the mapping F, and this completes the proof of the lemma.

A.5 Proof of Lemma 5.5.3

We first show that there is a sequence of nonempty sets $\mathcal{P}(n)$ with

$$\cdots \subset \mathcal{P}(n+1) \subset \mathcal{P}(n) \subset \cdots \mathcal{P}(0) \subset \mathcal{P}$$

satisfying the following two conditions.

1. Synchronous convergence condition: we have

$$F(\mathbf{p}) \in \mathcal{P}(n+1) \quad \text{for all } n \text{ and } \mathbf{p} \in \mathcal{P}(n) \ .$$

Furthermore, if $\{\mathbf{p}^k\}$ is a sequence such that $\mathbf{p}^k \in \mathcal{P}(k)$ for every k, then every limit point of $\{\mathbf{p}^k\}$ is the unique fixed point of the mapping F.

2. Box condition: for every n, there exists sets $\mathcal{P}_l(n), l \in \widetilde{\mathcal{L}}$, such that

$$\mathcal{P}(n) = \prod_{l \in \widetilde{\mathcal{L}}} \mathcal{P}_l(n) \ .$$

Let $\mathcal{P}_l(0) = \mathcal{P}$. Define for all $n \geq 1$, $\mathcal{P}^{'}(n) = \{F(\mathbf{p}) | \mathbf{p} \in \mathcal{P}(n-1)\}$. Take the projection for each $l \in \widetilde{\mathcal{L}}$

$$\mathcal{P}_l(n) = \{\mathbf{p}_l \mid \mathbf{p}_l \text{ is the } l\text{-th element of some } \mathbf{p} \in \mathcal{P}_l^{'}(n)\}$$

and define

$$\mathcal{P}(n) = \prod_{l \in \widetilde{\mathcal{L}}} \mathcal{P}_l(n) \ .$$

Then, it is plain $\mathcal{P}_l(n) \subset \mathcal{P}_l(n-1)$ and, hence, $\mathcal{P}(n) \subset \mathcal{P}(n-1)$. Furthermore, one can show that $\mathcal{P}(n)$ satisfies the *synchronous convergence condition*. From its construction $\mathcal{P}(n)$ satisfies the *box condition*. Now the lemma follows from [43].

BIBLIOGRAPHY

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

[2] X. Qiu and K. Chawla, "On the performance of adaptive modulation in cellular systems," *IEEE Transactions on Communications,* vol.47, no.6, pp.884-895, June, 1999.

[3] B. Vucetic, "An adaptive coding scheme for time-varying channels," *IEEE Transactions on Communications,* vol.39, no.5, pp.653-663, May, 1991.

[4] M. Andrews, K. Kumaran, K. Ramanan, A. Stolyar and P. Whiting, "Providing quality of service over a shared wireless link," *IEEE Communications Magazine,* vol.39, no.2, pp.150-154, Feburary 2001.

[5] J. Chuang, X. Qiu and J. Whitehead, "Data throughput enhancement in wireless packet systems by improved link adaptation with application to the EDGE system," *IEEE VTC*, Amsterdam, Netherlands, September, 1999.

[6] M. L. Puterman, *Markov decision processes: discrete stochastic dynamic programming*, Wiley, 1994.

[7] T. Ren, I. Koutsopoulos and L. Tassiulas, "QoS provisioning for real time traffic in wireless packet networks," *IEEE GLOBECOM*, Taipei, Taiwan, November 2002

[8] P. Viswanath, D. Tse and R. Laroia, "Opportunistic beamforming using dumb antennas," *IEEE Transactions on Information Theory,* vol.48, no.5, pp.1277-1294, June, 2002.

[9] A.K. Parekh and R.G. Gallager, "A Generalized Processor Sharing Approach to Flow control in Integrated Services Networks: The Single-node case," *IEEE/ACM Trans. Networking,* vol.1, no.3, pp.344-357, June 1993.

[10] A. Demers, S. Keshav and S. Shenker, "Analysis and simulation of a Fair Queueing Algorithm," *Internetworking: Research and Experience,* vol.1, pp.3-26, 1990.

[11] T. Eugene Ng, I Stoica and H. Zhang, "Packet fair queueing algorithms for wireless networks with location dependent errors," *IEEE INFOCOM*, San Francisco, CA 1998

[12] S. Lu, V. Bharghavan and R. Srikant, "Fair Scheduling in wireless packet networks," *IEEE/ACM Transactions on Networking,* vol.7, no.4, pp.473-489, Aug. 1999.

[13] L. Georgiadis, R. Guerin and A. Parekh, "Optimal Multiplexing on a single link: Delay and Buffer requirements," *IEEE Transactions on Information Theory,* vol.43, no.5, Sept. 1997.

[14] L. Tassiulas and A. Ephremides, "Dynamic server allocation to parallel queues with randomly varying connectivity," *IEEE Trans. Inf. Theory,* vol.39, no.2, pp.466-478, March 1993.

[15] S. Panwar, D. Towsley and J. Wolf, "Optimal scheduling policies for a class of queue with customer deadlines to the beginning of services," *Journal of ACM,* vol.35, no.4, pp.832-844, 1988.

[16] S. Shakkottai and R. Srikant, "Scheduling real-time traffic with deadlines over a wireless channel," *Proc.* 2^{nd} *ACM Int. Workshop on Wireless Mobile Multimedia,* pp.35-42, 1999.

[17] P. Bhattacharya and A. Ephremides, "Optimal scheduling of the transmission of messages with strict deadlines," *Conference on Information Sciences and Systems,* Princeton, 1988

[18] A. Acampora and S. Krishnamurthy, "A new adaptive MAC protocol for broadband wireless networks in harsh fading and interference environments," *IEEE/ACM Transactions on Networking,* vol.8, no.3, pp.328-336, June 2000.

[19] S. Krishnamurthy, A. Acampora, and M. Zorzi, "Polling based medium access control protocols for use with smart adaptive array antennas," *IEEE/ACM Transactions on Networking,* vol.9, no.2, pp.148-161, April 2001.

[20] K. Sheikh, D. Gesbert, D. Gore and A. Paulraj, "Smart antennas for broadband wireless access networks," *IEEE Communications Magazine*, vol.37, no.11, pp.100-105, November 1999.

[21] J. Ward and R.T. Compton, "Improving the performance of a slotted Aloha packet radio network with an adaptive array," *IEEE Transactions on Communications* vol.40, no.2, pp.292-300, Feburary 1992.

[22] J. Ward and R.T. Compton, "High throughput slotted Aloha packet radio networks with adaptive arrays," *IEEE Transactions on Communications,* vol.41, no.3, pp.460-470, March 1993.

[23] C. Sakr and T.D. Todd, "Carrier-sense protocols for packet-switched smart antenna basestations," *IEEE Int. Conf. Network Protocols,* pp.45-52, October 1997.

[24] A. Acampora, "Wireless ATM: a perspective on issues and prospects," *IEEE Personal Communications,*, vol.3, no.4, pp.8-17, August 1996.

[25] ANSI/IEEE Std 802.11, Local and metropolitan area networks - specific requirements, "Part 11: wireless LAN medium access control (MAC) and physical layer (PHY) specifications," 1999 Edition.

[26] R. Agrawal, A. Bedekar, R. J. La, and V. G. Subramanian, "C3WPF scheduler," *Proc. of 17th ITC*, December 2001.

[27] M. Bengtsson, "Jointly optimal downlink beamforming and base station assignment," *IEEE International Conference on Acoustics, Speech, and Signal Processing,* Salt Lake City, Ut, May 2001.

[28] J. Chuang, "Improvement of data throughput in wireless packet systems with link adaptation and efficient frequency reuse," *IEEE Vehicular Technology Conference,* Houston, Texas, May 1999.

[29] C. Farsakh and J. Nossek, "Spatial covariance based downlink beamforming in an SDMA mobile radio system," *IEEE Transactions on Communications,* vol.46, no.11, pp.1497-1506, November 1998.

[30] W. Janos, "Tail of the distribution of sums of log-normal variates," *IEEE Transactions of Information Theory,* vol.16, no.3, pp. 299-302, May 1970

[31] I. Koutsopoulos, T. Ren and L. Tassiulas, "The impact of space division multiplexing on resource allocation: a unified approach," *IEEE INFOCOM*, San Francisco, CA, April 2003.

[32] T. V. Lakshman and U. Madhow, "The performance of TCP/IP for networks with high bandwidth-delay products and random loss," *IEEE/ACM Transactions on Networking*, vol.5, no.3, pp. 336-350, June 1997.

[33] K. Leung, P. Driessen, K. Chawla and X. Qiu, "Link adaptation and power control for streaming services in EGPRS wireless networks," *IEEE Journal on Selected Areas in Communications,* vol.19, no.10, pp. 2029-2039, October 2001.

[34] K. Leung, "Power control by interference predictions for wireless packet networks," *IEEE Transactions on Wireless Communications,* vol.1, no.2, pp. 256 - 265, April 2002.

[35] J. Liberti, Jr., and T. Rappaport, *Smart antennas for wireless communications: IS-95 and third generation CDMA applications,* Prentice Hall, 1999.

[36] M. Rajih and S. Sarkar, "Reference link level curves for Qualcomm cdma2000 Revision D R-ESCH," *ftp://ftp.3gpp2.org,* May 2003.

[37] T. Rappaport, *Wireless communications: principles and practice,* Prentice Hall, 2002.

[38] F. Rashid-Farrokhi, K. R. Liu and L. Tassiulas, "Transmit beamforming and power control for cellular wireless systems," *IEEE Journal on Selected Areas in Communications,* vol.16, no.10, pp.1437-1450, October 1998.

[39] M. Schubert and H. Boche, "Solution of the multiuser downlink beamforming problem with individual SINR constraints," *IEEE Transactions on Vehicular Technology,* vol.53, no.1, pp.18-28, January 2004.

[40] F. Shad, T.D. Todd, V. Kezys and J. Litva, "Dynamic slot allocation (DSA) in indoor SDMA/TDMA using a smart antenna basestation," *IEEE/ACM Transactions on Networking,* vol.9, no.1, pp.69-81, February 2001.

[41] W. Yang and G. Xu, "Optimal downlink power assignment for smart antenna systems," *IEEE International Conference on Acoustics, Speech, and Signal Processing,* Seattle, Washington, May 1998.

[42] P. Dubey, "Inefficiency of Nash equilibria," *Mathematics of Operations Research*, vol. 11, pp. 1-8, 1986.

[43] D. P. Bertsekas and J. N. Tsitsiklis, *Parallel and distributed computation: numerical methods*, Prentice Hall Professional Technical Reference, 1988

[44] P. Gupta and P. Kumar, "The capacity of wireless networks," *i.e.,* IEEE Transactions on Information Theory, vol. 46, no. 2, pp.388-404, March 2000.

[45] S. Ramanathan and E. Lloyd, "Scheduling algorithms for multi-hop radio networks," *IEEE/ACM Transactions on Networking*, vol. 1, no. 2, pp.166-177, April 1993.

[46] V. Kumar, M. Marathe, S. Parthasarathy and A. Srinivasan, "End-to-end packet scheduling in wireless ad-hoc networks," *ACM-SIAM Symposium on Discrete Algorithms*, British Columbia, Canada, January 2005.

[47] J. Chang and L. Tassiulas, "Energy conserving routing in wireless ad-hoc networks," *IEEE Infocom*, Tel Aviv, Israel, March 2000.

[48] V. Rodoplu and T. Meng, "Minimum energy mobile wireless networks," *IEEE Journal on Selected Areas in Communications*, vol. 17, no. 8, pp.1333-1344, August 1999.

[49] R. Wattenhofer, L. Li, P. Bahl and Y. Wang, "Distributed topology control for power efficient operation in multihop wireless ad hoc networks," *IEEE Infocom*, Anchorage, Alaska, April 2001.

[50] M. Grossglauser and D. Tse, "Mobility increases the capacity of ad hoc wireless networks," *IEEE/ACM Transactions on Networking*, vol. 4, no. 10, pp. 477-486, August 2002.

[51] T. ElBatt and A. Ephremides, "Joint scheduling and power control for wireless ad-hoc networks," *IEEE Infocom*, New York City, New York, June 2002.

[52] R. Gupta, Z. Jia, T. Tung and J. Walrand, "Interference-aware QoS routing (IQRouting) for ad-hoc networks," *http://walrandpc.eecs.berkeley.edu/Papers/RG_IQRouting.pdf*

[53] X. Lin and N. B. Shroff, "The impact of imperfect scheduling on cross-Layer rate control in multihop wireless networks," *IEEE Infocom*, Miami, Florida, March 2005.

[54] Y. Yi and S. Shakkottai, "Hop-by-hop congestion control over a wireless multi-hop network," *IEEE Infocom*, Hong Kong, China, March 2004.

[55] J. Zander, "Performance of optimum transmitter power control in cellular radio systems," *IEEE Transactions on Vehicular Technology*, vol. 41, no. 1, pp. 57-62, Feburary 1992.

[56] L. Tassiulas and A. Ephremides, "Stability properties of constrained queueing systems and scheduling policies for maximum throughput in multihop radio networks," *IEEE Transactions on Automatic Control,* vol. 37, no. 12, pp.1936-1949, December 1992.

[57] L. Tassiulas, "Scheduling and performance limits of networks with constantly changing topology," *IEEE Transactions on Information Theory,* vol.43, no.3, pp.1067-1073, May 1997

[58] M. Neely, E. Modiano and C. Rohrs, "Power and server allocation in a multi-beam satellite with time varying channels," *IEEE INFOCOM,* New York, NY, June, 2002

[59] P. R. Kumar and S. P. Meyn, "Duality and linear programs for stability and performance analysis of queueing networks and scheduling policies," *IEEE Transactions of Automatic Control*, vol.41, pp4-17, January 1996

[60] S. Asmussen, *Applied Probability and Queues*, Wiley 1987

[61] I. C. Paschalidis, "Large deviations in high speed communication networks," Ph.D Dissertation, MIT LIDS, May 1996

[62] T. Ren and R. J. La, "Downlink beamforming algorithms with inter-cell interference in cellular networks," *to appear in the proc. of IEEE INFOCOM 2005*

(longer version submitted to IEEE Transactions on Wireless Communications), http://www.ece.umd.edu/ hyongla/PAPERS/twc04_ren.pdf

[63] X. Liu, E. Chong, and N. Shroff, "Opportunistic transmission scheduling with resource-sharing constraints in wireless networks," *IEEE Journal on Selected Areas in Communications,* vol. 19, no. 10, pp. 2053-2064, October, 2001.

[64] J. Lee, R. Mazumdar and N. Shroff, " Opportunistic power scheduling for multi-server wireless systems with minimum performance constraints," *IEEE Infocom*, Hong Kong, China, March 2004.

[65] X. Qin and R. Berry, "Opportunistic splitting algorithms for wireless networks," *IEEE Infocom*, Hong Kong, China, March 2004.