A CROSS-LAYER DESIGN FRAMEWORK FOR RESOURCE ALLOCATION IN WIRELESS DATA NETWORKS

by

Ming Hu

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

ARIZONA STATE UNIVERSITY

July 2004

A CROSS-LAYER DESIGN FRAMEWORK FOR RESOURCE ALLOCATION IN WIRELESS DATA NETWORKS

by

Ming Hu

has been approved

July 2004

APPROVED:

_____, Chair

_____

_____

_____

_____

Supervisory Committee

ACCEPTED:

_____

Department Chair

_____

Dean, Division of Graduate Studies

ABSTRACT

This thesis explores cross-layer design for resource allocation in wireless data networks, such as cellular networks and ad hoc networks. It consists of three major thrusts. The first thrust is on data communications in code division multiple access (CDMA) cellular networks. It is shown that in the downlink of CDMA networks, the multi-access interference (MAI) is self-similar, under standard assumptions on ON/OFF traffic flows and fading channels. Building on this, the corresponding predictive temporal structure of the MAI process is exploited for adaptive resource allocation, such as rate control and admission control. In the second thrust, smooth admission control (SAC) and traffic aided opportunistic scheduling (TAOS) are proposed for channel-access control in opportunistic communication systems. The SAC scheme carries out admission control effectively while meeting quality of service (QoS) requirements. By making use of file size information and channel variation in a unified manner, the TAOS algorithms can reduce the completion time significantly. In the third thrust, the utility of multiple-input multiple-output (MIMO) techniques in ad hoc networks is studied. Specifically, medium access control (MAC) protocols using spatial diversity or directional antennas are devised. Then, the corresponding saturation throughput is evaluated via analytical methods and GloMoSim simulations. Furthermore, joint design of MIMO MAC and routing is investigated. The impacts of the hop lengths, rate adaptation and contention levels on the end-to-end delay in MIMO ad hoc networks are examined; and the optimal hop length is characterized accordingly. In summary, this thesis research has made some promising steps towards a cross-layer design framework for resource allocation in wireless data networks.

To my parents, Lei, and Kelvin

ACKNOWLEDGMENTS

I wish to thank a multitude of people who have been helping me. In particular, I would like to show my sincerest gratitude to my advisor Dr. Junshan Zhang for his devoted guidance, constant encouragement, and constructive criticism. Working with Dr. Junshan Zhang has been my invaluable and delightful experience, and his keen advice has helped me both professionally and personally. I would also thank Dr. Tolga M. Duman, Dr. Antonia Papandreou-Suppappola, Dr. Martin Reisslein, and Dr. Cihan Tepedelenlioglu for being on my advisory committee and sharing their wisdoms.

I am indebted to my friends and colleagues Chunyu Bi, Zheng Zhang, Yiwen Wu, Tansal Gucluoglu, Mustafa N. Kaynak, Subhadeep Roy, Eric C. Wang, Shabana Jabeen, Raja Tupelly, Bo Wang, Dong Zheng, Qian Ma, Ping Gao, Ning He, Kai Bai, Zhifeng Hu, and Weiyan Ge, for their insightful discussions and warm friendship. Many thanks also go to Mrs. Cynthia Moayedpardazi for her devoted professional service in our group.

Finally, I wish to thank my entire extended family for providing a loving environment for me. I would like to thank my wife, Lei Cheng, for her love, understanding, patience, and encouragement when it was most required. I am very grateful to my parents for their everlasting love and encouragement, and to Kelvin Hu for being my another accomplishment. Very special thanks to Jim and Liz Lancaster for their friendship, love, and support.

This thesis research is partially supported by National Science Foundation through the grant ANI-0238550 and by a gift grant from the Intel Research Council.

TABLE OF CONTENTS

Page

Page

Page

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

Figure Page

CHAPTER 1

Introduction

Propelled by the dream of supporting communications with anyone, anywhere, at any time, researchers have been dedicated to develop wireless multimedia communication systems and deploy wireless communication services. In the past ten years, there has been a tremendous growth of wireless personal communications. In this thesis, we consider two types of wireless networks. The first class of networks use a fixed network infrastructure, and typically consist of fixed base stations (or access points) and wireless mobile stations. The base stations are connected via the wireline (or wireless) backbones, and the mobile stations communicate with a base station via shared wireless links. A typical application of such centralized systems is cellular networks. The second class of networks, commonly referred as mobile ad hoc networks, do not need a fixed infrastructure or central administrators. Such a network is organized in an ad hoc manner. That is, the whole network is constructed dynamically, and each node independently makes its own decision on medium access and network routing. Each node operates not only as a source/destination but also as a router, forwarding the packets for other nodes whenever necessary. Due to the easiness in deployment, ad hoc networks have great potentials for being used in critical situations such as battlefields and disaster-relief events, and in small-area networks such as wireless LANs.

1.1. Cross-Layer Design in Wireless Networks

In traditional communication networks, the open systems interconnection (OSI) layer architecture has been widely adopted, and the performance optimization is conducted largely within each individual protocol layer. Although the layered structure is still of great importance for next generation wireless networks, it is known that the conventional OSI architecture may not work well in many wireless data applications. For instance, consider TCP traffic over wireless links. The TCP protocol being used may have no knowledge of the channel condition at the physical layer: it assumes that the channel is reliable and interprets all packet losses as being congestion related. Whenever a loss occurs in the wireless link, the TCP source reacts to this as if it was due to congestion and decreases the packet transmission rate, resulting in throughput degradation. To resolve this issue, an explicit congestion notification (ECN) mechanism is proposed [9], [27]. In this scheme, one ECN bit is added to the packet head. If the router detects congestion, it will mark the ECN bit. When the marked packet eventually reaches the destination, the destination can inform the source about the congestion. With this explicit information, the source can identify the causes of the loss — due to either fading or congestion, and react accordingly. By this means, the degradation incurred by the "misunderstanding" can be mitigated [9]. Clearly, when the higher layer can retrieve some information of the physical layer, the system performance may be improved. Thus, we believe that optimizing the system design in a cross-layer manner can yield significant gains, and this is the theme encountered throughout this thesis.

Cross-layer design in cellular networks has recently garnered much attention. For instance, channel-aware approaches, such as rate adaptation and incremental redundancy

transmission, have been developed and eventually adopted into the 3G standards [62]. With such channel-aware schemes, with 5MHz bands, the WCDMA can support data rates up to 2.048 Mb/s (combining 6 code channels), while cdma2000 can achieve a data rate 614.4 kb/s. Along a different avenue, opportunistic scheduling, which incorporates channel state information into medium access control, has also been studied extensively. In particular, Tse [94] proposes a scheduling scheme based on the proportional fairness criterion. Liu, Chong, and Shroff [59] present a resource-based fair sharing opportunistic scheduling scheme. In [20], Borst and Whiting study a class of revenue-based scheduling strategies. In [3], Agrawal, Bedekar, and et al. develop class and channel condition based scheduling schemes for EDGE/GPRS. In [7], Andrews, Kumaran, and et al. develop a modified largest weighted delay first (M-LWDF) scheduling to address QoS provisioning. In [82], an exponential scheduling rule is proposed by Shakkottai, Srikant, and Stolyar. In [101], Viswanath, Tse and Laroia propose an opportunistic beamforming scheme, which can achieve both beamforming gain and multiuser diversity gain.

Cross-layer design in ad hoc networks has also received much attention. To name a few, in [77], Sadeghi, Kanodia, and et al. propose an opportunistic auto rate (OAR) protocol. The main idea of the OAR is to transmit more packets when the channel of a source-destination pair is good, in contrast to sending one packet each time as in the IEEE802.11 standards. In [22], Choudhury, Yang, and et al. incorporate the direction of arrival (DOA) information from the directional antennas into the medium access control (MAC), and devise basic directional MAC (DMAC) and multi-hop RTS MAC (MMAC) accordingly. A similar approach is also developed by Korakis, Jakllari, and Tassiulas in [54]. In [12], a receiver-oriented multiple access (ROMA) protocol is introduced by Bao and Garcia, to fully utilize the multiple-beam forming capability of antenna arrays.

1.2. Overview of Resource Allocation in Wireless Networks

It is well known that in wireless communications, radio frequency spectrum is scarce, and many users may have to share a limited bandwidth (channel). Meanwhile, fuelled by the explosive growth of the Internet, there has been a global demand for tetherless wireless data access. To meet the rapidly growing demand for wireless communications, significant efforts have been made on resource management to improve the wireless spectrum efficiency. In what follows, we begin with a brief introduction on some key techniques.

1.2.1. Multiple Access

Multiple access techniques allow the communication media to be shared among different users. In particular, three basic multiple access techniques, i.e., frequency division multiple access (FDMA), time division multiple access (TDMA), and code division multiple access (CDMA), are used in centralized networks. The first generation analog cellular system uses FDMA. Both TDMA and CDMA techniques has been implemented in second generation digital cellular systems, such as GSM and IS-95. The third generation standards (e.g., cdma2000 and UTRA-WCDMA) are mainly based on CDMA techniques. Providing smooth evolution to the future generation networks, the 2.5G standards, IS-856 in North America (a.k.a., CDMA/HDR or 1xEV-DO) [13] and high speed downlink packet access (HSDPA) in Europe [70], use a hybrid CDM/TDM (i.e., time-division multiplexing) structure in the downlink. That is, spreading codes are used to "separate" different cells; and within one cell, users share the radio spectrum in a TDM manner.

In contrast to the above cases where a centralized coordinator provides orthogonality among the transmissions of different users, some wireless data networks (referred as packet radio systems in the past) require little coordination and adopt contention-based multiple-

access schemes, such as ALOHA and carrier sensing multiple access (CSMA) protocols. In an ALOHA system, each user transmits when it has data to send, and then waits for an acknowledgement. If a collision occurs, the user backs off for a random period and retransmits the message. Different from ALOHA, in CSMA protocols, each user listens to the channel before engaging in transmissions. If the channel is idle, the user is allowed to sent packets.

1.2.2. Admission Control

In wireless communications, admission control is needed to provide quality of service (QoS) for users. In voice-centric cellular systems, admission control is utilized to admit as many users as possible (to maximize the revenue of the system) while maintaining a certain QoS level for ongoing connections. Since voice traffic can be modeled as a poisson process, this control issue is relatively well understood [74]. In wireless data systems, however, admission control is more challenging. In particular, multimedia traffic can be very bursty, and is highly heterogenous in terms of the QoS requirements, ranging from a small text message with weak requirements to high-data-rate video steaming with stringent delay requirements. Furthermore, the wireless channels often exhibit time-varying fading. Such high variations in both channel conditions and traffic flows, make admission control very challenging in wireless multimedia networks.

1.2.3. Rate Control

Voice-centric cellular systems are designed to provide good coverage for telephony services. In such a system, a minimum required signal to interference plus noise ratio (SINR) is guaranteed over 90–95 percent of the coverage area. As a result, in a large

portion of the area, the SINR can be much greater than the minimum requirement. In contrast, for packet data services, the higher SINR can be used to provide higher data rates. Research shows that cellular spectral efficiency (in terms of b/s/Hz/sector) can be increased by a factor of two or more if users with better links are served at higher data rates [62]. Rate adaptation can be achieved through a combination of variable spreading, coding, modulation, and code aggregation. To support rate adaptation, a key component is channel condition estimation. This has also been considered in the standards for wireless data services. For instance, in CDMA/HDR systems, in each slot there are two pilot bursts, which provide mobile users a means to estimate the channel conditions. Since the time slot duration is only 1.67ms, often smaller than the coherent time in practice, such estimation is fast enough to track the variation of the channels.

1.2.4. Opportunistic Scheduling

Opportunistic scheduling is a multiplexing strategy that makes use of the channel variation across users. It originates from a holistic view. Roughly speaking, in a multiuser wireless network, at each moment, it is most likely that there exists a user whose channel is boosted by constructive fading. By transmitting data to only the instantaneous "on-peak" user(s), opportunistic scheduling can efficiently utilize the wireless resources and thus dramatically improve the overall system throughput (e.g., [46], [94]). This gain achieved via channel-aware scheduling across the users is called *multiuser diversity* gain. To provide wireless data services, fairness among users also has to be considered in the design of opportunistic scheduling algorithms. Indeed, a proportional fair (PF) scheduling scheme has been implemented in CDMA/HDR and HSDPA systems [5], [13], [94].

1.3. Scope of The Thesis

In this thesis, we address cross-layer optimization and design for resource allocation in wireless data networks. In what follows, we outline our contributions and the organization of this thesis.

In Chapter 2, we study data communications in the downlink of CDMA systems. We exploit the recently discovered predictive temporal structure of the multi-access interference (MAI) for adaptive resource allocation, particularly rate control and admission control. Specifically, we first present our result that the MAI process in CDMA data networks is "self-similar", under standard assumptions on ON/OFF traffic flows and fading channels. The MAI self-similarity indicates that there exists a nontrivial predictive MAI structure, which enables more accurate interference prediction. The predictive MAI temporal structure is used to construct a multiple time-scale interference predictor. Rate adaptation is carried out based on the predicted MAI level. Our numerical results show that this rate control scheme achieves better performance than that based on the packet-level MAI prediction only. Building on the rate adaptation, we then devise a joint rate control and admission control scheme. We also investigate the impact of fading and traffic burstiness on the system performance.

In Chapter 3, we study access control in opportunistic communication systems. We first propose a traffic-aided smooth admission control (SAC) scheme that aims to guarantee throughput provisioning. Simply put, in the SAC scheme, the admission decision is "spread" over a trial period, by increasing gradually the amount of the time resource allocated to incoming users. Specifically, using the modified weighted proportional fair (WPF) scheduling, we devise a QoS driven weight adaptation algorithm, and the weights assigned

to new users are increased in a guarded manner. Then an admission decision is made based on the measured throughput within a time-out window. Our results show that the proposed SAC scheme works well in opportunistic communication systems.

Next, we explore the possibility of reducing the completion time by incorporating traffic information into opportunistic scheduling. In particular, we first establish convexity properties for opportunistic scheduling with the file size information. Then, we develop new traffic aided opportunistic scheduling (TAOS) schemes by making use of file size information and channel variation in a unified manner. We also derive lower and upper bounds on the total completion time. Our results show that the proposed TAOS schemes can yield significant reduction of the total completion time. The impact of fading, file size distributions, and random arrivals and departures of users, on the system performance, is also investigated.

In Chapter 4, we turn our attention to ad hoc networks with multiple antennas. Specially, we study multiple-input multiple-output (MIMO) ad hoc network in heavy-loaded regimes. When the spatial channels experience independent fading, we investigate the utility of spatial diversity for medium access control (MAC) design, and develop the corresponding MAC protocol, namely SD-MAC. In contrast, when the wireless channel has a strong line of sight (LOS), we exploit smart antennas to improve spatial reuse in ad hoc networks. We propose to use *directional listening* to resolve the hidden terminal problem incurred by the asymmetry in antenna gain. Building on this, we develop a MAC protocol using smart antennas, namely DA-MAC. The proposed DA-MAC takes into account a general directional antenna model with sidelobes, and makes use of directional listening, directional transmission, and directional reception. We also present analytical methods for characterizing the saturation throughput for ad hoc networks with these two multiple-antenna techniques. The performance of MIMO ad hoc networks using SD-MAC or DA-MAC is evaluated via

theoretical methods and GloMoSim simulations. Our results show that the cross-layer design using smart antennas and spatial diversity can yield significant throughput gains in ad hoc networks.

In Chapter 5, we study joint optimization of MAC design and routing in MIMO ad hoc networks. We note that an isolated cross-layer strategy may have unintended results, when such a strategy interacts with other layer protocols. Thus motivated, we extend our study to the joint consideration of MIMO MAC and routing. More specifically, assuming a homogeneous MIMO ad hoc network, we examine the impact of MIMO techniques on the routing. We first investigate the impact of hop length, rate adaptation, and contention on the end-to-end delay. Building on this, we characterize the hop length that minimizes the end-to-end delay. Note that in wireless communications, there exists a tradeoff between the hop length and the transmission rate. Moreover, in ad hoc networks using CSMA/CA (collision avoidance), the behavior of each user may be affected by the whole networks. The tradeoff, together with the inter-dependence, makes this optimization problem challenging. Our contribution here is to take some first steps to quantify these effects and characterize the optimal hop lengths. Our results show that under certain conditions there exists such an optimal solution for the hop length, and we provide algorithms to characterize it. We conclude that the gain from spatial diversity can be used to improve both transmission rates and hop-lengths.

CHAPTER 2

A Cross-Layer Approach for Bursty Traffic Over CDMA

In this chapter, we take a cross-layer approach and study bursty data communications in CDMA systems. It is well known that CDMA systems are interference-limited, indicating that the multi-access interference (MAI) is a key parameter that governs the system performance (see, e.g., [100], [102], [112]). Thus motivated, we first present a new approach for characterizing the MAI (consisting of both intercell and intracell interferences). More specifically, we incorporate explicitly both the bursty nature of data traffic and fading channel conditions, and characterize the MAI from a stochastic process perspective (in contrast to the marginal distributions). This approach cuts across the physical layer, medium access layer, and network layer, and opens a dimension to understand the MAI temporal correlation structure. Our finding reveals that the MAI exhibits scale-invariant burstiness and is "self-similar" across multiple time scales [115], [116]. The *MAI self-similarity* indicates, by definition, that the MAI is long-range dependent, i.e., there exists extended periods of either strong or weak interference. Clearly, the performance in any system with fixed rates may degrade significantly for a long period, either overloaded (corresponding to strong interference) or resource-underutilized (corresponding to weak interference). On the flip side, the MAI self-similarity implies that there exists a nontrivial predictive MAI structure at coarse time scales, which can be exploited for interference management to improve the sys-

tem performance. It is worth noting that the MAI (particularly intercell MAI) is known to be very difficult to cope with.

To exploiting the predictive MAI temporal structure, we first develop a rate adaptation scheme to ameliorate the system performance. Specifically, we exploit the predictive MAI temporal structure to construct a multiple time-scale interference predictor. Based on the predicted MAI level, our rate control can be summarized as follows: If the (predicted) future interference is weak, we increase the transmission rate via decreasing the spreading gain or increasing the code rate or a combination thereof; if the (predicted) future interference is strong, we decrease the transmission rate accordingly. This rate control scheme can be viewed as an example of joint adaptation between the link layer and physical layer. Our results show that rate adaptation can yield significant performance gain, and rate control using a multiple time-scale MAI predictor achieves significantly better performance than that with a packet-level MAI predictor only.

The MAI temporal correlation structure can also be utilized for admission control. In particular, we devise a joint admission control and rate control scheme. Specifically, we propose a sliding observation window scheme which has a two-tier flavor: each observation window is divided into many time slots; rate control based on the interference prediction is conducted within each slot, and the corresponding throughput in the observation window is used for admission control. Based on the predicted (available) throughput, the system makes admission decisions, i.e., a new user is admitted if its throughput and delay requirements can be met, and vice versa if the opposite is true. Our results show that the above algorithm achieves high prediction accuracy and may be very useful in bursty data CDMA systems.

Our above exploitation of the MAI temporal correlation structure is reminiscent of multiuser detection [100], which exploits the MAI snapshot structure at the *symbol level*—

a much finer time scale, whereas rate control and admission control decisions are taken on *coarser time scales*. Clearly, resource allocation based on the predictive MAI temporal structure is a transmission technique, whereas multiuser detection is a signal processing technique implemented at the receiver end. These two exploitations of the MAI structure complement each other. We note that the exploitation of the MAI temporal correlation structure has barely been studied, and the main goal of this chapter is to provide some steps along this line.

The impact of fading and traffic burstiness on the performance of rate control and admission control is also investigated. Our results show that the rate adaptation scheme using a multiple time-scale MAI predictor performs significantly better than that with a packet-level predictor only, in fast fading channels. Also, we show that the more bursty the traffic is, the more gain our rate control and admission control yield.

In related work, there has been a great deal of research studying resource allocation in CDMA networks. For example, schemes with variable spreading gains to provide multiple rates can be found in [36], [66]. Joint rate control and power control have been studied extensively (see, e.g., [44], [65], [86]). Uplink access control for CDMA systems can be found in [21], [57], [84], [110]. A call admission algorithm using the shadow cluster concept is given in [55]. A recent work [23] presents a joint admission control and flow control algorithm for wireless web browsing. Joint admission control and power control have been studied in [6, 8], [108] and the references therein. A throughput maximization scheme is provided for systems with two user classes, one consisting of data users and the other real-time users [73]. (We also note that self-similarity in network traffic has been utilized for congestion control in wireline networks [68], [96]).

2.1. Model Description

2.1.1. MAI Processes in the Downlink

Consider the downlink of a cellular CDMA network with many ON/OFF data users. (Similar studies can be carried our for the uplink.) By ON/OFF we means that the transmission of each user is ON (active) and OFF (idle) alternatively. As is standard [106], we assume that the ON/OFF periods are heavy-tailed and exhibit the *Noah Effect* (i.e., have high variability or infinite variance; see [106] and the references therein). Intuitively, the Noah Effect for an individual ON/OFF source model yields ON and OFF periods that can be very large with non-negligible probability.

Let J be the number of cells under consideration. Assume that there are totally K_j ON/OFF users in cell j, $j = 1, \cdots, J$. Without loss of generality, consider user 1 in cell 1. Assume that the matched filter is employed to process the received signal. In a direct-sequence (DS) CDMA system, the interference to the desired user is the superposition of all other signals from the base stations in the network. Then, the total received power at time t due to the transmissions in cell j (in the downlink), denoted $\boldsymbol{I}_{1,j}$, is given by

$$\boldsymbol{I}_{1,j}(t) = \left(\sum_{k \in \text{cell } j} P_{k,j}(t) X_{k,j}(t) \right) g_{1,j}(t), \tag{2.1}$$

where $P_{k,j}(t)$ is the transmission power from the base station for user k in cell j at time t, $g_{1,j}(t)$ denotes the fading coefficient from the base station in cell j to the user under consideration, and $X_{k,j}(t) \in \{0, 1\}$ is the activity indicator for user k in cell j and $X_{k,j}(t) = 1$ if and only if the user is ON at time t.

The MAI, consisting of both intercell interference and intracell interference, is a key parameter that limits the capacity of CDMA systems. For convenience, we use $\boldsymbol{I}(t)$ to denote the total MAI. In a large network, $\boldsymbol{I}(t)$ is well approximated by $\sum_{j=1}^{J} \boldsymbol{I}_{1,j}(t)$. (Strictly

speaking, the intracell MAI is equal to $\sum_{k\neq 1, k\in \text{cell } 1} P_{k,1}(t)X_{k,1}(t)g_{1,1}(t)$. However, the signal from one single user is negligible in a large network.) Also we note that the intercell interference always exists, even though in theory the intracell interference can be eliminated by using orthogonal spreading if there is no multipath [32],[74].

The above model has a root in an earlier work [60]. One key difference between our study and theirs is that [60] presented a snapshot analysis for power control algorithms in the uplink, whereas here we take into account the burstiness of the ON/OFF users, and characterize the MAI from a stochastic process perspective for the downlink. This new approach enables us to understand the MAI temporal correlation structure across multiple time scales. Indeed, as shown below, the heavy-tailedness of the ON/OFF sources results in the self-similarity of the MAI process, indicating the existence of a nontrivial predictive structure of the MAI across multiple time scales, which is exploitable for efficient resource allocation [115].

2.1.2. Fading Channel Models

In a wireless system, fading effects can be classified as large-scale (slow) fading effects and small-scale (fast) fading effects [45], [72], [111]. Large-scale fading includes distance-related attenuation and slow-shadowing fading, and its duration is on the order of seconds. On the other hand, fast fading is due to the scattering of the transmitted signals off surrounding objects, and is on the order of milliseconds. Small-scale fading is superimposed on top of the large-scale fading.

We consider both slowly fading and fast fading. Specifically, we assume that fading is due to distance-related attenuation, log-normal shadowing, and fast fading. The propagation attenuation is in the form of $d^{-\beta}$ where d is the distance, β is the path loss exponent,

and the slow-shadowing fading is log-normal with standard deviation $\sigma_\Omega = 8\,\text{dB}$ [87]. We use Gudmundson's auto-regressive (AR) model [33], [87] for log-normal shadowing. Suppose that the signal strength is sampled every time interval T_f, and ξ is the correlation coefficient of two consecutive samples. Then, the AR model has the form of

$$\Omega_{n+1} = \xi\Omega_n + (1-\xi)v_n, \tag{2.2}$$

where Ω represents the log-normal fading (in dB), v_n is zero-mean white Gaussian noise with variance $\sigma_\Omega^2(1+\xi)/(1-\xi)$, and ξ is the correlation coefficient with $0 < \xi \leq 1$. Fast Rayleigh fading (with the Doppler shift f_m [87]), is superimposed on top of slowly fading. A filtered Gaussian noise model for Rayleigh fading is utilized in our study [87].

2.1.3. Data Traffic Models

For web traffic, it has been shown that the probability of large file size is not negligible, and that the ON duration is effectively characterized by heavy-tailed models [105]. The OFF duration is determined by the user's *thinking time*, which is also modeled as heavy-tailed [24]. Along this line, we assume throughout that the data traffic of each user is an ON/OFF process, where both ON and OFF periods are Pareto distributed (it is worth pointing out that our result on the MAI long-range dependence holds for general heavy-tailed distributions, as is clear in the proof of Theorem 2.2.1), that is,

$$Pr\{T > t\} = (T_{\min}/t)^\alpha, \tag{2.3}$$

where, $T_{\min}$ denotes the smallest possible value random variable T can take. The parameters in this traffic model are specified as follows:

a) $T_{\min,1}$: Minimal ON duration, which is determined by the minimal file size and transmission rate. According to [24] and [106], the minimal file size for web traffic is about

2 kBytes. (For example, assume that the wireless system provides an average service of about 100 kb/s for each user. Then $T_{\min,1}$ is about 0.2 s for each burst transmission.)

b) $T_{\min,2}$: Minimal OFF duration, which is mainly determined by the user's think time, according to Crovella [24]. It varies from about 1 to 30 seconds.

c) α_{on}: It is determined by the slope of file size distribution, and is 1.3 in this study.

d) α_{off}: It is determined by the slope of think time distribution, and is 1.5 in this study.

For convenience, we define $\alpha_{\min} = \min(\alpha_{\text{on}}, \alpha_{\text{off}})$. For simplicity, we assume that the distributions of ON/OFF-periods are the same for all users. Our study here can be easily generalized to the cases where the distributions of ON/OFF-periods are different across users (see, e.g., [90, Theorem 2]).

2.2. MAI Self-Similarity and Predictive MAI Temporal Structures

2.2.1. A Brief Overview of Self-Similar Models

In what follows, we provide a brief overview of heavy-tailed distributions and self-similar processes (see, e.g., [2], [69]).

Definition 2.1 A random variable Y has a *heavy-tailed* distribution if

$$Pr\{Y > y\} \sim \ell_1 y^{-\alpha} \tag{2.4}$$

as $y \to \infty$, where $0 < \alpha < 2$, and ℓ_1 is some constant.

Roughly speaking, the asymptotic shape of the distribution follows a power law, in contrast to the exponential decay. Heavy-tailed distribution, by definition, implies that a "larger" portion of the probability mass moves to the tail of the distribution, as α decreases.

Definition 2.2 For a given stationary time series $\{\boldsymbol{I}_t, t \in Z^+\}$, define the corresponding aggregated series $\boldsymbol{I}_i^{(m)}$ as

$$\boldsymbol{I}_i^{(m)} = \frac{1}{m}(\boldsymbol{I}_{im-m+1} + \cdots + \boldsymbol{I}_{im}). \tag{2.5}$$

Let $r(k)$ and $r^{(m)}(k)$ denote auto-correlation functions of $\{\boldsymbol{I}_t\}$ and $\{\boldsymbol{I}_i^{(m)}\}$, respectively. We say that $\{\boldsymbol{I}_t\}$ is *asymptotically self-similar* (with Hurst parameter $1/2 < H < 1$) if the following conditions are satisfied:

$$r(k) \sim \ell_2 k^{-\beta}, \tag{2.6}$$

$$r^{(m)}(k) \sim r(k), \tag{2.7}$$

as $k \to \infty$, where $\beta = 2(1 - H)$ and ℓ_2 is some constant.

A key parameter associated with self-similar processes is the above mentioned *Hurst parameter*, with range $1/2 < H < 1$. Indeed, The Hurst parameter H is sometimes called the index of self-similarity. Roughly speaking, the further H deviates from $1/2$, the more "long-range dependent" the $\{\boldsymbol{I}_t\}$ is.

2.2.2. MAI Self-Similarity

In the following, we characterize the MAI process in the downlink. The MAI process, $\sum_{j=1}^{J} \boldsymbol{I}_{1,j}(t)$, is a superposition of the interference processes generated by both intercell and intracell interferers. It is clear that the interference process from a user, say user k in cell j, is a product of its ON/OFF process $X_{k,j}(t)$ and the corresponding fading process $g_{1,j}(t)$. The fading coefficients can be treated as "rewards" for the underlying renewal process (i.e., the ON/OFF process). It has been showed that the superposition of many ON/OFF heavy-tailed Ethernet sources is self-similar [24], [105] where the "rewards" in

each ON/OFF period are assumed as constants. Clearly, fading makes the characterization of the MAI process more challenging. Indeed, we cannot hope that the MAI process can be approximated by a Fractional Brownian Motion or even a Gaussian process [90], because the factor $g_{1,j}(t)$ in the MAI process is common for all users in cell j and there might be strong correlation between the summands of the MAI.

For technical reasons, we impose the following assumptions:

Condition 1 The sample paths of the fading process $g_{1,j}(t)$ are continuous, and the correlation function of $g_{1,j}(t)$, denoted as $\varrho_{1,j}(u)$, is $o(u^{2H-2})$ as $u \to \infty$, $j = 1, \ldots, J$.

Condition 2 The empirical distribution of the transmission powers of the users in cell j, converges weakly to a distribution function F_p with mean μ_p, $j = 1, \cdots, J$.

We note that the above conditions hold in many practical systems of interest (see [115, 116] for more details). We are now ready to present our result on the MAI self-similarity.

Theorem 2.2.1 *Suppose conditions 1–2 hold. As both T and W increase, the total accumulated MAI, $\frac{1}{TW}\int_0^{Tt}\sum_{j=1}^{J}\boldsymbol{I}_{1,j}(u)\,du$, is asymptotically self-similar with Hurst parameter $H = (3-\alpha_{\min})/2$.*

Proof: We provide a proof of part a) in the following. Part b) follows by using [90, Theorem 2].

Consider the MAI from cell i. For convenience, define

$$\begin{aligned}\boldsymbol{U}_{K_i}(t) &\triangleq \sum_{k\in C_i} P_{k,i}(t)X_{k,i}(t),\\ \tilde{\boldsymbol{U}}_{K_i}(t) &\triangleq \frac{1}{\sqrt{K_i}}\left(\boldsymbol{U}_{K_i}(t) - \mathbb{E}[\boldsymbol{U}_{K_i}(t)]\right),\end{aligned}$$

and for any $x > 0$,

$$\boldsymbol{S}_{K_i}(x) \triangleq \int_0^x g_{1,i}(t)\tilde{\boldsymbol{U}}_{K_i}(t)dt.$$

First consider $x \in [0,1]$. By the central limit theorem, $\tilde{\boldsymbol{U}}_{K_i}(t)$ converges weakly, as $K_i \to \infty$, to a Gaussian process with mean zero. (We note that here weak convergence is interpreted as weak convergence of probability measures on $C[0,1]$ with the uniform topology.) Let $\tilde{\boldsymbol{U}}(t)$ denote the corresponding limiting Gaussian process, and define the process

$$\boldsymbol{S}(x) \triangleq \int_0^x g_{1,i}(t)\tilde{\boldsymbol{U}}(t)dt.$$

Then, the accumulated MAI from cell i over the interval $[0,x]$ is given by

$$\begin{aligned}
\frac{1}{\sqrt{K_i}}\int_0^x \boldsymbol{I}_{1,i}(t)dt &= \int_0^x g_{1,i}(t)\tilde{\boldsymbol{U}}_{K_i}(t)dt + \frac{1}{\sqrt{K_i}}\int_0^x g_{1,i}(t)\mathbb{E}[\tilde{\boldsymbol{U}}_{K_i}(t)]dt \\
&= \boldsymbol{S}_{K_i}(x) + \frac{1}{\sqrt{K_i}}\int_0^x g_{1,i}(t)\mathbb{E}[\tilde{\boldsymbol{U}}_{K_i}(t)]dt \\
&= \boldsymbol{S}_{K_i}(x) + \frac{1}{\sqrt{K_i}}\frac{\mu_1}{\mu_1+\mu_2}\int_0^x g_{1i}(t)\sum_{k=1}^{K_i}\mathbb{E}[P_{k,i}(t)]dt. \qquad (2.8)
\end{aligned}$$

A direct application of the monotone convergence theorem [76] yields that

$$\lim_{K_i\to\infty}\frac{1}{K_i}\int_0^x g_{1i}(t)\sum_{k=1}^{K_i}\mathbb{E}[P_{k,i}(t)]dt = \mu_p\int_0^x g_{1i}(t)dt.$$

Furthermore, by Condition 1, it is straightforward to see that

$$\text{var}\left[\int_0^x g_{1i}(t)dt\right] \sim o(x^{2H}) \quad \text{as } x\to\infty. \qquad (2.9)$$

Therefore, it suffices to characterize $\text{var}[\boldsymbol{S}_{K_i}(x)]$ as $K_i \to \infty$ and $x \to \infty$.

Next, we show, by constructing the processes on a common probability space[1], that $\boldsymbol{S}_{K_i}(x)$ converges weakly to the process $\boldsymbol{S}(x)$. Recall that we assume that x ranges on a finite interval $[0,1]$. (We will elaborate on the case when x ranges over $[0,\infty)$.) By Condition 1, $g(t)$ is bounded for $0 < t \le x$. Since $\tilde{\boldsymbol{U}}_{K_i}(t)$ converges weakly to $\tilde{\boldsymbol{U}}(t)$, by appealing to Skorohod's Theorem [17, Theorem 25.6] (see also [85, Theorem 3; p. 357]),

[1]This is so called "the method of a single probability space" [85].

there exists a probability space supporting $\tilde{\boldsymbol{U}}^*_{K_i}(t)$, $K_i = 1, 2, ...$ and $\tilde{\boldsymbol{U}}^*(t)$ jointly (where $\tilde{\boldsymbol{U}}^*_{K_i}(t) \stackrel{d}{=} \tilde{\boldsymbol{U}}_{K_i}(t)$ and $\tilde{\boldsymbol{U}}^*(t) \stackrel{d}{=} \tilde{\boldsymbol{U}}(t)$), so that $\sup_{0\le t\le 1} |\tilde{\boldsymbol{U}}^*_{K_i}(t) - \tilde{\boldsymbol{U}}^*(t)| \to 0$ as $K_i \to \infty$, almost surely. Since $g_{1,i}(t)$ is independent from the $\tilde{\boldsymbol{U}}_{K_i}(t)$'s, we can enlarge the probability space to support this $g(t)$ as well. Then, the continuity of $g_{1,i}(t)$ assures that

$$\begin{aligned} & \sup_{0\le x\le 1} \left| \int_0^x g_{1,i}(t)\tilde{\boldsymbol{U}}_{K_i}(t)dt - \int_0^x g_{1,i}(t)\tilde{\boldsymbol{U}}(t)dt \right| \\ \stackrel{d}{=} & \sup_{0\le x\le 1} \left| \int_0^x g_{1,i}(t)\tilde{\boldsymbol{U}}^*_{K_i}(t)dt - \int_0^x g_{1,i}(t)\tilde{\boldsymbol{U}}^*(t)dt \right| \to 0, \end{aligned} \tag{2.10}$$

as $K_i \to \infty$, which dictates directly that $\boldsymbol{S}_{K_i}(x)$ converges to $\boldsymbol{S}(x)$, for $0 < x < 1$. Now let us turn to the case when x ranges over $[0, \infty)$. Because weak convergence on $C[0, \infty)$ is, by definition, convergence with respect to the topology of uniform convergence on each finite interval (see, e.g., [104, p. 500]), i.e., a process $\{\boldsymbol{S}_{K_i}(x), x > 0\}$ converges weakly to $\{\boldsymbol{S}(x), x > 0\}$ if and only if, for each $x < \infty$, $\{\boldsymbol{S}_{K_i}(t), 0 < t \le x\}$ converges weakly to $\{\boldsymbol{S}(t), 0 < t \le x\}$. We conclude that $\{\boldsymbol{S}_{K_i}(x), x > 0\}$ converges weakly to the process $\{\boldsymbol{S}(x), x > 0\}$.

It remains to show that $\text{var}[\boldsymbol{S}(x)] \sim \ell_4 x^{2H}$ as $x \to \infty$ for some contant ℓ_4. Recall that

$$\boldsymbol{S}(x) = \int_0^x g_{1i}(t)\tilde{\boldsymbol{U}}(t)dt \tag{2.11}$$

Let ζ denote that the covariance function of $\boldsymbol{U}$. Then, we have that

$$\text{var}[\boldsymbol{S}(x)] = 2\int_0^x \int_0^v \zeta(u)\varrho(u)\ du\ dv + 2\left(\mathbb{E}[g_{1,i}(0)]\right)^2 \int_0^x \int_0^v \zeta(u)\ dudv \tag{2.12}$$

Since

$$\int_0^x \int_0^v \zeta(u)\ dudv \sim c_2 x^{2H},$$

it follows that

$$\text{var}[\boldsymbol{S}(x)] \sim \ell_4 x^{2H}, \quad \text{as } x \to \infty$$

We conclude that as both T and W go to infinity, the (normalized) MAI from cell i, $T^{-H}W^{-1/2}\int_0^{Tt}\boldsymbol{I}_{1,i}(u)du$ is asymptotically self-similar with the Hurst parameter $H = (3 - \alpha_{\min})/2$ (see, e.g., [75]). ■

2.2.3. Predictive MAI Temporal Structures

The MAI self-similarity, by definition, indicates that the MAI is long-range dependent and therefore there exists a MAI temporal correlation structure. Specifically, Theorem 2.2.1 reveals that the MAI process $\{\boldsymbol{I}(t), t \in \mathbb{R}\}$ is long-range dependent. Intuitively speaking, the MAI levels are highly correlated at coarser time scales, and there exists a nontrivial predictive MAI temporal structure.

A few observations are worth noting: 1) The predictive MAI (particularly intercell MAI) temporal structure exists in many realistic systems. For instance, in a cellular CDMA system, each cell has six neighboring cells, so it is likely that there are hundreds of co-channel users. Therefore, the approximation of the MAI process by long-memory models holds well. 2) The predictive MAI temporal structure exists in systems with heterogenous multimedia traffic, because it is the heavy-tailedness of traffic that results in the MAI temporal correlation structure. In a nutshell, we expect that the predictive MAI temporal structure exists in many multimedia CDMA systems, as long as the fading is not long-range dependent (which is applicable to many practical systems).

2.2.4. Numerical Examples

In what follows, we illustrate the above findings via numerical examples. In our simulation, we consider a cell with six adjacent neighboring cells. The base station in each cell uses omni-directional antennas. The total number of ON/OFF users in each cell is 120 (the

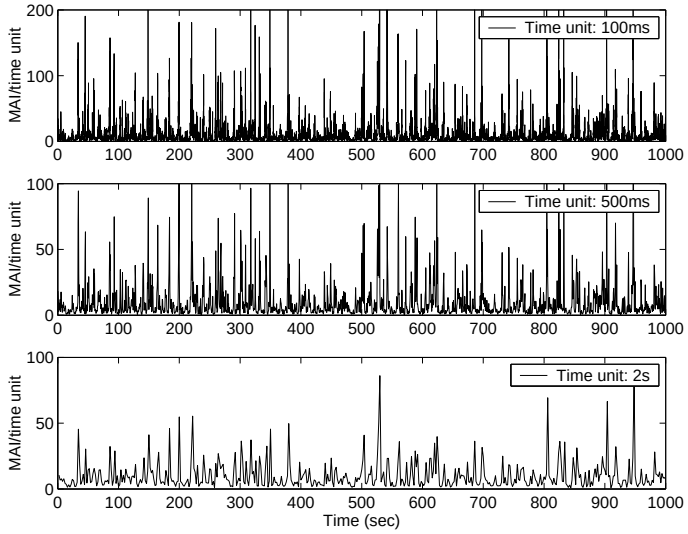

Figure 2.1. A pictorial "proof" of self-similar MAI: aggregated MAI exhibits "similar" burstiness on three time scales (100ms, 500ms, 2s). Some parameters of the fading channel are specified as follows: $T_f = 0.01$s, $\xi = 0.99$, $f_m = 5$Hz.

average number of active (ON) users is around 15). Assume that the average transmission rate for each user is around 100 kb/s. For the MAI sequence $\{\boldsymbol{I}_n, n = 1, 2, \cdots, N\}$ in the downlink, we define the corresponding aggregated sequence at the time scale $(10m)$ ms [89]:

$$\boldsymbol{I}_i^{(m)} = \frac{1}{m}(\boldsymbol{I}_{im-m+1} + \cdots + \boldsymbol{I}_{im}), \quad i = 1, 2, \cdots, [N/m].$$

Figure 2.1 depicts a sequence of simple plots of the average MAI for three time units (100 ms, 500 ms, and 2 s). As is evident in Figure 2.1, the MAI exhibits scale-invariant burstiness at multiple time scales and "looks" self-similar at coarser time scales.

As noted before, the Hurst parameter H is sometimes called the index of self-similarity. Estimating H plays a crucial role in diagnosing self-similarity. In Figure 2.2, we use the variance method [89] to estimate the Hurst parameter. Specifically, we compute the sample variance at the time scale $(10m)$ ms:

$$\text{var}(\boldsymbol{I}_i^{(m)}) = \frac{1}{[N/m]} \sum_{i=1}^{[N/m]} \left(\boldsymbol{I}_i^{(m)} - \bar{\boldsymbol{I}}\right)^2, \tag{2.13}$$

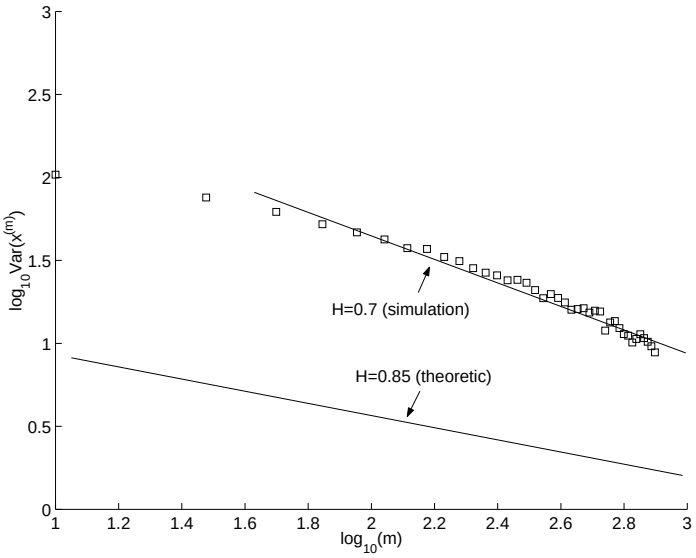

Figure 2.2. Estimating Hurst parameter via the variance method ($T_f = 0.01$s, $\xi = 0.99$, $f_m = 5$Hz)

where $\bar{\boldsymbol{I}}$ is the sample mean of the whole sequence $\{\boldsymbol{I}_n, n = 1, 2, \cdots, N\}$. The estimated Hurst parameter is 0.70, indicating that the MAI process is self-similar. (We note that the variance method is an informal diagnostic tool for checking if the Hurst parameter H is larger than 1/2 or not, and typically has biases [2], [14].)

2.3. Interference Prediction

Since CDMA systems are interference-limited, it is of vital importance to conduct effective interference management. To this end, a key step is to obtain accurate interference prediction, based on which we carry out resource allocation for interference management.

The MAI self-similarity ($H > 1/2$) indicates directly that there exists a predictive MAI temporal correlation structure at coarse time scales. In this section, we explore MAI prediction based on the MAI self-similarity. (We will utilize the predictive MAI temporal structure for feedback control at coarse time scales, particularly rate control and admission

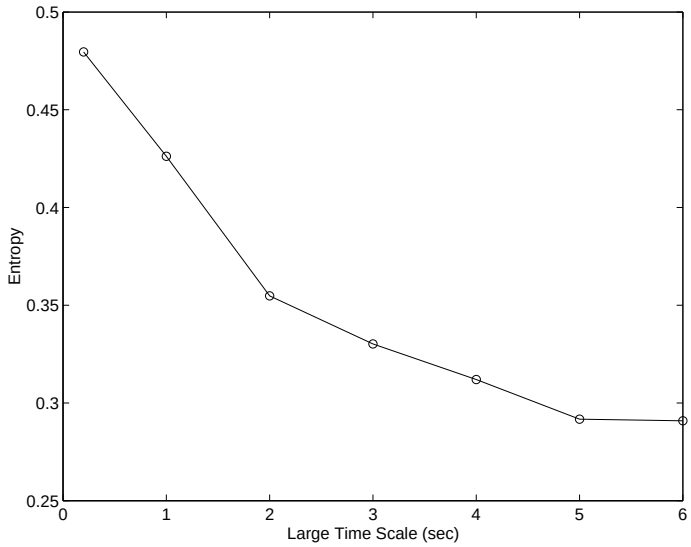

Figure 2.3. Entropy $\bar{S}$ versus large time scale T_m

control in Section 5 and 6.)

2.3.1. Optimum Time Scales for Exploiting Predictive MAI Temporal Structures

The time scale T_m for MAI prediction is a critical parameter because it determines the sample size m used to predict the future interference. If T_m is too small, the long-range dependence of the MAI process may not hold; on the other hand, if T_m is too large, it is difficult to imagine a mechanism that would be able to effectively exploit the prediction, simply because during such a long period, too many changes could occur, yielding zero net gain for control decisions. There are many methods useful for determining the time scale T_m. In what follows we use two methods, namely the variance method and the entropy method, to determine T_m.

Variance method: In the following, we first use the variance method to determine T_m. Using (2.13), we compute the sample variance at the time scale $(10m)$ ms, and plot it

against m (in the "log" scale). When the MAI is well modeled as self-similar, the plot becomes a straight line with a slope of $2H-2$. As shown in Figure 2.2, the plot approaches a line as the time scale grows (beyond $\log_{10}(m) > 2$). Therefore, it is reasonable to choose $T_m = 1$s (corresponding to $\log_{10}(m) = 2$), as a large time scale for MAI prediction to utilize the MAI self-similarity.

Entropy method: Next, we use the off-line entropy method to determine T_m, along the lines of [96]. More specifically, given a time scale $T_m > 0$, we define

$$V_1 = \sum_{i\in[t-T_m,t]} \boldsymbol{I}(i), \quad V_2 = \sum_{i\in[t,t+T_m]} \boldsymbol{I}(i). \tag{2.14}$$

We introduce two random variables L_1, L_2 as the quantization of V_1, V_2, i.e.,

$$L_j = L_j(V_j),\ L_j \in [1, M],\ j = 1, 2, \tag{2.15}$$

where M denotes the number of the quantization levels. The conditional probability density is denoted as $Pr\{L_2|L_1 = l\}$. Define

$$S_l = -\sum_{l'} Pr\{L_2 = l'|L_1 = l\} \log Pr\{L_2 = l'|L_1 = l\}. \tag{2.16}$$

Then the entropy is given by

$$\bar{S} = \frac{1}{M}\sum_{l=1}^{M} S_l. \tag{2.17}$$

Figure 2.3 gives the plot of $\bar{S}$ as a function of the time scale T_m, with $\alpha_{\text{on}} = 1.3$ and $\alpha_{\text{off}} = 1.5$. In Figure 2.3, we find that the entropy $\bar{S}$ begins to"converge" at 2 s in the sense that the entropy becomes "flat" when the time scale is larger than 2 s, indicating that the MAI becomes more "predictable" at time scales of 2 s or larger.

Comparing the variance method and the entropy method, we notice that the estimated large time scales are on the same order. We choose $T_m = 2$ s as the large time scale for MAI prediction in the following numerical examples.

2.3.2. Multiple Time-Scale MAI Predictors

Since rate control in our study is implemented at the packet level (see Section 5), we devise a multiple time-scale MAI predictor which combines the MAI predictions at the packet-level and T_m. Specifically, we have that

$$\hat{\boldsymbol{I}}(i) = \zeta \hat{\boldsymbol{I}}^P(i) + (1-\zeta)\hat{\boldsymbol{I}}^L(i),\ 0 < \zeta \leq 1, \tag{2.18}$$

where $\hat{\boldsymbol{I}}^P(i)$ is the MAI prediction at the packet level, and $\hat{\boldsymbol{I}}^L(i)$ is the MAI prediction at T_m. In the following, we elaborate on the MAI prediction at T_m and at the packet-level.

MAI Prediction at large time scale T_m: We first devise a MAI predictor at a large time scale T_m. In light of the limited computing and storage capabilities at the mobile user end, we propose to use the following simple-to-implement MAI prediction at T_m:

$$\hat{\boldsymbol{I}}^L(i) = \frac{1}{m}\sum_{n=i-m}^{i-1} \boldsymbol{I}_n, \tag{2.19}$$

where $\boldsymbol{I}_n$ is the measured MAI level for the nth packet, and m is the number of samples within T_m.

MAI prediction at packet level: We adopt the following simple MAI predictor at the packet level:

$$\hat{\boldsymbol{I}}^P(i) = \boldsymbol{I}_{i-1}, \tag{2.20}$$

where $\hat{\boldsymbol{I}}^P(i)$ is the MAI prediction at the packet level.

2.4. Rate Control

In a large CDMA network with many users, the signal-to-interference-plus-noise ratio (SINR) can be well approximated as

$$\mathrm{SINR}_{1,1}(t) = \frac{P_{1,1}(t)g_{1,1}(t)}{\sigma^2 + \frac{1}{G_{1,1}(t)}\sum_{j=1}^{J}\boldsymbol{I}_{1,j}(t)}, \tag{2.21}$$

where $G_{1,1}(t)$ is the processing gain of the desired user in cell 1, and σ^2 is the variance of the ambient additive white Gaussian noise. Furthermore, $G_{1,1}(t) = W/R_{1,1}(t)$, where $R_{1,1}(t)$ is the transmission rate, and W denotes the bandwidth.

As noted before, MAI exhibits scale-invariant burstiness at multiple time scales, and the "DC" component of interference can be either strong or weak for a relatively long period. Therefore, for systems with fixed transmission rates, there exist concentrated periods where performance is poor (strong MAI) or resources are under-utilized (weak MAI). In the following, we explore rate control to improve the system performance.

2.4.1. Rate Adaptation Algorithms

Rate control is a central technique for interference management in bursty data systems. Next we devise an easy-to-implement rate adaptation scheme based on the multiple time-scale MAI prediction. The underlying rationale can be summarized as follows: If the (predicted) future MAI is weak, we increase the transmission rate via decreasing the spreading gain or increasing the code rate or a combination thereof; if the (predicted) future MAI is strong, we decrease the transmission rate accordingly.

We assume that a continuous transmission rate can be achieved. However, in practical systems, the spreading gain can take values 2^k (k is an integer) only. So, in order to achieve a continuous transmission rate, we need to combine spreading gain control with adaptive coding. Specifically, we decompose the bandwidth redundancy (expansion) into two parts [99]:

$$G = G_s + G_c \ (dB), \tag{2.22}$$

where G_s is the bandwidth redundancy corresponding to spreading, and G_c denotes the bandwidth redundancy corresponding to coding.

For adaptive coding, we can choose rate compatible punctured convolutional codes or Turbo codes (the design of coding schemes is beyond the scope of this research). Suppose that there are N_r types of code rate modes available. We can construct a look-up table, via calculating the processing gain G for each combination of spreading gain G_s and code rate R_c. During the course of on-line control, the base station determines the pattern $\{G_s, R_c\}$ in the look-up table for certain desired SINR. Figure 2.4 gives a simple block diagram for the above algorithm.

Our rate control algorithm can be summarized briefly as follows:

1) The mobile user measures the strength of its downlink signal and the MAI. At the end of each packet, it makes a prediction of the MAI level for the next packet transmission, using our multiple time-scale MAI predictor.

2) The mobile user feeds the signal strength and the predicted MAI level back to the base station.

3) The base station determines the transmission rate to achieve the target SINR, and the corresponding spreading gain and code rate for the next packet transmission.

4) The base station adjusts its transmission rate accordingly.

2.4.2. Numerical Examples

In what follows, we evaluate the performance when the above rate control algorithm is applied to the example in Section 3. We assume that if the SINR of one entire packet is higher than the target SINR (denoted as γ), this packet is transmitted successfully, otherwise the packet is lost. The corresponding throughput is also called *Goodput* [78].

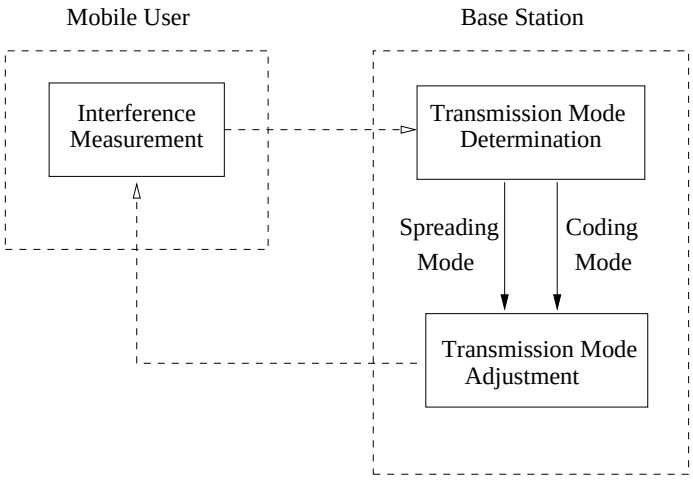

Figure 2.4. A simple diagram for transmission rate control

Then, the average throughput is given by

$$\widetilde{\mathrm{Th}} = \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} \mathrm{sgn}(\mathrm{SINR} - \gamma) \cdot R_i, \tag{2.23}$$

where

$$\mathrm{sgn}(z) = \begin{cases} 1 & \text{when } z > 0 \\ 0 & \text{otherwise} \end{cases},$$

R_i is the transmission rate of packet i, and T_p is the packet duration.

Let $\widetilde{\mathrm{Th}}^p$ denote the average throughput of a system using the packet-level MAI predictor only and $\widetilde{\mathrm{Th}}^m$ denote the one using the multiple time-scale MAI predictor. Define the relative throughput gain δ_{Th} as

$$\delta_{\mathrm{Th}} = \frac{\widetilde{\mathrm{Th}}^m - \widetilde{\mathrm{Th}}^p}{\widetilde{\mathrm{Th}}^p}. \tag{2.24}$$

We now compare the performance of two schemes: Scheme 1 employs the multiple time-scale MAI predictor with $T_m = 0.2\,\mathrm{s}$, and Scheme 2 uses the multiple time-scale MAI predictor with $T_m = 2\,\mathrm{s}$. In this example, we assume that the packet duration T_p is 10 ms, the bandwidth is 10 MHz, and the target SINR threshold is 4 dB for data services. Our

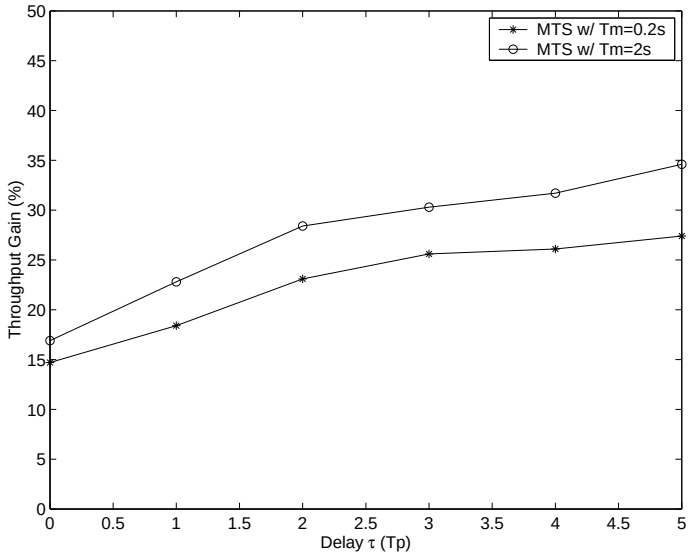

Figure 2.5. Throughput gain versus feedback delay ($T_f = 0.01$s, $\xi = 0.99$, $f_m = 30$Hz)

results are given in Figure 2.5. As would be expected, Scheme 2 achieves a larger throughput gain than Scheme 1. The underlying rationale is that we can get more accurate MAI prediction using the multiple time-scale method, and the choice of the large time scale also affects the prediction accuracy.

Feedback delay may exist in practical systems. It is natural to expect that feedback delay would diminish the accuracy of prediction. The effect of feedback delay on the system performance deserves consideration. In our simulation, the feedback delay is on the order of tens of milliseconds [62]. Figure 2.5 shows that the relative throughput gain increases with the feedback delay. Our intuition is that the larger the feedback delay is, the less accurate the packet-level MAI prediction is, and the more useful the large time-scale MAI prediction at T_m is.

2.5. Joint Admission Control and Rate Control

In this section, we exploit the predictive MAI structure at coarser time scales to explore admission control for data applications. For systems with voice users, admission control is based on the network capacity, which is defined as the maximum number of users that can be supported without violating their average SINR requirements [74], [78]. This approach, however, cannot be applied directly to systems with data users because of the highly bursty MAI (and hence SINR) even at large time scales. In particular, we expect that admission control schemes using the mean value of the SINR (or equivalently MAI) over one session would not work well. To resolve this issue, we propose a joint admission control and rate control scheme that has a two-tier flavor: Rate control based on the MAI prediction is conducted within a large observation window τ, and the corresponding throughput in τ is used for admission control.

Before we proceed to elaborate on the admission control algorithm, we need to set admission control criteria for data applications. In general, admission control involves the following two criteria [78]: 1) when a new user is admitted, the system should be able to guarantee that its QoS requirement is met; and 2) the QoS requirements of the other users already in the system should not be violated. In this study, since each user is assumed to adapt its transmission rate to the time-varying MAI, we expect that the impact of this new user on the other active transmissions is marginal. (Recall that a key feature of CDMA systems is that the performance degrades gracefully.) In light of this observation, we focus primarily on the QoS requirement of the new user. Specifically, we assume that a new user can be admitted only if its throughput requirement in a time window τ can be satisfied,

that is,

$$\psi_{\text{req}} \leq \hat{\psi}_{\text{av}}, \tag{2.25}$$

where ψ_{req} denotes the required throughput, and $\hat{\psi}_{\text{av}}$ denotes the predicted (available) throughput within τ. We also assume that new users can tolerate delay up to some value D, and τ (less than D) is chosen according to the type of the data application.

We now present our joint admission control and rate control scheme. Specifically, a time window τ (we call it *sliding observation window*) is divided into many time slots with length T_m. The average MAI in each slot is predicted via the large time-scale MAI predictor developed in Section 4, based on which the transmission rate is adjusted and the corresponding throughput within T_m is calculated. Recall that T_m is the large time scale for interference prediction based on the MAI self-similarity. Denote $N_s = \lceil \tau / T_m \rceil$, that is, N_s denotes the number of slots in τ. The total (predicted) throughput within a sliding observation window τ is given by

$$\hat{\psi}_{\text{av}}^{m} = \sum_{j=1}^{N_s} \text{Th}_j, \tag{2.26}$$

where Th_j is the predicted throughput in slot j, $j = 1, \cdots, N_s$. We note that the rate control in the above scheme is done at the time scale T_m. We show in the following that even this simple scheme yields significant performance gain. We expect that more sophisticated rate adaptation, would improve the performance even more.

Our admission control algorithm is summarized as follows:

1) The new user submits a connection request and also its QoS requirement in terms of $\{\psi_{\text{req}}, \tau, D\}$.

2) This user predicts the MAI level for every T_m, and feeds it back to the base station.

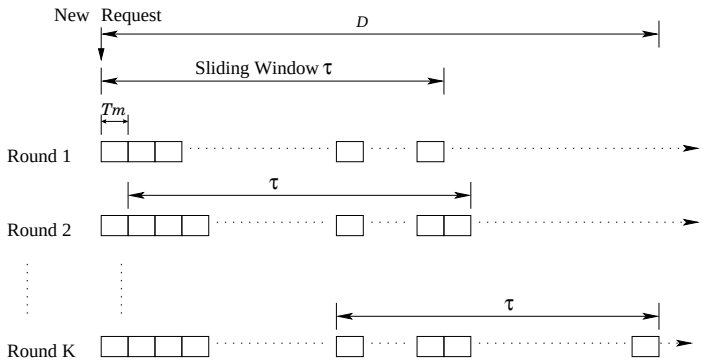

Figure 2.6. A simple diagram for admission control

3) Based on the predicted MAI level, the base station calculates the available throughput $\hat{\psi}_{\text{av}}$ in every sliding observation window, using (2.26).

4) The base station makes an admission control decision using (2.25), that is, the user is admitted if (2.25) is satisfied. If not, the base station checks if the waiting time is within D, moves to the next time window if it is true, and starts Step 3 again. If the waiting time exceeds D, the request from this new user is dropped.

A simple diagram for the above scheme is shown in Figure 2.6.

In the following, we illustrate that the above scheme performs significantly better than the one using the mean value of the SINR (MAI) over one session. For convenience, we call the former Scheme A, and the latter Scheme B. Define the relative throughput prediction accuracy as

$$\eta = 1 - \frac{1}{N_\tau} \sum_{n=1}^{N_\tau} \frac{|\hat{\psi}_{\text{av},n} - \psi_{\text{av},n}|}{\psi_{\text{av},n}}, \tag{2.27}$$

where $\psi_{\text{av},n}$ is the actual available throughput in sliding window n, $\hat{\psi}_{\text{av},n}$ is the predicted one, and N_τ is the number of sliding windows. In this example, τ is 5 minutes and T_m is 2 seconds. Table 2.1 gives the performance of two prediction schemes. Several observations are in order: First, the prediction accuracy of Scheme A is always better than Scheme B;

α	1.1	1.3	1.5	1.7	1.9
Scheme A	82.7%	83.5%	86.1%	87.5%	88.3%
Scheme B	43.2%	43.7%	56.1%	59.2%	70.5%

Table 2.1. Prediction accuracy η versus α ($T_f = 0.01$s, $\xi = 0.99$, $f_m = 5$Hz)

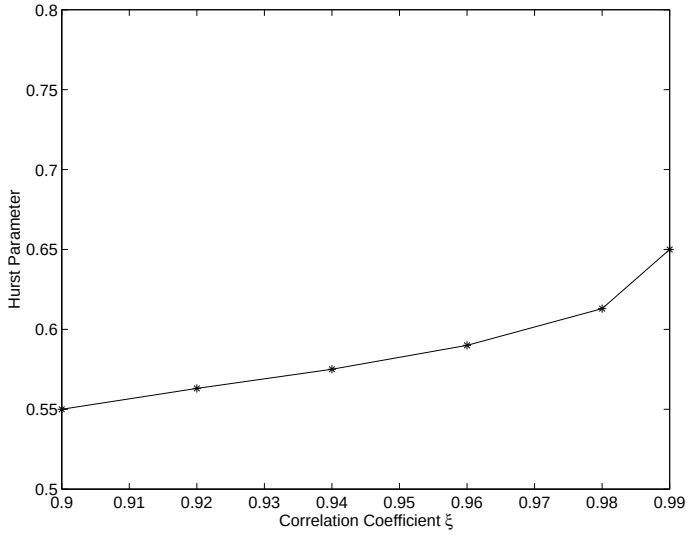

Figure 2.7. Hurst parameter versus correlation coefficient ξ ($T_f = 0.01$s, $f_m = 30$Hz)

second, the performance of Scheme B degrades dramatically with α, whereas Scheme A is more or less not sensitive to the degree of heavy tailedness. In a nutshell, the throughput prediction in Scheme A is significantly more accurate than that in Scheme B, and may be very useful for admission control for bursty data applications.

2.6. Impact of Fading and Traffic Burstiness

In the preceding sections, we have shown that the multiple time-scale MAI prediction can enhance system performance significantly. In the following, we seek to understand two important aspects related to the predictive MAI temporal structure: 1) What is the impact of fading? 2) How would the degree of traffic burstiness impact the system performance?

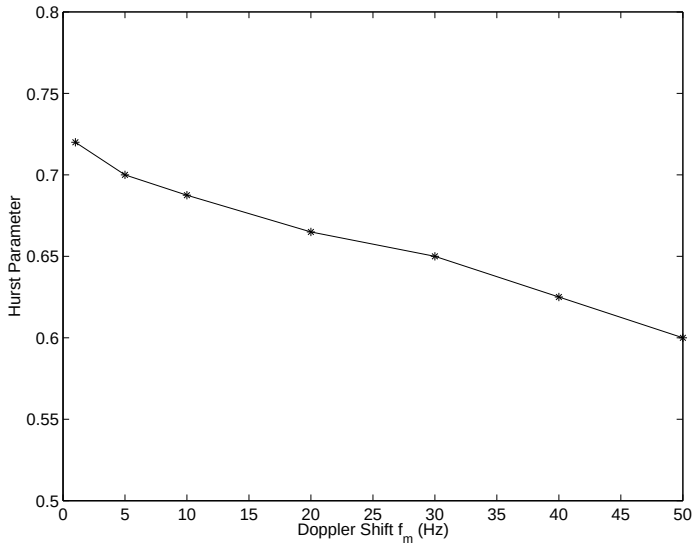

Figure 2.8. Hurst parameter versus Doppler shift f_m ($T_f = 0.01$s, $\xi = 0.99$, $f_m = 30$Hz)

To start with, we examine the impact of fading on the MAI long-range dependence. In the AR model for log-normal shadowing given in (2.2), ξ is the correlation coefficient of the channels, and T_f is the time granularity. For a fixed T_f, the smaller ξ is, the faster the fading is. In Rayleigh fading, the Doppler shift f_m governs the fading speed. Figure 2.7 shows that the Hurst parameter increases with the correlation coefficient ξ, and Figure 2.8 reveals that the Hurst parameter decreases with f_m. As is shown above, loosely speaking, the slower the shadowing is, the stronger the MAI long-range dependence is. Our intuition is as follows: Recall that the MAI is a superposition of many interference processes, each of which is a product of a fading process and an ON/OFF process. We would expect that both the fading process and the ON/OFF process affect the level of the MAI self-similarity. Then it is natural to expect that the degree of the long-range dependence lies in between that of the fading process and that of the ON/OFF process. Generally speaking, the shadowing fading is short-range dependent compared to ON/OFF processes of data traffic. Therefore, the Hurst parameter of the MAI process in a "slower" fading channel will be larger than

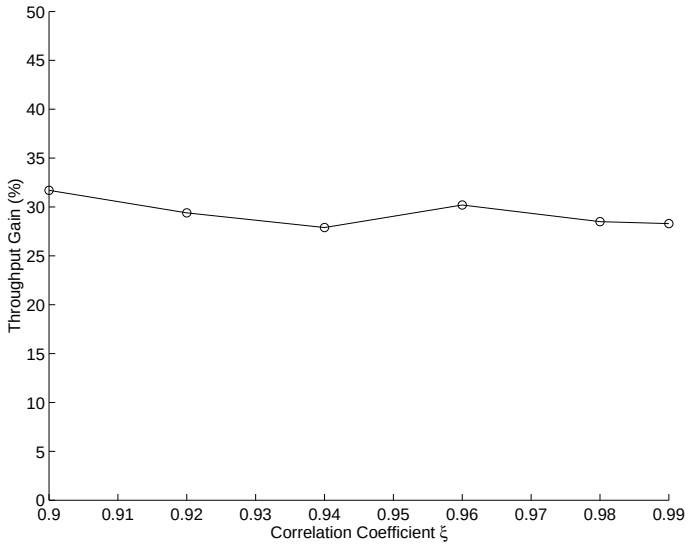

Figure 2.9. Rate adaptation: throughput gain versus ξ ($T_f = 0.01$s, $f_m = 30$Hz)

that in a "faster" fading channel, as is the cases in Figures 2.7 and 2.8. We also notice that the impact of Rayleigh fading is not as significant as the slow log-normal shadowing. This is because Rayleigh fading "modulates" the bursty traffic at smaller time scales than shadowing, and the short-range effect of Rayleigh fading can be averaged out at large time scales.

In Figures 2.9 and 2.10, we examine the impact of fading on rate adaptation. Figure 2.9 shows that rate adaptation is not sensitive to the correlation coefficient ξ of log-normal shadowing. This is because the transmission rate adaptation is fast enough to trace the changing of the SINR in slowly fading channels. In Figure 2.10, we observe that the throughput gain increases with the Doppler shift, and the rate adaptation scheme with a multiple time-scale MAI predictor can improve system performance significantly when f_m is large. The underlying rationale is as follows: When f_m increases, the accuracy of the packet-level prediction decreases. Furthermore, since the feedback delay is much smaller than T_m, its impact on the large time-scale prediction is negligible. Therefore, the rate adaption with a multiple

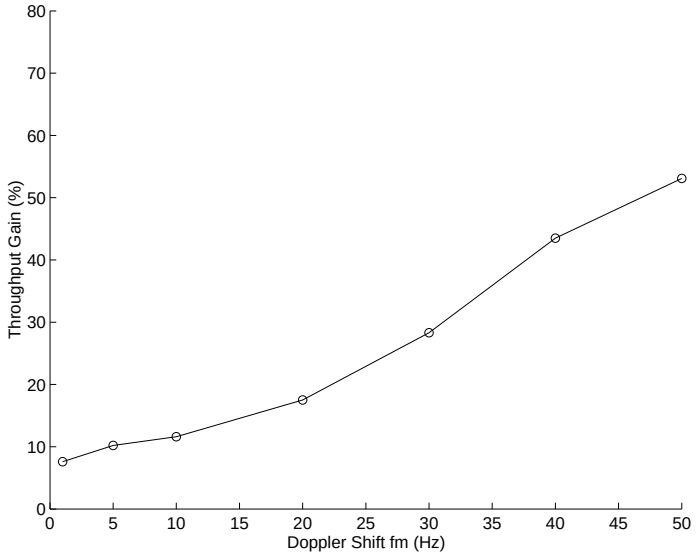

Figure 2.10. Rate adaptation: throughput gain versus Doppler shift f_m ($T_f = 0.01$s, $\xi = 0.99$)

time-scale predictor achieves significant throughput gain, when f_m becomes large.

To examine the impact of traffic burstiness, we assume for simplicity that $\alpha_{\text{on}} = \alpha_{\text{off}} = \alpha$. Figure 2.11 shows that the relative improvement will increase when α decreases. Our intuition is that the smaller α is, the more long-range dependent the MAI will be. This leads the large time-scale predictor to achieve higher prediction accuracy. As a result, the relative improvement increases.

Next, we examine the fading effect on the joint rate adaptation and admission control scheme. Figure 2.12 depicts the impact of Rayleigh fading on the scheme using the multiple time-scale predictor with $T_m = 2\,\text{s}$. Our results show that the prediction accuracy for admission control decreases with f_m. We also investigate the impact of traffic burstiness on the above joint rate adaptation and admission control scheme. Figure 2.13 shows that the smaller α is, the less the prediction accuracy is.

Worth noting is that we have also investigated the impact of fading and traffic bursti-

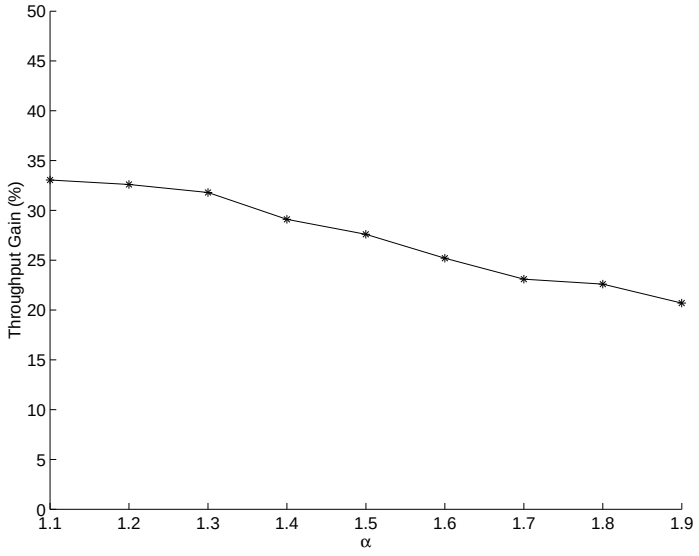

Figure 2.11. Rate adaptation: throughput gain versus α ($T_f = 0.01$s, $\xi = 0.99$, $f_m = 30$Hz)

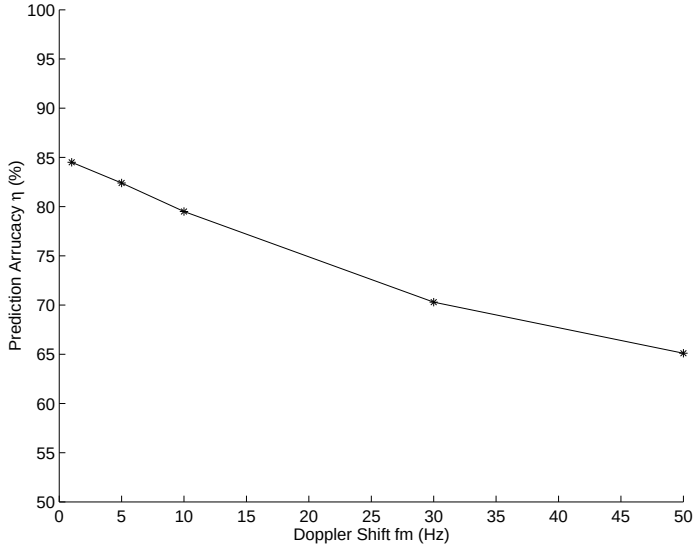

Figure 2.12. Admission control: prediction accuracy η versus Doppler shift f_m ($T_f = 0.01$s, $\xi = 0.99$)

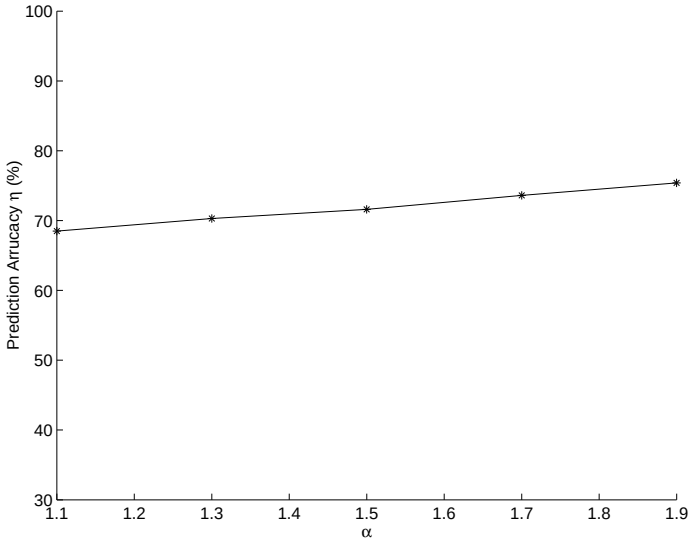

Figure 2.13. Admission control: prediction accuracy η versus α ($T_f = 0.01$s, $\xi = 0.99$, $f_m = 30$Hz)

ness on the performance, for both systems with power control and without power control in the downlink. It turns out that the above conclusions are applicable to both cases.

2.7. Conclusions

In this chapter, we first take a new approach to characterizing the MAI process in CDMA networks. This approach simultaneously takes into account time-varying channel conditions and the burstiness of data traffic, and opens a dimension to understand the MAI temporal correlation structure. Our findings show that the MAI process exhibits scale-invariant burstiness and is self-similar across multiple time scales, indicating the existence of a nontrivial predictive MAI structure at coarser time scales. This predictive MAI structure is then exploited for adaptive resource allocation to achieve efficient interference management, which is the key to achieving high spectral efficiency in CDMA systems.

Next, we utilize the predictive MAI structure to explore rate control and admission con-

trol. Specifically, we propose a multiple time-scale MAI predictor for interference sensing, built on which we devise a rate adaptation scheme. The rate adaptation uses a combination of spreading gain control and adaptive coding. Our result shows that rate control using the multiple time-scale MAI predictor performs better than that using the packet-level MAI predictor only, and can improve the system throughput significantly. Furthermore, we also exploit the MAI structure to improve admission control for data applications. In particular, we introduce a sliding observation window within which rate control is conducted, and the predicted throughput in each sliding window is used for admission control. Our results reveal that this admission control scheme may be very useful for bursty data applications. We also investigate the impact of fading and traffic burstiness on the system performance. Our results show that the rate adaptation using a multiple time-scale MAI predictor yields significant throughput gain in fast fading channels. The more bursty the traffic is, the more gain we can expect from rate control and admission control.

CHAPTER 3

Opportunistic Communications: Traffic Aided Smooth Admission Control and Opportunistic Scheduling

In wireless systems, the channel is typically error-prone due to fading. To improve the link quality, extensive wireless techniques have been proposed to combat fading. In contrast, *multiuser diversity* which makes use of (more or less independent) time-varying fading cross users has recently been revealed (see, e.g., [3], [7], [20], [53], [58], [59], [82], [94], [101]). Multiuser diversity originates from a cross-layer view. Roughly speaking, in a multiuser wireless network, at each time slot, with high probability there exists a user whose channel is boosted by constructive fading. By transmitting data to the "on-peak" user only, *opportunistic scheduling* can efficiently utilize the wireless resources and thus dramatically improve the overall system throughput (e.g., [46], [94]).

Fuelled by the rapid growth of Internet, there have been extensive demands for wireless data access. We note that data traffic exhibits high variation in terms of size and QoS requirements. In this chapter, we strive to find new methods in opportunistic communications to fulfill QoS provisioning, and improve the entire system performance. Thus motivated, we take cross-layer approaches to address the above issues.

We first study throughput provisioning, particularly admission control in opportunistic

communication systems. We note that the traditional trunk reservation policy of making "hard" admission decisions, is limited in two ways: 1) it is difficult to calculate explicitly the system capacity for opportunistic multi-access systems, since the capacity depends on not only channel conditions across all users, but also the specific scheduling scheme; and 2) the "hard" admission of new users may cause a sudden overload of the network, and degrade the QoS performance of active users. To resolve these problems, we propose a traffic aided smooth admission control (SAC) scheme. The basic idea of SAC is to increase gradually the amount of the time resource allocated to each new user over a trial period. Specifically, we first propose to use an adaptive resource allocation algorithm, namely QoS driven weight adaptation, for weighted proportional fair (WPF) opportunistic scheduling [50]. Building on this algorithm, we allocated more time resources to the new users by increasing their weights adaptively, while ensuring the throughput of original active users. Based on the observed throughput, an admission decision is made within a time-out window: the system admits an incoming user if its throughput is above the threshold; otherwise, the user drops out and requests access again after a back-off time. A key feature is that the back-off time is designed by using the traffic information.

We then explore a cross-layer approach to reducing the completion time by exploiting both the file size information and channel state information in a unified manner. In particular, we first examine two scheduling schemes, i.e., the wireless shortest remaining processing time first (W-SRPT, see also [81]) scheme and "riding on the channel peak" scheme, which can be viewed as two extreme cases utilizing the traffic information only or channel conditions only. Building on the insights gained from these two schemes, we then develop new traffic aided opportunistic scheduling (TAOS) schemes, including TAOS-1, TAOS-1a, TAOS-1b, and TAOS-2. To evaluate the performance, we also derive a lower

bound and an upper bound on the total completion time. The results show that the TAOS schemes can reduce the completion time of the entire system significantly.

3.1. Traffic Aided Smooth Admission Control

3.1.1. System Models

We consider the downlink, and focus on data transmissions in a time-slotted system, i.e., the transmissions of data users within a cell are time division multiplexed (TDM). More specifically, at each time slot, all users utilize pilot signaling to estimate the channels and feed the information back to the base station. Based on channel conditions and QoS requirements, an opportunistic scheduler is introduced to arrange the transmissions.

A. Throughput Requirement

Following [7], we assume that the QoS requirement of user k is given by

$$\overline{R}_k(t) \geq \eta_k, \tag{3.1}$$

where $\overline{R}_k(t)$ is the average throughput of user k within a given observation window T_k^c, and η_k is its throughput requirement threshold.

B. Channel Model

As is standard, we assume that the time-varying fading is due to distance-related attenuation, log-normal shadowing, and Rayleigh fading. Specifically, we use a multi-slope model [67] to calculate the propagation attenuation:

$$PL(\text{dB}) = \begin{cases} L_b + 20\log(\frac{d}{r_b}) & d \leq r_b \\ L_b + 40\log(\frac{d}{r_b}) & d > r_b, \end{cases} \tag{3.2}$$

with

$$\begin{cases} r_b = \frac{4h_b h_m}{\lambda} \\ L_b = |20\log(\frac{\lambda^2}{8\pi h_b h_m})|, \end{cases} \tag{3.3}$$

where λ is the carrier's wavelength, h_b is the height of the transmitter's antenna, and h_m is the height of the receiver's antenna. We also assume that if $d \leq r_b$, the standard deviation for log-normal fading σ_Ω is 4dB; otherwise, σ_Ω is 8dB. A filtered Gaussian noise (FGN) model [87] for Rayleigh fading is utilized in numerical experiments.

C. Weighted Proportional Fair Scheduling

In order to address QoS provisioning, we use a weighted proportional fair scheme [50] for downlink scheduling in a cellular network (which can be viewed as an extension of proportional fair opportunistic scheduling in [94], [101]). Suppose there are K users in an opportunistic communication system, and each user is assigned a weight according to its QoS requirement and channel condition. Two main steps in the weighted proportional fair opportunistic scheduling are outlined as follows.

WPF scheduling:

1) The scheduler (at the base station) transmits data to user k^* with the highest priority, i.e.,

$$k^* = \arg\max_k \left(w_k(t) \frac{R_k(t)}{\overline{R}_k(t)} \right), \tag{3.4}$$

where $w_k(t)$ is the weight assigned to user k, $R_k(t)$ is current data rate supportable by the channel state of user k, and $\overline{R}_k(t)$ is the average throughput.

2) The scheduler updates the average throughput with an exponentially weighted low-pass

filter (see, e.g., [94], [101]), i.e.,

$$\overline{R}_k(t+1) = \begin{cases} (1-\frac{1}{T^c})\overline{R}_k(t) + \frac{1}{T^c}R_k(t) & k = k^* \\ (1-\frac{1}{T^c})\overline{R}_k(t) & k \neq k^*, \end{cases} \tag{3.5}$$

where T^c is the observation window in terms of time slots.

3.1.2. Traffic Aided Smooth Admission Control

As mentioned before, the emerging multimedia applications in 3G and beyond typically have highly heterogeneous QoS requirements [62]. Therefore, it is of great interest address the QoS provisioning in opportunistic communications systems, particularly when admission control is involved.

Thus motivated, we propose an easy-to-implement traffic aided SAC scheme for opportunistic multiuser communications. The basic idea of the SAC scheme is to increase gradually the amount of the network resource allocated to each new user, so as to "spread" the admission decision over a trial period. Then, it is natural to ask: 1) what step size should be used to increase the amount of the resource for the requesting user; and 2) how can an admission decision be made? To this end, we incorporate the network traffic and channel variation into the admission control. Specifically, we first propose an adaptive resource allocation algorithm — QoS driven weight adaptation for WPF opportunistic scheduling. Building on this algorithm, we increase the time resources of the incoming users by adaptively raising their weights, while ensuring the throughput fulfillment of the original active users. Based on the measured throughput, an admission decision is made within a time-out window: the system admits this incoming user if its throughput is above the threshold; otherwise, the user drops out and requests admission again after a back-off time. In particular, we employ the information of backlogged traffic for the design of the back-off

time.

This traffic aided smooth admission control algorithm is reminiscent of distributed power control with active link protection (DPC/ALP) [11], which gradually powers up new links entering the channel. A key difference distinguishing our study from [11] is that DPC/ALP is for interference-limited wireless systems, whereas our SAC scheme is designed for opportunistic communication systems, and the increment step of the weight for each user is chosen adaptively using the information of traffic, fading, and throughput requirements.

A. Traffic Aided Admission Control

Suppose there are K users in the system. We classify the users into two sets: $A(t)$ denotes the set of (admitted) active users, and $B(t)$ denotes the set of (new) trial users. Roughly speaking, we can gradually allocate more resources to the trial users by increasing their weights in a guarded manner. Because the time resource of the network is fixed, the throughput requirements of all users may not be simultaneously satisfied. Therefore, we admit a new user if its throughput requirement is fulfilled within a trial period; otherwise, we let the "inadmissible" user drop out and request admission later.

The time-out window T_{out} is a key parameter for the admission decision. Intuitively speaking, the time-out window can not be too small or too large. If T_{out} is very small, the measurement may not accurately represent the average throughput, and therefore may cause some unnecessary drop-outs. If T_{out} is too large, the "inadmissible" user may keep staying in the system for a long time, and thus reduce other trial users' possibility of getting admission. Therefore, we select $T_{\text{out}} = \beta T^c$, $\beta \geq 1$, where T^c is the throughput observation window in opportunistic multi-access communications.

The back-off time T_{off} can be computed by using traffic information. Specifically, we

assume that the scheduler has the knowledge of the backlogged file size distribution F. Then, we can set $T_{\text{off}} = \xi\mu_F/\tilde{R}$, $\xi > 0$, where μ_F is the mean corresponding to the distribution F, and $\tilde{R}$ is the average throughput across active users. For example, we can choose $\xi = 1/K_A$, where K_A is the number of active users. The intuition is that after $\mu_F/(K_A\tilde{R})$, with high probability some active user will complete its session and exit the system, and thus there will be more "surplus" of time resources to accommodate new users.

B. Adaptive Resource Allocation in SAC

Adaptive resource allocation is a key element of the proposed traffic aided SAC scheme. In a WPF opportunistic multi-access system, the resource allocation is governed by the weight assignment. Hence, we devise a QoS driven weight adaptation scheme to address two key problems of our SAC scheme: 1) how to ensure the throughput of active users, and 2) how to increase the time resources allocated to trial users?

For the SAC scheme, it is important to obtain a clear understanding of the relationship between the weight and the throughput for one user. We observe that at steady states, the time resource assigned to user k is roughly proportional to $w_k/\sum_{i=1}^{K} w_i$ (see also [59], [94], [109]). Then, the average throughput of user k can be approximated by

$$\overline{R}_k \simeq \alpha_k \overline{C}_k \frac{w_k}{\sum_{i=1}^{K} w_i}, \ k = 1, \ldots, K, \tag{3.6}$$

where $\overline{C}_k$ is the average channel capacity of user k, and α_k is a constant. In a system with many users, each user's opportunistic transmission is near its channel peak with high probability. Then, α_k can be approximated to $R_k^p/\overline{C}_k$, where R_k^p denotes the peak transmission rate of user k.

Simply put, the proposed QoS driven weight adaptation is that if the throughput of one user is greater than its requirement threshold, its weight is decreased to "donate" some

surplus resources; if the opposite is true, its weight is increased to get more time resources. Since the departures and arrivals of users take place at coarse time scales, we assume that in each adaption window T^w ($T^w \geq T^c$), the system can achieve steady state performance. The weights are adapted as follows:

$$w_k(n+1) = w_k(n) - \Delta_k(n)\frac{\overline{R}_k(n) - \eta_k}{\overline{C}_k(n)}, \text{ for } k \in A \cup B, \tag{3.7}$$

where $\overline{R}_k(n)$ denotes the (steady state) average throughput in the nth adaptation window and $\Delta_k(n)$ (with a positive value) denotes the adaptation parameter. Moreover, to guarantee that the average throughput of each user monotonically increases with its weight, we constrain the summation of weights to be a constant, i.e.,

$$\sum_{k \in A \cup B} w_k(n) = \sum_{k \in A \cup B} w_k^0, \tag{3.8}$$

where w_k^0 is the nominal weight of user k, i.e., $w_k^0 = \eta_k / \overline{C}_k$.

Proposition 3.1.1 *If the steady state average throughput $\overline{R}_k(n) \geq \eta_k,\ k \in A$, for any $\Delta_k(n) \in \left(0,\ \sum_{i \in A \cup B} w_i^0/\alpha_k\right)$, $\overline{R}_k(n+1) \geq \eta_k$.*

Proof: For convenience, we denote $\sum_{i \in A \cup B} w_i^0$ as $\sum_i w_i^0$. Then,

$$\begin{aligned}
\overline{R}_k(n+1) &= \alpha_k \overline{C}_k \frac{w_k(n+1)}{\sum_i w_i^0} \\
&= \frac{\alpha_k \overline{C}_k}{\sum_i w_i^0}\left[w_k(n) - \Delta_k(n)\frac{\overline{R}_k(n) - \eta_k}{\overline{C}_k}\right] \\
&\geq \frac{\alpha_k \overline{C}_k}{\sum_i w_i^0}\left[w_k(n) - \frac{\sum_i w_i^0}{\alpha_k}\frac{\overline{R}_k(n) - \eta_k}{\overline{C}_k}\right].
\end{aligned}$$

Recall that $w_k(n) = \frac{\sum_i w_i^0 \cdot \overline{R}_k(n)}{\alpha_k \overline{C}_k}$. Therefore,

$$\begin{aligned}
\overline{R}_k(n+1) &\geq \alpha_k \overline{C}_k \left[\frac{\overline{R}_k(n)}{\alpha_k \overline{C}_k} - \frac{\overline{R}_k(n) - \eta_k}{\alpha_k \overline{C}_k}\right] \\
&= \eta_k,
\end{aligned}$$

completing the proof. ■

The above proposition shows that if we select $\Delta_k < \sum_{i\in A\cup B} w_i^0/\alpha_{\max}$ for $k \in A$, where $\alpha_{\max} = \max_i\{\alpha_i\}$, our weight adaption algorithm can still fulfill throughput requirements of active users. This nice property provides us a guideline in designing resource adaptation taking into account the departures and arrivals of users. That is, the system can ensure the throughput of active users by choosing a proper decreasing step for their weights, and meanwhile let trial users share the "surplus" weights.

In what follows, we elaborate the adaptive resource allocation algorithm in SAC. To simplify the presentation, we assume that at each slot, only one user arrives or departs from the system. Our resource adaptation scheme can be specified as follows.

Adaptive resource allocation algorithm:

1) The base station picks the user with the highest priority $w_k(t)R_k(t)/\overline{R}_k(t)$, and schedules transmission for it.

2) The base station updates the average throughput of all users according to (3.5).

3) The base station adapts the weights. At the jth slot, for active user k in set $A(j)$, the base station decreases the weight $w_k(j)$ to "donate" part of the "surplus" network resources, when the average throughput of user k is greater than the requirement threshold. For trial users in set $B(j)$, the base station gradually increases their weights and lets them share the "donation". Particularly, we let all active users have the same adjustment parameter $\delta_A(j)$ and all trial users have the same $\delta_B(j)$, and keep the summation of weights invariant. More specifically, we choose a proper step size $\delta_A(j) = \Delta/T^w,\ \Delta < \sum_{i\in A\cup B} w_i^0/\alpha_{\max}$, and calculate $\delta_B(j)$ according to (3.9),

$$\delta_A(j)\sum_{k\in A}\frac{\overline{R}_k(j)-\eta_k}{\overline{C}_k(j)} = -\delta_B(j)\sum_{k\in B}\frac{\overline{R}_k(j)-\eta_k}{\overline{C}_k(j)}. \tag{3.9}$$

Then, the base station adjust weights using

$$w_k(j+1) = \begin{cases} w_k(j) - \delta_A(j)\frac{\overline{R}_k(j)-\eta_k}{\overline{C}_k(j)}, & k \in A \\ w_k(j) + \delta_B(j)\frac{\eta_k-\overline{R}_k(j)}{\overline{C}_k(j)}, & k \in B. \end{cases} \tag{3.10}$$

4) The base station deals with the dynamics of users. If a trial user q enters the system, i.e, $B(j+1) = B(j) \cup \{q\}$, the scheduler assigns an initial weight $w_{q,\min}$ to this user, and scales up weights of all other users:

$$w_k(j+1) = w_k(j)\frac{\zeta + (w_q^0 - w_{q,\min})}{\zeta},\ k \in A \cup B, \tag{3.11}$$

where $w_{q,\min}$ is the minimal weight of trial user q, w_q^0 is its nominal weight, and $\zeta = \sum_{i \neq q} w_i^0$. If user q exits the system, the base station deletes its registration in either A or B and scales the weights of all other users as

$$w_k(j+1) = w_k(j)\frac{\zeta}{\zeta + (w_q^0 - w_q(j))},\ k \in A \cup B. \tag{3.12}$$

Remark: In the adaptive resource allocation for SAC, the weights are updated every time slot. Therefore, we make an approximation, i.e., spread Δ to the adaptation window T^w, and thus fulfill the throughput requirements of active users in a guarded manner.

3.1.3. Numerical Examples

In this section, we investigate the dynamics of the SAC scheme. Generally speaking, due to the limitation of network resources, some new users may succeed in receiving admission, while others may saturate at certain throughput level (below the thresholds) for extensive periods. In the following, we illustrate the performance of the proposed SAC scheme for opportunistic multi-access via numerical experiments.

Suppose that the average SNR on the cell boundary SNR_b is 0 dB, the maximal Doppler shift f_m is 10 Hz, and all users are uniformly scattered within a cell. We assume that all

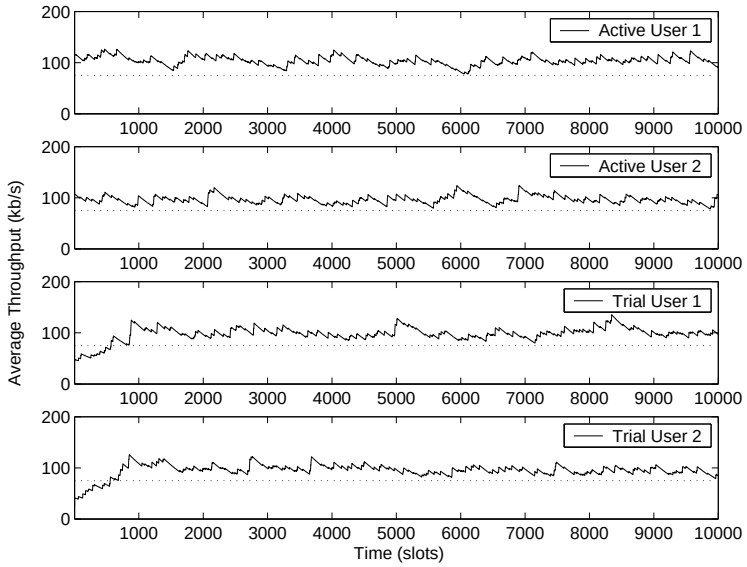

Figure 3.1. Throughput evolution for fully admissible users

transmissions have the same observation and adaptation window, specifically $T^c = T^w$ consists of 1000 slots, and the average throughput threshold η =75kb/s. Figures 3.1, 3.2, and 3.3 depict the evolution of average throughput for three cases: 1) fully admissible, 2) totally inadmissible, and 3) partially admissible with drop-out. In each figure, we plot the average throughput of four users: the active user with the smallest average channel capacity, the active user with the largest average channel capacity, and two trial users. Comparing these figures, we can see that our SAC scheme can always ensure the throughput of original active users, while trying to admit new users. Also, the drop-out of the inadmissible user can increase other users' share of the time resource, and thus makes it possible to admit other trial users.

In Figure 3.4, we utilize a key metric — average admission delay, to evaluate the SAC scheme. We compare the performance of two schemes: Scheme 1 is without drop-out, and Scheme 2 employs a time-out window $T_{\text{out}} = 1.2T^c$ and back-off time $T_{\text{off}} = \mu_F/(K_A\tilde{R})$. We use the bulk-data ftp transmission in this example. Along the line of [71], we assume

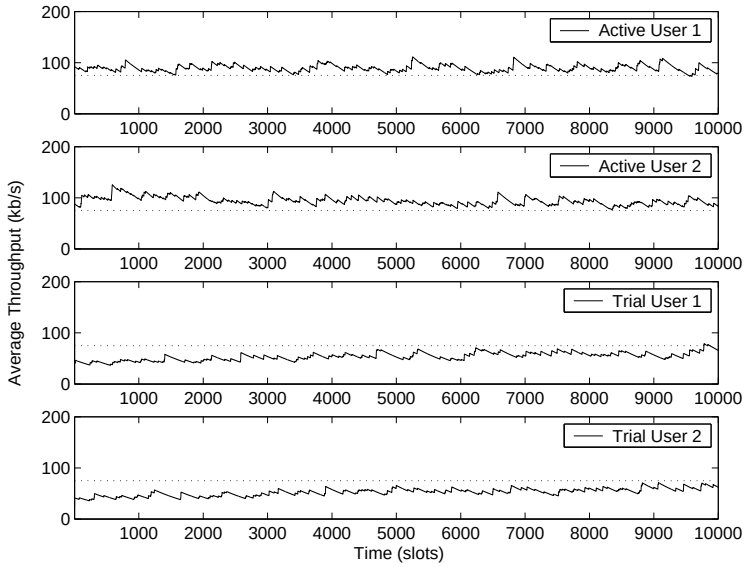

Figure 3.2. Throughput evolution for inadmissible users

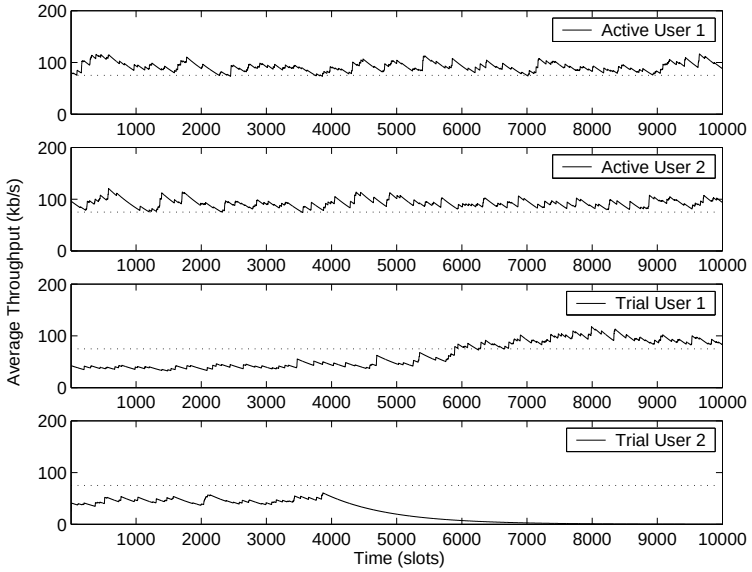

Figure 3.3. Throughput evolution for partially admissible users using drop-out mechanism

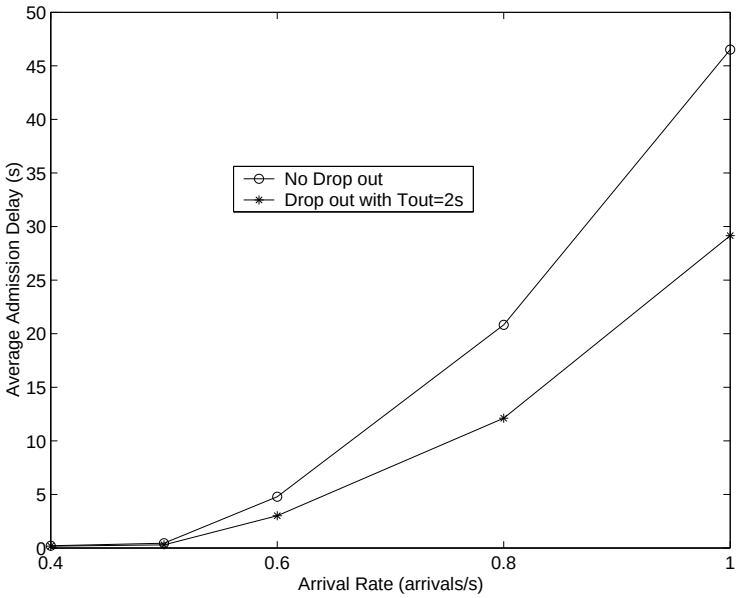

Figure 3.4. Average admission delay versus arrival rate

that the backlogged ftp file size is log-normally distributed with the mean $\mu_F = 500$kB. We note that in Scheme 1, as the arrival rate increases, the average admission delay rises dramatically, indicating that the total system gets congested quickly. Intuitively, without drop-out, the more requesting users come into the system, the more trial users compete for the network resources, the more difficult for each trial user to get sufficient resource to meet its throughput requirement, resulting in heavy congestion. Meanwhile, with a time-out window and a proper back-off time, the competition is scattered. Thus at each moment, there are fewer trial users. Therefore, the average admission delay can remain at a lower level.

3.2. Traffic Aided Opportunistic Scheduling

Recent years have witnessed a tremendous growth in the demand for ubiquitous information access. Intense demands in wireless networks can induce possibly large delay, and degrade the system performance. Therefore, it is of great interest to investigate the problem of minimizing the total (or average) *completion time* which consists of both processing time and waiting time (delay) [34].

In wireline networks, the problem of reducing the completion time has been studied extensively. A general review on wireline scheduling algorithms can be found in [48]. In [81], a "shortest remaining processing time first (SRPT)" scheduling scheme is presented. In recent studies [25], [34] (see also the references therein), the authors develop traffic size based scheduling schemes for connection control at Web servers, and show that these schemes can yield significant reduction in the average completion time. Simply put, by properly exploiting the file size information in scheduling (e.g., picking the user with the shortest file), the overall system performance can be improved.

In wireless networks, opportunistic scheduling (see, e.g., [3], [7], [19], [20], [53], [58], [59], [82], [83], [94], [101]) offers a promising solution to reducing the completion time. Opportunistic scheduling originates from a holistic view. Roughly speaking, in a multiuser wireless network, at each moment it is likely there exists a user with "good" channel conditions. By picking the instantaneous "on-peak" user for data transmission, opportunistic scheduling can utilize the wireless resource efficiently and thus improve the overall system throughput dramatically. Hence, the total completion time can be shortened, due to the enhancement of the system throughput.

We note that "picking the user with the shortest file" and "picking the user on the

channel peak" may not coincide. As a result, these two approaches may lead to conflicting scheduling. Then, a natural question to ask is "what is the optimal scheme in the sense of minimizing the total completion time?" It is clear that the completion time is correlated across the users, i.e., one user's processing time is related to the delay of the other users. The tight coupling among the transmissions of different users, together with the channel variation in wireless networks, makes the task more challenging.

In this section, we take a cross-layer design approach to reducing the completion time by exploiting both the file size information and channel state information in a unified manner. Our contributions can be summarized as follows. We first establish general convexity properties for opportunistic scheduling with file size information, which provides a basis for devising scheduling schemes (see also [42]). Then, building on the insights gained from two existing scheduling schemes, namely the wireless shortest remaining processing time first (W-SRPT) scheme and "riding on the channel peak" scheme, we develop new traffic aided opportunistic scheduling (TAOS) schemes, including TAOS-1, TAOS-1a, TAOS-1b, and TAOS-2. Roughly speaking, the TAOS-1 scheme can be viewed as a generalization of the well-known "proportional fair scheduling" by taking into account the file size information; and the TAOS-2 scheme is a locally optimal scheme (we will elaborate on this further in Section 3.2). The third key contribution consists of a lower bound and an upper bound on the total completion time, which serve as benchmarks for examining the performance of the TAOS schemes. The results show that the TAOS schemes can reduce the completion time of the entire system significantly. We then extend the study to the cases with random arrivals and departures. As expected, the proposed TAOS schemes yield significant reduction of the total completion time when the arrival rate is high.

In related work, a framework for scheduling in wireless networks is presented in [58].

The authors of [47] study scheduling in power-controlled CDMA data networks, assuming that the channel remains static during the entire transmission of each user. An interesting work by Tsybakov studies file transmission over wireless fading channels in [95]. More specifically, a recursive method using dynamic programming is used to compute the expected completion time, based on which optimal and suboptimal scheduling algorithms are devised accordingly. One key difference between [95] and our study here lies in the fact that [95] considers the average performance and focuses on the expected completion time, whereas we are interested in an (arbitrary) "sample" case and study the completion time for a given set of files. Needless to say, different performance metrics used would lead to different scheduling strategies. A recent work [83] proposes a Foreground-Background (FB) scheduling algorithm for the heavy-tailed data traffic, and develops a new hierarchical approach using both proportional fair opportunistic scheduling ([94], [101]) and FB scheduling. The authors of [7] have investigated the scheduling for possible simultaneous transmissions, taking into account both channel conditions and delay constraints.

3.2.1. Background

We consider the downlink and assume that the transmissions of data users within a cell are time division multiplexed (TDM). Simply put, at each time slot, all users utilize pilot signaling to estimate the channel condition and feed the information back to the base station. Based on the channel conditions and file size information, an opportunistic scheduler at the base station arranges the transmissions, aiming to minimize the completion time of the whole system.

For simplicity, we consider the transmission of one backlogged file for every user (cf. [103]). (In the case when there are multiple backlogged files for one user, we assume that

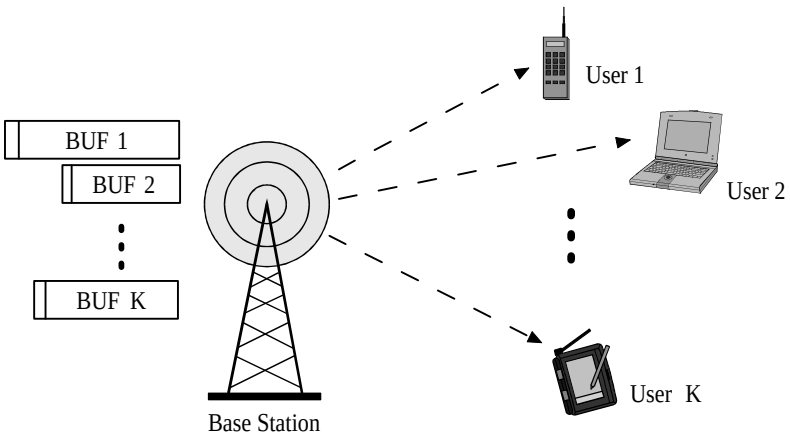

Figure 3.1. Downlink transmissions in a cellular system

this user finishes one file before transmitting the next. We will address the case with random arrivals and departures in Section C.) In this context, we use "file" and "user" interchangeably. Without loss of generality, we call the user with the shortest initial file size, user 1, and so on, i.e., $F_k = F_{(k)}$, where F_k denotes the initial file size of user k, and $F_{(k)}$ denotes the kth order statistic of the initial file sizes. A simple diagram for this system is given in Fig. 3.1.

A. Channel Model

In a wireless channel, the received signal can be expressed as

$$y = gs + w, \tag{3.13}$$

where s denotes the transmitted signal, g denotes the channel gain (which can be time-varying), and w denotes the white Gaussian noise. We assume that the time-varying fading is due to distance-related attenuation and small-scale fading. Specifically, we consider two kinds of fading channels: Rayleigh fading and Rician fading [87]. Let $R(t)$ denote the instantaneous data rate supported by the fading channel at time slot t. For simplicity, we

assume ideal coding, and thereby achievable channel capacity.

B. Completion Time

As is standard [34], we define the completion time of file k, denoted ψ_k, as the duration from the moment the file arrives at the system to the moment the file departs the system completely. Accordingly, the total (system) completion time is defined as

$$\Psi = \sum_{k=1}^{K} \psi_k, \tag{3.14}$$

where K is the number of files. It is worth pointing out that in many practical wireless systems, the transmission time of each file is on the order of seconds or even minutes, which is much greater than the slot duration and the coherent time of fading; and this is an assumption we impose throughout.

C. Wireless SRPT and "Riding on the Channel Peak"

In this section, we aim to reduce the total completion time by making use of file size information and channel state information in a unified manner. In what follows, we first recapitulate two scheduling schemes, i.e., the wireless SRPT (see also [81]) scheme and "riding on the channel peak" scheme, which can be viewed as two extreme cases utilizing the traffic information only or channel conditions only.

Wireless SRPT: Suppose that the scheduler at the base station utilizes only the knowledge of the average data rates and file size information of mobile users. The wireless SRPT scheduler picks the user with the shortest (*expected*) remaining processing time (see also [48], [81]), i.e.,

$$k^* = \arg\min_k \Big(\frac{X_k(t)}{E[R_k]}\Big), \tag{3.15}$$

where $X_k(t)$ denotes the residual backlogged file size of user k at time t, and $E[R_k]$ denotes the average data rate of user k. The corresponding total completion time is presented in [42].

"Riding on the Channel Peak": If the scheduler utilizes only the knowledge of the instantaneous channel conditions, "riding on the channel peak" is then a plausible approach to improving the total system throughput, thereby reducing the completion time. Specifically, at each time slot, the "riding on the peak" scheduler (see, e.g., [20], [35], [59], [94]) arranges the transmission for user k^* with

$$k^* = \arg\max_k \Big(\frac{R_k(t)}{E[R_k]}\Big), \tag{3.16}$$

where $R_k(t)$ denotes the instantaneous data rate of user k.

3.2.2. Traffic Aided Opportunistic Scheduling

Given a set of users, our goal is to minimize the total completion time. As noted in the Introduction, compared with the simple Round-Robin algorithm, both "picking the user with the shortest file" and "picking the user on the channel peak" can lead to reduced completion time. Then, it is natural to expect that we can improve the performance further by exploiting both the file size information and channel variation. Thus motivated, we devise traffic aided opportunistic scheduling schemes.

A. Convexity Properties of Scheduling Schemes

We start with characterizing general properties for opportunistic scheduling with file size information. Let $X_k(t)$ and $R_k(t)$ denote the backlog size and instantaneous data rate of user k at time slot t, respectively. ($X_k(0) = F_k$ is the initial file size of user k.) Suppose

that allocating partial time resource within one slot is applicable. Let $\gamma_k(t)$ denote the portion of time resource assigned to user k at time slot t, and define

$$\gamma(t) \triangleq (\gamma_1(t), \cdots, \gamma_K(t))^T, \tag{3.17}$$

where $(\cdot)^T$ denote the vector transpose. We assume that $R_k(t)$ and $\gamma(t)$ do not change within each slot. By neglecting the edge effect, we define the completion time of user k corresponding to a scheduling function $\gamma(t)$ as

$$\psi_k(\gamma) \triangleq \inf\Big\{t : \sum_{n=1}^{t} R_k(n)\gamma_k(n)T_s \geq X_k(0)\Big\}, \tag{3.18}$$

where T_s denotes the slot duration, and t is the time in terms of slots. Then, a scheduling function $\gamma(t)$ is said to be admissible if for $k = 1, \cdots, K$,

$$0 \leq \gamma_k(t) \leq 1, \tag{3.19}$$

$$\sum_{k=1}^{K} \gamma_k(t) \leq 1, \tag{3.20}$$

$$\gamma_k(t) = 0 \text{ for all } t \geq \psi_k(\gamma). \tag{3.21}$$

Note that

$$\sum_{n=1}^{\psi_k(\gamma)} R_k(n)\gamma_k(n)T_s = X_k(0). \tag{3.22}$$

Let Γ denote the set of all admissible scheduling functions. We have the following propositions on the general properties of opportunistic scheduling with file size information.

Proposition 3.2.1 *Γ is a convex set. Moreover, for any $\gamma, \gamma' \in \Gamma$ and $\theta \in (0, 1)$,*

$$\psi_k\big(\theta\gamma + (1-\theta)\gamma'\big) = \max\big(\psi_k(\gamma), \psi_k(\gamma')\big), k = 1, \cdots, K. \tag{3.23}$$

Proof: It is straightforward to see that Γ is a convex set. Without loss of generality,

suppose $\psi_k(\gamma) \leq \psi_k(\gamma')$. Then, for $0 < \theta < 1$,

$$\begin{aligned}
&\sum_{t=1}^{\psi_k(\gamma')} R_k(t)\Big(\theta\gamma_k(t) + (1-\theta)\gamma'_k(t)\Big)T_s \\
&= \theta \sum_{t=1}^{\psi_k(\gamma')} R_k(t)\gamma_k(t)T_s + (1-\theta)\sum_{t=1}^{\psi_k(\gamma')} R_k(t)\gamma'_k(t)T_s \\
&\overset{(a)}{=} \theta X_k(0) + (1-\theta)X_k(0) \\
&= X_k(0),
\end{aligned} \tag{3.24}$$

where (a) follows from that fact that $\psi_k(\gamma) \leq \psi_k(\gamma')$ and $\gamma_k(t) = 0$ for $t > \psi_k(\gamma)$, which indicates that

$$\sum_{t=1}^{\psi_k(\gamma')} R_k(t)\gamma_k(t)T_s = \sum_{t=1}^{\psi_k(\gamma)} R_k(t)\gamma_k(t)T_s = X_k(0). \tag{3.25}$$

We define

$$\psi_k(\theta) \triangleq \psi_k\big(\theta\gamma(t) + (1-\theta)\gamma'(t)\big).$$

Moreover, if $\psi_k(\theta) < \psi_k(\gamma')$, then

$$\begin{aligned}
&\sum_{t=1}^{\psi_k(\theta)} R_k(t)\Big(\theta\gamma_k(t) + (1-\theta)\gamma'_k(t)\Big)T_s \\
&= \theta \sum_{t=1}^{\psi_k(\theta)} R_k(t)\gamma_k(t)T_s + (1-\theta)\sum_{t=1}^{\psi_k(\theta)} R_k(t)\gamma'_k(t)T_s \\
&= \theta \sum_{t=1}^{\psi_k(\gamma')} R_k(t)\gamma_k(t)T_s - \theta \sum_{t=\psi_k(\theta)+1}^{\psi_k(\gamma')} R_k(t)\gamma_k(t)T_s \\
&\quad + (1-\theta)\sum_{t=1}^{\psi_k(\gamma')} R_k(t)\gamma'_k(t)T_s - (1-\theta)\sum_{t=\psi_k(\theta)+1}^{\psi_k(\gamma')} R_k(t)\gamma'_k(t)T_s.
\end{aligned} \tag{3.26}$$

Since

$$\sum_{t=\psi_k(\theta)+1}^{\psi_k(\gamma')} R_k(t)\gamma_k(t)T_s \geq 0,$$

and

$$\sum_{t=\psi_k(\theta)+1}^{\psi_k(\gamma')} R_k(t)\gamma'_k(t)T_s > 0,$$

we have that

$$\begin{aligned}
&\sum_{t=1}^{\psi_k(\theta)} R_k(t)\Big(\theta\gamma_k(t) + (1-\theta)\gamma'_k(t)\Big)T_s \\
&< \theta \sum_{t=1}^{\psi_k(\gamma')} R_k(t)\gamma_k(t)T_s + (1-\theta)\sum_{t=1}^{\psi_k(\gamma')} R_k(t)\gamma'_k(t)T_s \\
&= X_k(0),
\end{aligned} \tag{3.27}$$

which contradicts (3.22). Therefore, for $k = 1, \cdots, K$,

$$\psi_k\big(\theta\gamma + (1-\theta)\gamma'\big) = \max\big(\psi_k(\gamma), \psi_k(\gamma')\big). \tag{3.28}$$

■

Proposition 3.2.2 *Define* $\Psi(\gamma) \triangleq \sum_{k=1}^{K} \psi_k(\gamma)$. *Then,* $\Psi(\gamma)$ *is a concave function on* Γ.

Proof: For any $\theta \in (0,1)$,

$$\begin{aligned}
\Psi(\theta\gamma + (1-\theta)\gamma') &= \sum_{k=1}^{K} \psi_k(\theta\gamma + (1-\theta)\gamma') \\
&= \sum_{k=1}^{K} \max\Big\{\psi_k(\gamma), \psi_k(\gamma')\Big\} \\
&\geq \sum_{k=1}^{K} \Big\{\theta\psi_k(\gamma) + (1-\theta)\psi_k(\gamma')\Big\} \\
&= \theta\Psi(\gamma) + (1-\theta)\Psi(\gamma').
\end{aligned} \tag{3.29}$$

Hence, $\Psi(\gamma)$ is a concave function on Γ. ■

Propositions 3.2.1 and 3.2.2 reveal that the optimal solution γ^* is an extreme point of Γ, i.e., $\gamma_k^*(t) = 0$ or 1, for $k = 1, \cdots, K$. Roughly speaking, assigning the entire time slot to one user while keeping others silent can lead to shorter processing time.

We note that the globally optimal scheduling scheme is highly non-trivial. In what follows, we present several suboptimal opportunistic scheduling schemes, and derive lower bounds and upper bounds on the total completion time. Specifically, building on the insights

gained from the two extreme cases studied in Section C, we develop traffic aided opportunistic scheduling schemes, namely TAOS-1, TAOS-1a, TAOS-1b, and TAOS-2, which inherit the merits of both algorithms above.

B. Traffic Aided Opportunistic Scheduling

TAOS-1: Clearly, giving high priority to users with either small file sizes or "good" channel conditions would yield reduction in the total completion time. Thus, it is plausible to devise a cost function (or priority function), which increases with the file size and decreases with the instantaneous data rate, and develop cost function based scheduling accordingly. To provide fairness (see [94], [101]), we can also take into account the average throughput in designing the function. Along this line, we construct the cost function as $\frac{F_k U_k(t)}{R_k(t)}$, where F_k denotes the initial backlogged file size of user k, and U_k is the average throughput of user k. The corresponding scheduling scheme, TAOS-1, picks user k^* with

$$k^* = \arg\min_k \Big(\frac{F_k U_k(t)}{R_k(t)} \Big). \tag{3.30}$$

The average throughput U_k can be calculated as (see [94], [101])

$$U_k(t+1) = \begin{cases} (1 - \frac{1}{T^c})U_k(t) + \frac{1}{T^c} R_k(t) & k = k^* \\ (1 - \frac{1}{T^c})U_k(t) & k \neq k^*, \end{cases} \tag{3.31}$$

where T^c is the sliding observation window in terms of the number of slots.

Building on the TAOS-1 scheme above, we have also derived two other suboptimal schemes: TAOS-1a and TAOS-1b. TAOS-1a does not consider the fairness, and the corresponding cost function is simply $\frac{F_k}{R_k(t)}$. That is, the TAOS-1a scheme schedules the transmission for user k^* with

$$k^* = \arg\min_k \Big(\frac{F_k}{R_k(t)} \Big). \tag{3.32}$$

TAOS-1b takes into account the dynamics of the file size by replacing F_k in (3.32) with $X_k(t)$. Then, we obtain a scheduling scheme totally relying on the instantaneous information at time t. More specifically, the TAOS-1b scheme picks user k^* with

$$k^* = \arg\min_k \Big(\frac{X_k(t)}{R_k(t)}\Big). \tag{3.33}$$

TAOS-2: In the above heuristic schemes, the cost functions cannot characterize the completion time. In what follows, we develop an opportunistic scheduling scheme, via devising a cost function directly related to the completion time. Observe that wireless SRPT scheduling utilizes only the average data rate, not the instantaneous channel state information. Then, if the instantaneous channel information is incorporated into wireless SRPT, the scheduling scheme can exploit the dynamics of file size and channel in a more integrated manner, thereby achieving more reduction in the total completion time. Thus motivated, we devise a new TAOS scheme which evolves in two phases. Specifically, the TAOS-2 algorithm can be outlined as follows:

Phase I:

i) Sort all users in the ascending order of $\frac{X_k(t)}{E[R_k]}$, and let $I_k(t)$ denote the rank of user k among the ordered variates.

Phase II:

ii) Compute the cost function for user k as

$$D_k(t) = \big(I_k(t) - 1\big) - \big(M(t) - I_k(t) + 1\big)\Big(\frac{R_k(t)}{E[R_k]} - 1\Big), \tag{3.34}$$

where $M(t)$ denotes the number of remaining users in the system.

iii) Schedule the transmission for user k^* which has the smallest cost function, i.e.,

$$k^* = \arg\min_k \Big(D_k(t)\Big). \tag{3.35}$$

Worth noting is that this constructive opportunistic scheduling scheme is locally optimal, and we also call it locally optimal TAOS (LO-TAOS). In the following, we elaborate further on this.

Consider the scheduling at the nth slot. Observe that when only the average data rate $\{E[R_i],\ i = 1, \cdots, K\}$ and remaining file size $\{X_i(n),\ i = 1, \cdots, K\}$ are utilized, wireless SRPT would be the best scheduling scheme (see also [48]). We adopt the wireless SRPT criterion in Phase I of the TAOS-2 scheme, and let $Z_0(n)$ denote the corresponding expected system time for the remaining files. In Phase II, building on the rank $\{I_i(n), i = 1, \cdots, K\}$ obtained in Phase I, the scheduler optimizes the scheduling by exploiting the knowledge of the instantaneous data rate $\{R_i(n),\ i = 1, \cdots, K\}$. (Note that $\{R_i(t),\ t \geq n+1\}$ is not available at the nth slot.) If the scheduler picks user k for the transmission, the corresponding expected remaining time in term of slots can be expressed as

$$Z_k(n) = Z_0(n) + D_k(n) + \epsilon_k(n), \tag{3.36}$$

where $D_k(n)$ is given in (3.34) and $\epsilon_k(n)$ denotes the correction term if the realization of $\{I_i(n+1), i = 1, \cdots, K\}$ is different from $\{I_i(n), i = 1, \cdots, K\}$ derived in Phase I. Specifically, $\epsilon_k(n)$ can be expressed as

$$\epsilon_k(n) = Y_k(n+1) - Y_0(n+1), \tag{3.37}$$

where $Y_k(n+1)$ denotes the expected remaining time if the system transmits the remaining files ($t \geq n+1$) according to the order given by $\{I_i(n+1), i = 1, \ldots, K\}$, and $Y_0(n+1)$ denotes the expected remaining time if the system transmits the remaining files ($t \geq n+1$) according to the order of $\{I_i(n), i = 1, \ldots, K\}$. We note that the impact of $\epsilon_k(n)$ is negligible, because the scheduler updates the rank $\{I_i(n), i = 1, \ldots, K\}$ in each slot and the changing rate of

the ranks is much slower. Therefore, (3.36) can be approximated as

$$Z_k(n) \cong Z_0(n) + D_k(n). \tag{3.38}$$

Note that compared with $Z_0(n)$ in Phase I, the first term of $D_k(n)$ in (3.34), $I_k(t) - 1$, denotes the "cost" experienced by users whose rank is smaller than $I_k(n)$, and the second term, $\big(M(t) - I_k(t) + 1\big)\Big(\frac{R_k(t)}{E[R_k]} - 1\Big)$, represents the "saving" of the other users. Then, by scheduling the transmission for user k^* with the smallest D_k:

$$k^* = \arg\min_k \Big(D_k(n)\Big),$$

the expected remaining time (and thus the expected total completion time) is minimized "locally".

3.2.3. Performance Bounds on the Total Completion Time

As noted in the Introduction, different users experience different channels, and the completion time is tightly coupled across the users. This makes it difficult (if not impossible) to find a closed-form solution for the total completion time. Thus, we turn to derive lower bounds and upper bounds on the total completion time. In particular, we first examine a hypothetical case shown in Fig. 3.2, and characterize the corresponding (optimal) total completion time; this serves as a lower bound of all scheduling schemes (we note that this lower bound is not achievable). Next, we derive an upper bound on the total completion time corresponding to the "riding on the channel peak" scheme. Since the TAOS schemes perform better than "riding on the channel peak", this bound is an upper bound on the completion time corresponding to the TAOS schemes.

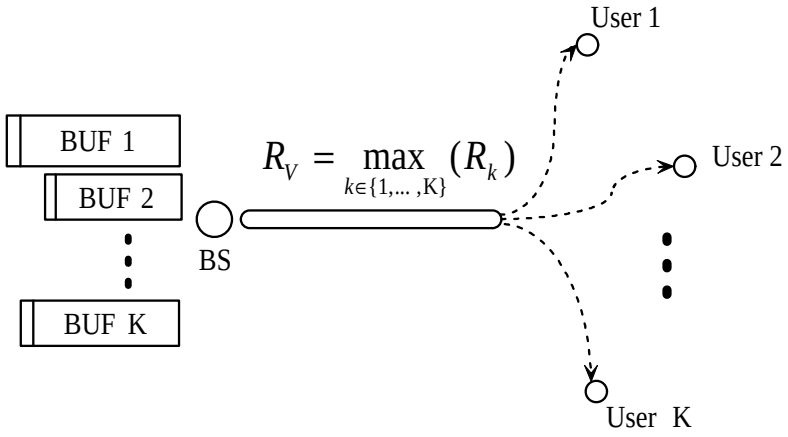

Figure 3.2. A hypothetical channel model for the lower bound

A. A Lower Bound for TAOS

In what follows, we derive a lower bound on the total completion time. Observe that the data rate between the base station and the user scheduled for transmission can not exceed R_{V}, where

$$R_{\mathrm{V}} \triangleq \max_{k\in\{1,\cdots,K\}} (R_k). \tag{3.39}$$

(Note $R_{\mathrm{V}}(t) = R_{(K)}(t)$, where $R_{(K)}(t)$ is the Kth order statistic of the instantaneous channel capacities of all K users.) We consider a hypothetical case in Fig. 3.2, where the base station transmits all K files by using R_{V}, regardless of the destination (user). Accordingly, the optimal scheduling scheme in this hypothetical context would be wireless SRPT, which serves as a lower bound. Hence, the lower bound of the total completion time is given by

$$\Psi_{\mathrm{LB}} = \frac{1}{E\big[R_{\mathrm{V}}\big]} \sum_{k=1}^{K} (K-k+1)F_k. \tag{3.40}$$

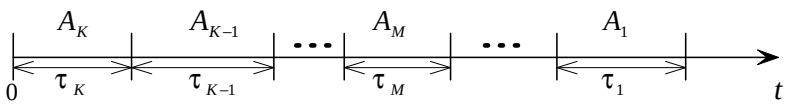

Figure 3.3. Transmission dynamics corresponding to the "riding on the channel peak" scheme

With the knowledge of file size distributions, we can also examine the average-case performance. Specifically, the expectation of the total completion time is

$$E\big[\Psi_{\mathrm{LB}}\big] = \frac{1}{E\big[R_{\mathrm{V}}\big]} \sum_{k=1}^{K} (K-k+1) E\big[F_k\big]. \tag{3.41}$$

where $E[F_k]$ can be derived by using results on order statistics [26]. Worth pointing out is that this lower bound is not achievable.

B. An Upper Bound for TAOS

Now, we examine "riding on the channel peak", and thus find an upper bound for TAOS. Under the assumption that the mobile users have i.i.d. (statistically symmetric) channels, the "riding on the peak" scheduling criterion can be written in the form of

$$k^* = \arg\max_k \Big(R_k(t) \Big). \tag{3.42}$$

Define $R_{\max} = \max_{k \in B}(R_k)$, where B denote the set of active users remaining in the system. In a system with many users, $R_{\max}$ can be well approximated as (see, e.g., [35])

$$R_{\max}\big(K\big) \backsimeq E[R] + \sqrt{2\sigma_R^2 \ln K}, \tag{3.43}$$

where $E[R]$ and σ_R^2 denote the mean and the variance of the data rate, respectively.

Consider a transmission process in Fig. 3.3. Let τ_M denote the duration when there are M active users remaining in the system, and A_M denote the system throughput within

the period τ_M. Then, $\sum_{k=1}^{K} A_k = \sum_{k=1}^{K} F_k$. Assuming the channel is ergodic, we have that

$$\tau_M = \frac{A_M}{E\big(R_{\max}(M)\big)}, \tag{3.44}$$

where $R_{\max}(M)$ denotes the maximum data rate across M remaining users. The total completion time is given by

$$\Psi = \sum_{k=1}^{K} \frac{kA_k}{E\big(R_{\max}(k)\big)}. \tag{3.45}$$

Using the results of [26], we get

$$E\big(R_{\max}(k)\big) = k \int_0^{\infty} R[P(R)]^{k-1} p(R) dR, \tag{3.46}$$

where $P(R)$ and $p(R)$ denotes the cumulative distribution function (CDF) and probability density function (PDF) for the (unordered) data rate R, respectively. Thus,

$$\frac{k}{E\big(R_{\max}(k)\big)} = \frac{1}{\int_0^{\infty} R[P(R)]^{k-1} p(R) dR}. \tag{3.47}$$

Since $0 \leq P(R) \leq 1$,

$$\int_0^{\infty} R[P(R)]^{K-1} p(R) dR = \min_k \Big(\int_0^{\infty} R[P(R)]^{k-1} p(R) dR \Big). \tag{3.48}$$

It follows that

$$\frac{K}{E\big(R_{\max}(K)\big)} = \max_k \Big(\frac{k}{E\big(R_{\max}(k)\big)} \Big). \tag{3.49}$$

Therefore, we conclude that

$$\Psi \leq \max_k \Big(\frac{k}{E\big(R_{\max}(k)\big)} \Big) \sum_{k=1}^{K} A_k = \frac{K}{E\big(R_{\max}(K)\big)} \sum_{k=1}^{K} F_k. \tag{3.50}$$

3.2.4. Numerical Examples

We illustrate the performance gain of the traffic aided opportunistic scheduling schemes via numerical examples. Using the simple Round-Robin algorithm as the baseline of performance evaluation, we define a *normalized* total completion time as $\Psi_{\mathrm{S}}/\Psi_{\mathrm{R}}$, where Ψ_{S} and

Table 3.1. Normalized total completion time

Schemes	W-SRPT	L-bound
Simulation	0.546	0.225
Theoretic	0.545	0.204

Ψ_{R} denote the total completion time of a given scheduling scheme and the Round-Robin algorithm, respectively.

Following [94], we assume that the system bandwidth is 1.25MHz, the slot duration t_s is 1.67ms, and the observation window T^c consists of 1000 slots. In the first example, we assume $K = 20$. All users in the system experience i.i.d. Rayleigh fading channels with long-term (average) signal-to-noise ratio (SNR) 0dB, and the maximal Doppler shift $f_m = 10$Hz. Along the line of [24], [106], we assume that the file sizes for Web browsing follow a heavy-tailed (Pareto) distribution with the minimal file size of 50kB and shape parameter $\alpha = 1.2$ (hence, the mean $\mu_F = 300$kB). In Table 3.1, we observe that the analytical results coincide with that of the Monte Carlo simulation, i.e., the difference between analytical results and simulation results is negligible. Therefore, the following simulation results are reliable for evaluating the performance. Also, we note that the upper bound for TAOS may not always be tight, and the lower bound provides a benchmark on the performance of the TAOS schemes.

A. Impact of Fading

Next, we investigate the impact of fading on the performance of TAOS schemes. Suppose that the average SNR is 0dB. Table 3.2 depicts the normalized total completion time in Rayleigh fading channels. In this example, we assume that the receiver can estimate the channel conditions perfectly, regardless of the Doppler shift. We observe that in such

Table 3.2. Normalized total completion time of scheduling schemes in Rayleigh Fading Channels

Schemes	W-SRPT	On-peak	TAOS-1	TAOS-1a	TAOS-1b	TAOS-2	L-bound
Simulation	0.548	0.449	0.407	0.371	0.409	0.366	0.225

Table 3.3. Normalized total completion time of scheduling schemes ($\mathcal{K}$=50dB)

Schemes	W-SRPT	On-peak	TAOS-1	TAOS-1a	TAOS-1b	TAOS-2	L-bound
Simulation	0.551	0.993	0.801	0.551	0.551	0.551	0.541

a channel with high dispersion, TAOS-2 achieves the best performance, whereas W-SRPT has the worst performance. The performance of other schemes is between these two above. The physically appealing explanation is that those schemes other than W-SRPT, employ channel variation for opportunistic scheduling. Thus, they achieve significant multiuser diversity gains, even though the gain may vary across different schemes. Therefore, the opportunistic scheduling can outperform W-SRPT in Rayleigh fading channels.

Fig. 3.4 gives the normalized total completion time in Rician fading channels with respect to the Rice factors. Several observations are in order. First, as would be expected, the total completion time of W-SRPT does not vary with the Rice factor $\mathcal{K}$. In fact, the W-SRPT only utilizes the average data rate, and thus the channel variation has little impact on the W-SRPT scheme. Next, the total completion times corresponding to the "riding on the peak", TAOS-1, TAOS-1a, TAOS-1b, and TAOS-2 schemes, increase with $\mathcal{K}$. Our intuition is as follows: the greater the Rice factor $\mathcal{K}$, the less variation the fading has, and the less *multiuser diversity* (see, e.g., [94], [101]) gains in the system throughput can be exploited. Therefore, as the Rice factor $\mathcal{K}$ increases, more time is needed to complete the transmissions for these three scheduling schemes. Furthermore, after $\mathcal{K}$ is greater than certain values, "riding on the peak" and TAOS-1 perform worse than W-SRPT. Finally,

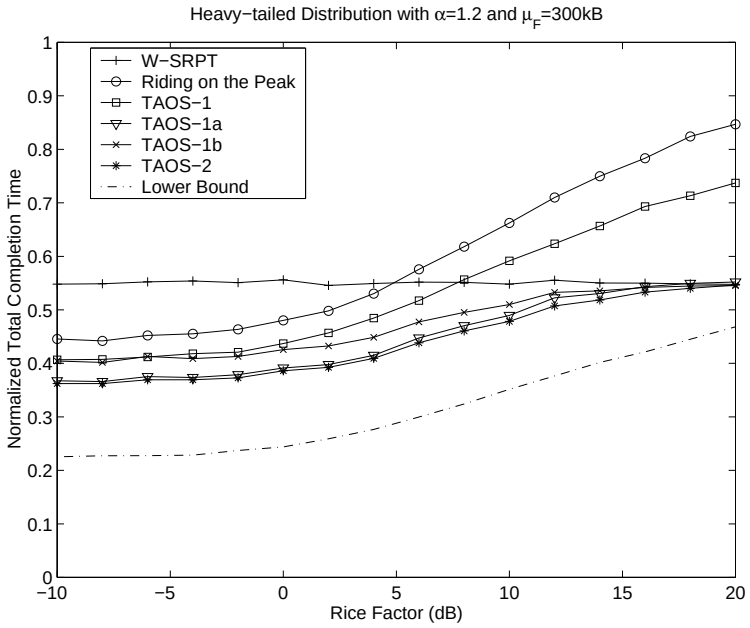

Figure 3.4. Normalized total completion time in Rician fading channels

the TAOS-2 always achieves the best performance, which is close to the lower bound, even though in general the lower bound is not achievable.

Table 3.3 describes the performance of scheduling schemes, when the Rice factor $\mathcal{K} =$ 50dB. From Table 3.3 and Fig. 3.4, we note that as $\mathcal{K} \to \infty$, i.e., multipath fading diminishes, the performance of TAOS-1a, TAOS-1b, and TAOS-2 converges to that of W-SRPT, whereas "riding on the channel peak" achieves only the same performance as Round-Robin. The performance of TAOS-1 lies in-between. We also observe that as $\mathcal{K} \to \infty$ the lower bound becomes tighter.

B. Impact of File Size Distribution

In the following, we examine the impact of network traffic on the performance of scheduling schemes. In particular, we compare the normalized total completion times cor-

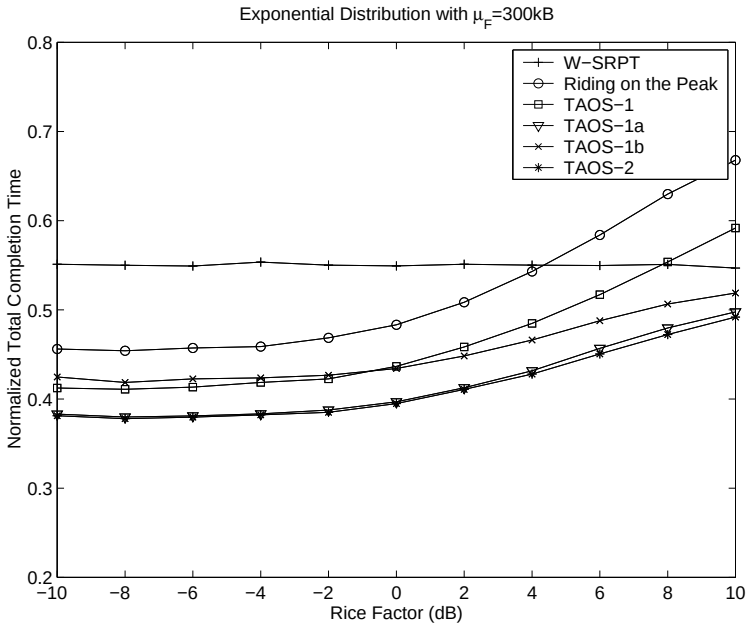

Figure 3.5. Normalized total completion time for exponentially distributed file size

responding to three different file size distributions: the heavy-tailed distribution, the exponential distribution, and the uniform distribution. We assume that they have the same mean $\mu_F = 300$kB. Comparing Fig. 3.5 and Fig. 3.6 with Fig. 3.4 (which is for the heavy-tailed distribution), we observe that our conclusions drawn above for the heavy-tailed distribution case are also applicable to the exponential and uniform distribution cases.

C. Impact of Random Arrivals and Departures

Next, we extend our study to the cases with random "users" arrivals and departures. We note that the theoretical analysis for the case with user dynamics remains open. In the following, we use simulation to evaluate the performance. Note that in each time slot, the proposed TAOS schemes make use of updated channel and file size information, and arrange the transmissions in an opportunistic manner. Since the TAOS schemes are capable of tracking the user dynamics, the TAOS schemes yield significant gains in the case with

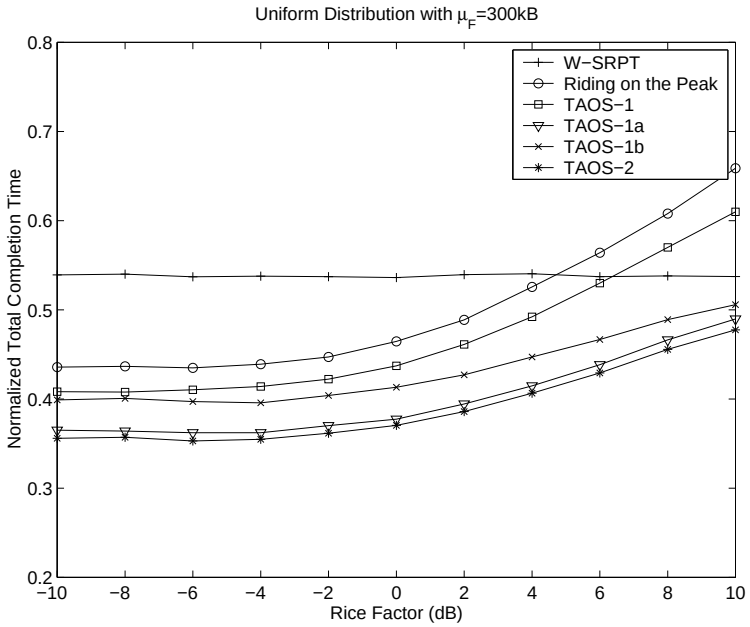

Figure 3.6. Normalized total completion time for uniformly distributed file size

random arrivals and departures as would be expected.

Assume that the arrival process is Poisson, all transmissions experience i.i.d. fading, and the file sizes obey the heavy-tailed distribution. Fig. 3.7 depicts the normalized total completion time with respect to the random arrival rate λ, when $\mathcal{K} = 0$dB. We can see that the higher the arrival rate, the shorter the normalized completion time of TAOS schemes, i.e., the more gains the TAOS schemes can yield. Our intuition is that if the arrival rate is high, at each instance it is likely that more users may join the contention. As a result, TAOS schemes can achieve possibly more multiuser diversity gains, leading to better performance.

D. A Comparison of System Throughput

Finally, we investigate the average system throughput corresponding to the TAOS schemes. We use the Round-Robin algorithm as the baseline, and define the throughput gain as $(U_{\mathrm{S}} - U_{\mathrm{R}})/U_{\mathrm{R}}$, where U_{S} and U_{R} denote the average system throughput of a given

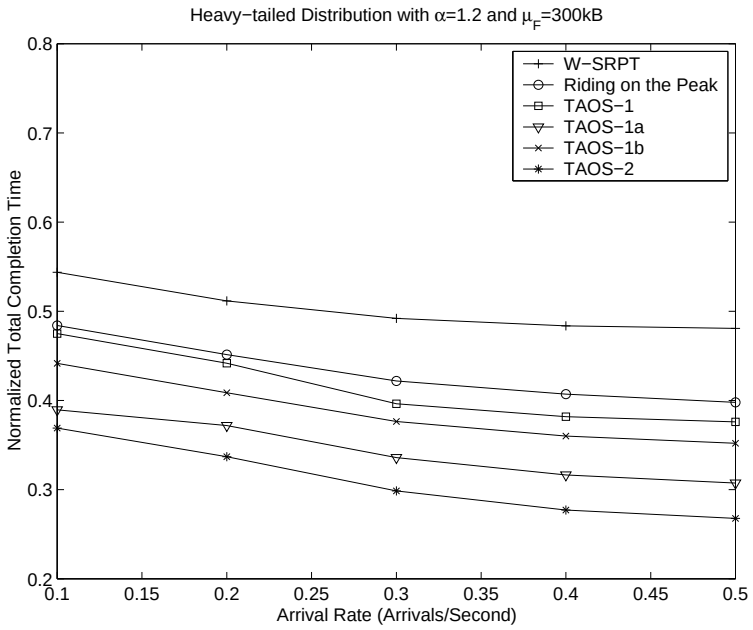

Figure 3.7. Normalized total completion time with respect to the arrival rate

scheduling scheme and the Round-Robin algorithm, respectively.

Fig. 3.8 depicts the throughput gain of the TAOS schemes in Rician fading channels, where $K = 20$. As expected, the throughput gain corresponding to the TAOS schemes lies in between that of the "riding on the channel peak" scheme and that of the W-SRPT scheme. More specifically, the throughput gain of the TOAS-2 scheme is about 15% lower than that of the "riding on the channel peak" scheme. This loss in throughput is due to the fact that the goal of TAOS schemes is set to minimize the total completion time. Another observation is that TAOS-2 achieves higher performance than the TAOS-1 schemes, in terms of both the system throughput and the completion time.

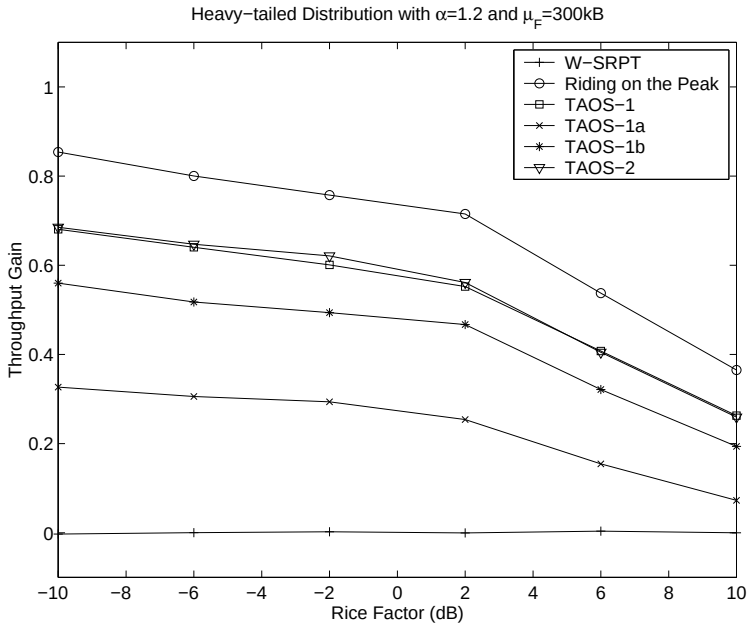

Figure 3.8. Throughput gain in Rician Fading Channels

3.3. Conclusions

In this chapter, we study cross-layer approaches in opportunistic communication systems to address QoS provisioning and system performance enhancement. Specifically, we first study admission control for opportunistic scheduling. We propose a traffic aided smooth admission control scheme, which increases the amount of the time resource allocated to trial users in a guarded manner and therefore spreads the admission decision to a trial period. Specifically, we use an adaptive resource allocation algorithm — QoS driven weight adaptation for WPF opportunistic scheduling, in which the base station allocates more resources to the incoming users by gradually increasing their weights. Based on the measured throughput, an admission decision is made within a time-out window. In particular, we employ explicitly the traffic information and throughput requirements in devising the back-off time. Our results show that our SAC scheme works well in opportunistic communication systems.

Then, we take a cross-layer optimization approach to minimize the total completion time (which consists of both processing time and waiting time). We establish general properties of scheduling schemes, and the identified convexity properties provide a basis for investigating opportunistic scheduling. We develop new traffic aided opportunistic scheduling schemes, namely TAOS-1, TAOS-1a, TAOS-1b, and TAOS-2. The TAOS schemes make use of both the file size information and channel variation, in a unified manner. As expected, they can yield significant reduction of the total completion time. We also derive lower and upper bounds on the total completion time. Furthermore, we investigate the impact of fading, file size distributions, and random arrivals and departures on the system performance. Our results show that the more channel variation, the more gains the TAOS schemes can achieve. This conclusion holds well for different file size distributions. We also observe that with random arrivals and departures, the higher arrival rate, the more reduction of the total completion time the TAOS schemes would yield. Therefore, the proposed TAOS schemes can perform well in heavy-loaded wireless networks. We observe that by making use of the traffic information and channel variation judiciously, the TAOS schemes, especially TAOS-2, achieve good performance in reducing the total completion time, at the cost of a little loss in throughput.

CHAPTER 4

MIMO Ad Hoc Networks: Medium Access Control and Saturation Throughput

In this chapter, we explore the utility of MIMO techniques for medium access control (MAC) in ad hoc networks. Recent years have witnessed a tremendous growth of interest in ad hoc wireless networks that can facilitate communications between wireless devices, without using a planned infrastructure. For example, robust ad hoc networks form a vital component in realizing the vision of communication-on-the-move in highly dynamic tactical environments, such as battlefields and disaster-relief events. Along a different avenue, lately the multiple-antenna technology has garnered much attention in wireless communications. It has been shown that using the multiple-input multiple-output (MIMO) techniques can boost up the channel capacity significantly and achieve high spectral efficiency [28], [92]. It is envisioned that the MIMO techniques can help to propel significant advances and lead to potential breakthrough towards robust ad hoc networks.

However, there has been little work on MIMO ad hoc networks, and it is unclear how to take advantage of MIMO techniques in ad hoc networks. It is well known that for point-to-point communications, the MIMO channel can offer spatial multiplexing gain, diversity gain, interference cancelling gain, and antenna array gain [18], [117]. We note that the

interference can reduce the number of effective reception antennas; accordingly, it is very difficult to achieve spatial multiplexing again in an inference-limited environment [18]. In this chapter, we focus on exploiting other MIMO techniques, including smart (directional) antenna and spatial diversity techniques, for MIMO ad hoc networks.

Spatial diversity and smart antennas are two important multiple-antenna techniques that have great potential utility in improving the performance of ad hoc networks. Specifically, spatial diversity can be deployed, when the antenna elements are placed far apart and the channels of different antennas are more or less uncorrelated. Spatial diversity (including both transmission diversity and reception diversity) can make use of i.i.d. spatial channels to combat fading and improve the reliability of the wireless links significantly (see [4], [64], [91]). In contrast, smart antennas are often used, when there exists a line of sight (LOS) and the channels corresponding to different antennas are highly correlated. Smart antennas can transmit and receive the signals in the directions of interest by forming narrow beams, and suppress the interference by forming nulls in corresponding directions (see [56] and the references therein). As a result, the smart antennas can increase the spatial reuse of the channel in the wireless networks. Moreover, due to beamforming, the smart antennas can concentrate the power on certain directions and thus yield antenna gains.

In this chapter, we focus on the MAC design and throughput analysis of ad hoc networks with multiple antennas. In particular, we devise MAC protocols using spatial diversity or directional antennas, namely SD-MAC and DA-MAC, and propose analytical methods to evaluate corresponding throughput performance. Our contributions to the MAC design can be summarized as follows. When the spatial channels experience independent fading, we propose a SD-MAC strategy based on IEEE802.11 DCF. We note that the exploitation of spatial diversity techniques for MAC has barely been studied [29], and we aim to pro-

vide some steps along this avenue. In the existence of LOS, we employ a general smart directional antenna model with both a mainlobe and sidelobes, in ad hoc networks. We demonstrate the utility of *directional listening* and incorporate it into the design of the DA-MAC protocol. One salient feature is that directional listening can resolve the hidden terminal problem due to the asymmetry in antenna gain (see [22], [54]). We also have studied the throughput analysis for ad hoc networks with multiple antennas. Specifically, assuming homogenous ad hoc networks using proposed MAC protocols, we characterize the saturation throughput. The proposed analytical methods takes into account the MIMO techniques, fading, and contention. In contrast to the method in [16], which works well *only* for one basic service set (BSS) of ad hoc networks with single antennas, our method can be applied to the multi-hop ad hoc networks with spatial diversity in fading channels, or with directional antennas in the channels of LOS. Furthermore, we evaluate the MIMO ad hoc networks with both theoretical methods and GloMoSim simulations.

4.1. System Models

We study the utility of multiple-antenna techniques for MAC design. Our MAC protocol design is built on the IEEE802.11 standard. In what follows, we first recapitulate the IEEE802.11 distributed coordination function (DCF) briefly and present the wireless channel model.

4.1.1. IEEE802.11 DCF

IEEE802.11 is a popular protocol for the ad hoc networks. In this standard, a DCF is developed to coordinate the multi-access by using CSMA/CA. Roughly, the DCF has two communication modes: 1) two-way handshaking, which involves DATA and ACK packets

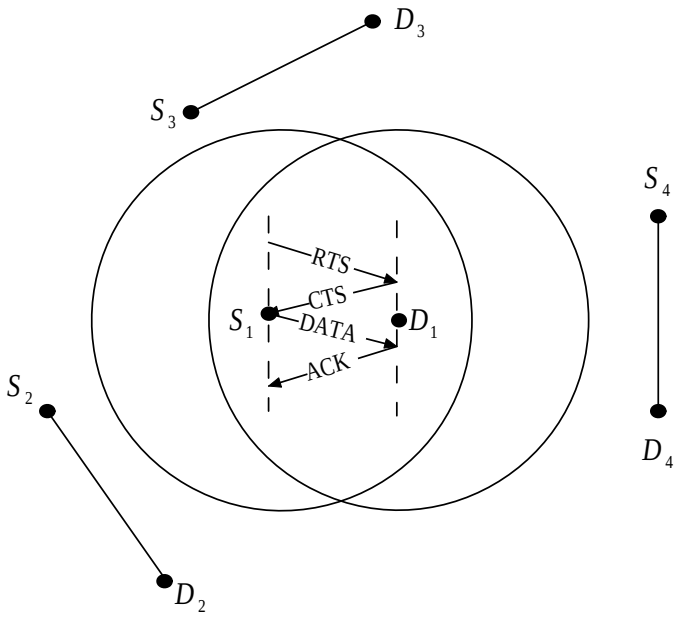

Figure 4.1. A simple diagram for RTS/CTS handshaking in IEEE802.11

only, and 2) four-way handshaking, which also adopts RTS and CTS packets, aiming to combat the so-called hidden terminal problem. In this chapter, we focus on the latter one. Simply put, in the RTS/CTS mechanism, each node (station) has a network allocation vector (NAV) table, which contains the remaining time of packet transmission of its neighbors. The nodes conduct virtual carrier sensing by using the NAV. More specifically, when the channel is physically sensed idle and the NAV is empty, the source node sends out a RTS packet. All other idle nodes, upon receiving the RTS packet, update their NAV tables, and defer their transmissions. When the destination node receives the RTS packet, it confirms the reception by sending back a CTS packet. Other idle nodes also overhear the CTS packet, and update their NAV tables correspondingly. After receiving the CTS packet, the source node transmits data, and then the destination node responds with the packet (see the IEEE802.11 standard [1] for more details). A simple diagram is given in Fig. 4.1.

4.1.2. Channel Model

Consider a generic channel model, where the fading coefficient can be expressed as

$$h = ae^{j\phi} + b, \tag{4.1}$$

where $ae^{j\phi}$ denotes the LOS component and is constant, and b denotes the time-varying component of the fading. It is known that for point-to-point communications, spatial diversity technique is suitable to the channels with i.i.d. fading, while smart antenna technique works well in the channels with LOS. In the following, we explore MIMO MAC design for the these two wireless channel models.

4.2. Ad Hoc Networks with Spatial Diversity

It is well known that in a point-to-point MIMO channel, the spatial degrees of freedom embedded in the antenna arrays can offer the following gains: spatial multiplexing gain, diversity gain, interference cancelling gain, and array gain [18], [117]. In a wireless communication environment, if the antenna elements are spaced far apart, the channels corresponding to different antenna elements may experience more or less independent fading. In such a scenario, spatial diversity can be utilized [4], [31], [91]. In this section, we assume i.i.d. fading across antenna elements, and study the utility of spatial diversity in ad hoc networks. A simple diagram is given in Fig. 4.2.

4.2.1. Spatial Diversity

Spatial diversity, including both transmission diversity and reception diversity, has been studied extensively to improve the reliability of wireless links. A comprehensive review on spatial diversity can be found in [31]. In an ad hoc networks where each node

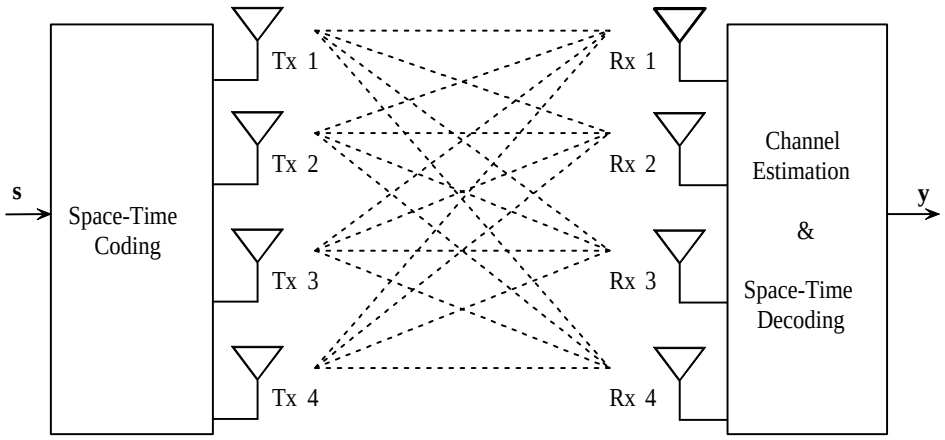

Figure 4.2. A MIMO link with 4-element antenna arrays for both transmission and reception

has an M-element antenna array, the MIMO systems can yield M^2 degrees of freedom for communications [18], [117]. We assume all degrees of freedom are utilized for spatial diversity. That is, each source-destination pair can possibly achieve a diversity gain of M^2. For example, with space-time block codes [4], [91], the system can perform maximal ratio combining (MRC) with a diversity order of M^2.

4.2.2. SD-MAC: A MAC Protocol with Spatial Diversity

Now, we generalize the IEEE802.11 DCF to take advantage of the spatial diversity technique. We assume that channel gains are obtained by using preamble symbols added to the each packet. In particular, the proposed MAC protocol corresponding to spatial diversity can be summarized as follows.

- *RTS transmission:* The source node, denoted S_k, receives a packet from its upper layer. Then, S_k performs virtual carrier sensing by checking the NAV table. If the NAV table is empty, S_k uses multiple antennas to carry out the physical carrier sensing. If the average received interference across reception antennas is lower than the

threshold for a period of DIFS, the channel is regarded as idle, and is available for the transmission. Then, the RTS packet with default data rate is transmitted by using spatial diversity techniques (e.g., space-time coding). If NAV table is not empty, or the channel is not sensed idle, the user needs to backoff a random period and defer its transmission. In particular, the user continues (virtual and physical) carrier sensing, and counts down the backoff counter only if the channel is idle. When the backoff counter becomes zero, the packet is sent out immediately.

- *RTS/CTS listening:* All idle nodes in the neighborhood overhear the RTS/CTS packets. Specifically, each idle node estimates the channels using the preamble symbols, conducts space-time decoding, obtains the transmission duration from the header of that packet, and then updates its NAV table.

- *RTS reception and CTS transmission:* For the destination node D_k, after receiving the RTS packet, D_k performs virtual and physical carrier sensing. If the NAV table is empty and the channel is idle for a duration SIFS, the channel is free. Then, D_k selects the rate control parameters for the following DATA packet from S_k based on the channel estimation, and transmits such information via the default-rate CTS packet to S_k using the spatial diversity technique. If the channel is busy, the CTS transmission is cancelled.

- *CTS reception and DATA transmission:* After the RTS transmission, the source node S_k waits for the CTS packet. Upon receiving the CTS packet, S_k senses the channel. If the channel is idle for a duration of SIFS, S_k adapts the transmission data rate according to the information from the CTS packet, and transmits the multi-rate DATA packet by using the spatial diversity technique. If the CTS packet does not arrive

within a time-out window, S_k would resend the RTS packet.

- *DATA reception and ACK transmission:* After sending out the CTS packet, the node D_k moves to the DATA reception phase. When the DATA packet is completely received, D_k confirms the reception by sending a default-rate ACK packet to S_k.

In summary, we explore the utility of spatial diversity in ad hoc networks (which has barely been studied before). The proposed SD-MAC takes into account the impact of spatial diversity on overhearing, the RTS/CTS dialogue, and data transmissions.

4.3. Ad Hoc Networks with Directional Antennas

As noted above, when the wireless channel has a LOS, smart antennas can yield gains for the desired signals while suppressing the interference. This property can be utilized to enhance the performance of the ad hoc networks. To achieve this goal, we address the smart antenna techniques briefly, and then develop a MAC protocol for ad hoc networks with smart directional antennas.

4.3.1. Smart Antennas

In wireless systems, smart antennas are often used if there exists a LOS. Roughly speaking, smart antennas have three forms: 1) switch-beam antennas, which consists of switchable narrow beam antennas; 2) smart directional antennas, whose antenna pattern has a fixed shape but the direction of the mainlobe is steerable; and 3) adaptive (pattern) antennas, whose antenna pattern is totally adaptive (see also, [88]). We note that switch-beam antenna technique can only select beam on some pre-determined directions, which may incur some loss of performance, whereas the adaptive antenna technique is more complicated

to be implemented in mobile terminals. In this chapter, we focus on smart directional antennas. The smart (directional) antenna technique has two key elements: direction of arrival (DOA) estimation and directional beamforming. Roughly, if there exists a LOS path, the antenna elements receive replica of the transmitted signal with different delays. By using the difference in the delays, the estimation algorithm (e.g., MUSIC [80]) can detect the DOA accurately. Based on the detected DOA, the smart antenna then can form a directional (transmission/reception) beam, thereby tuning its direction to the desired user.

To take advantage of the smart antenna technique in ad hoc networks, it is important to get a clear understanding of the directional beamforming technique of smart antennas. We outline the main ideas of directional transmission and reception in the following.

Directional transmission antennas: Consider a directional transmission antenna array with M antenna elements. Suppose that the directional antenna array uses a steering vector $\mathbf{w} = [c_1 e^{j\phi_1}, \cdots, c_M e^{j\phi_M}]^T$ to form a directional antenna pattern. The signal at a single reception antenna is the superposition of components from different (transmission) antenna elements, and can be expressed as

$$y = \mathbf{h}^T \mathbf{w} s + \mathbf{v}, \tag{4.2}$$

where $\mathbf{h} = [h_1, \cdots, h_M]^T$ are the fading coefficients corresponding to the links from antenna elements, s is the information-bearing signal, and $\mathbf{v}$ is the noise.

In an ideal case with LOS only (e.g., in an open area), the channels corresponding to different elements are highly correlated. That is, $a_1 = a_2 = \cdots = a_M$ and the difference in phase $(\phi_i - \phi_j),\ i \neq j$ is determined by the propagation delay [56]. Note that the propagation delay is a function of the transmission direction and the spatial separation of antenna elements. Then, by tuning the phases of the steering vector on transmission elements, the directional antenna can compensate for the propagation delay (corresponding

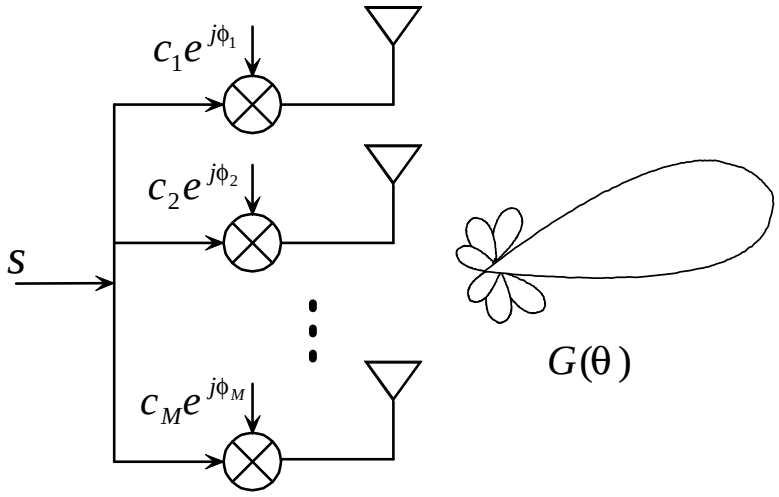

Figure 4.3. A sketch of smart directional antennas

to certain direction, denoted θ_k), and thus yield coherent combining of received components; due to non-coherent superposition, the antenna array may yield sidelobes and nulls in other directions. That is, the directional antenna array tunes its mainlobe pointing to θ_k. The directional antenna array can be characterized by the antenna gain pattern $G(\theta)$. A simple diagram is given in Fig. 4.3.

Directional reception antennas: Based on the reciprocal theorem [10], the reception antennas have the reciprocal behavior as the transmission antennas, and thus the above results are also applicable to reception antennas.

Remark: In this chapter, we assume that the DOA estimation is perfect, and that both transmission and reception antennas are used in ad hoc networks. In particular, we use a flat-top approximation for the directional antenna gain (see also, [79]), i.e.,

$$G(\theta) = \begin{cases} G_m & \text{for } -\frac{\Delta}{2} < \theta < \frac{\Delta}{2} \\ G_s & \text{otherwise,} \end{cases} \tag{4.3}$$

where Δ denotes the beamwidth of the directional antenna. Thus, when transmission and

reception antennas have their mainlobes pointing to each other, the signal can have both transmission antenna gain and reception antenna gain. Accordingly, the signal can possibly achieve a maximal antenna gain as G_m^2.

4.3.2. DA-MAC: A MAC Protocol Using Directional Antennas

Recently, there has been an increasing interest in ad hoc networks with directional antennas (e.g., [12], [22], [54], and the references therein). Recent works [22], [54], assume ideal beamforming and *omnidirectional listening*. In practice, however, the sidelobes may not be negligible. Moreover, since different antenna patterns lead to different antenna gains, the asymmetry in directional transmission/reception and omnidirectional listening, may result in the hidden terminal problem [22], [54]. For example, assume in Fig. 4.4 that node A transmits a directional RTS packet to node B, while node C is idle. Upon receiving RTS packet, node B sends a directional CTS to node A. Then, node A and B are engaged in DATA transmission, both using directional antennas with gain G_m. Note the coverage range is determined by both transmission and reception antenna gain. Let G_0 denote the gain of omnidirectional antennas. The total antenna gain taking into account both directional transmission and omnidirectional listening is $G_0 G_m$, smaller than G_m^2 when directional antennas are used for both transmission and reception. Therefore, node C using omnidirectional listening may not detect the directional CTS packet from node B. But, when the DATA packet from node A to node B is in progress, it is likely that the directional RTS from node C (with the directional beam toward node B) can cause a collision, leading to a hidden terminal problem.

We propose to use *directional listening* to resolve the hidden terminal problem. We note that by using the RTS signals as "pilot signals", each node can carry out directional

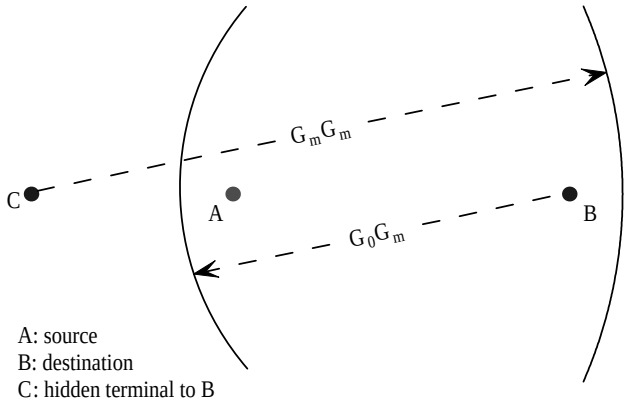

Figure 4.4. A hidden terminal problem due to asymmetry in antenna gain

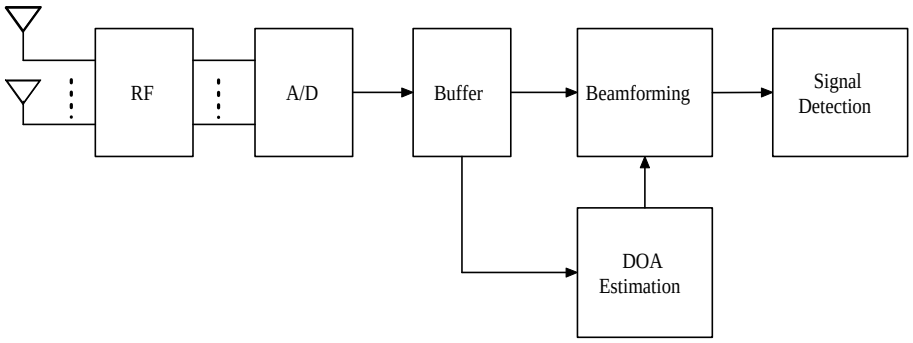

Figure 4.5. A block diagram for directional listening

listening via DSP techniques. By directional listening, we mean that each node is capable of listening to multiple nodes simultaneously with corresponding directional antenna patterns. More specifically, the RTS signals received by an antenna array can be stored in a buffer; the digital signal processor (DSP) uses a copy of the data in the buffer to perform the DOA estimation (see, [56], [80]); with the estimated DOAs and corresponding steering vectors, the DSP processes the data in the buffer again to perform the directional listening (beamforming). Moreover, with recent advances in DSP technologies, the node has the capability of exploiting multiple steering vectors within a packet duration. Therefore, roughly each node can be viewed as listening with smart directional beams pointing to multiple transmissions. A simple diagram is given in Fig. 4.5. The directional listening can be implemented with such a structure. Since listening, transmission, and reception are all directional and with the same antenna gain pattern, the hidden terminal problem aforementioned is thereby resolved. Indeed, the directional listening, together with a general directional antenna model with sidelobes, is incorporated into our MAC protocol; and this is a key feature of the proposed DA-MAC protocol below.

Worth pointing out is that connectivity is also an important issue of the ad hoc networks. In mobile ad hoc networks (MANET) with directional antennas, each node has to broadcast its information to keep connection to the networks. In principle, this goal can be achieved by sending out HELLO packets (see also, [61]) with circularly directional beams (see also, [54]). Upon receiving such packets, the neighboring nodes can estimate the DOA of the transmission node, render the contents of the packets, and update their routing tables when necessary. Therefore, in the MAC design, we assume that the DOA of the destination node is known for the RTS/CTS and DATA transmissions.

DA-MAC: Next, we develop a new MAC protocol for ad hoc networks with directional

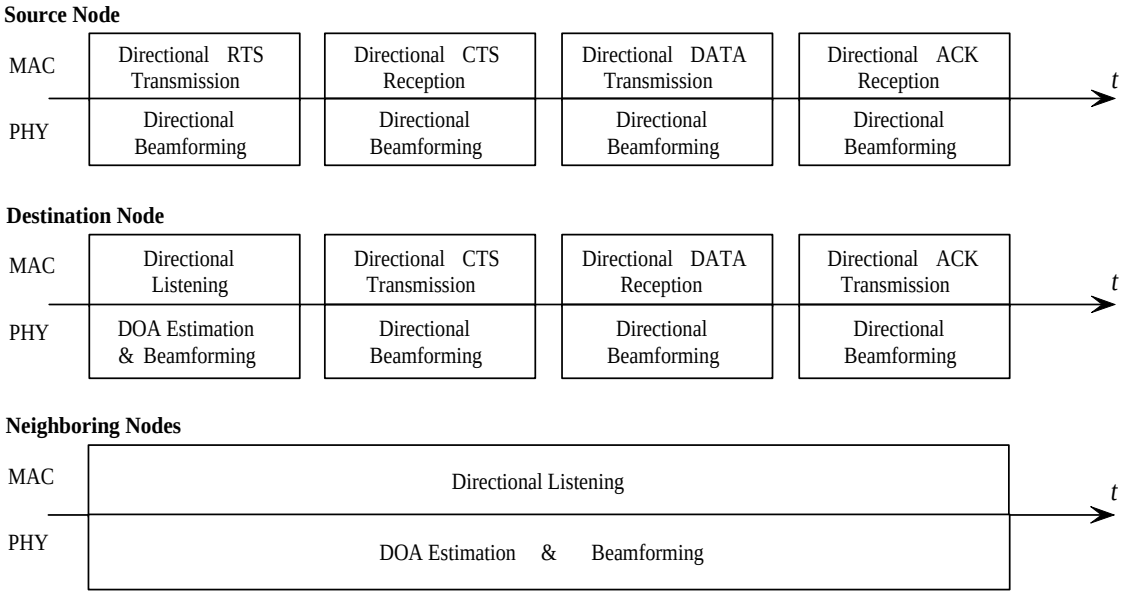

Figure 4.6. A block diagram for the four-way handshaking of DA-MAC: directional listening, directional transmission/reception

listening, directional transmission, and directional reception. A block diagram is given in Fig. 4.6 to depict the key steps in the directional RTS/CTS dialogue. For exploiting the benefits of directional antennas, two tables are used, namely the antenna pattern lookup table and the directional NAV (D-NAV) table. In the antenna pattern lookup table, the antenna gain is listed with respect to the azimuth direction. The D-NAV table consists of the RTS/CTS mode, node index, DOA, the signal power corresponding to each DOA, and NAV derived from the received RTS/CTS packet. The proposed DA-MAC protocol for directional antennas can be outlined as follows.

- *RTS transmission:* The source node, denoted S_k, receives a packet from the upper layer, and obtains the direction of the (next hop) destination node in its connectivity table. Then, the source node S_k performs virtual carrier sensing by using both its D-NAV table and antenna pattern table. Simply put, S_k calculates the effective

interference power for the nodes in the D-NAV as

$$P_e(\theta) = \frac{P_r(\theta)G\big(\theta - \theta^r_{kk}\big)}{G_m}, \tag{4.4}$$

where $P_r(\theta)$ is the received power of RTS/CTS in the DOA (denoted θ) by using directional listening, $G(.)$ is the reception antenna pattern, θ^r_{kk} denotes the angle of the mainlobe center, and G_m is the gain of the mainlobe. If for the desired direction, the corresponding $P_e(\theta)$ is below the threshold, the directional channel is viewed idle in the virtual carrier sensing. Then, the source node forms a narrow-beam antenna pattern, and performs physical carrier sensing. If the power of received interference is below the threshold for a period of DIFS, the channel is determined to be available for transmission, and the directional RTS packet is transmitted to the (next hop) destination, denoted D_k. Otherwise, the user S_k needs to backoff a random period and defer its transmission in this direction. In particular, the user continues directional carrier sensing, and counts down the backoff counter only if the channel is idle. When the backoff counter becomes zero, the packet is sent out immediately.

- *RTS/CTS listening:* All idle nodes in the neighborhood overhear the RTS/CTS packets directionally by using smart antenna techniques, and then update their D-NAV tables.

- *RTS reception and CTS transmission:* The destination node D_k overhears the RTS packet using directional reception beamforming. Upon receiving the RTS packet correctly, D_k conducts virtual carrier sensing as done for the *RTS transmission*. If the channel is viewed idle in virtual carrier sensing, D_k forms a directional beam and performs physical carrier sensing. If the channel is idle for a duration SIFS, the node transmits the directional CTS packet to S_k. If the channel is busy, the CTS

transmission is cancelled.

- *CTS reception and DATA transmission:* After the RTS transmission, S_k forms a directional reception antenna pattern and waits for the CTS packet. If S_k receives the CTS packet, it performs virtual carrier sensing and physical carrier sensing sequentially. If the channel is idle for a duration of SIFS, the DATA packet is then transmitted directionally. If the CTS packet does not arrive within a predetermined time-out window, S_k will resend the RTS packet.

- *DATA reception and ACK transmission:* After sending out the CTS packet, D_k moves to the DATA reception phase. When the DATA packet is received, D_k confirms the reception by sending a ACK packet to S_k directionally.

In a nutshell, we incorporate directional listening into the MAC design to resolve the hidden terminal problem. Furthermore, the proposed DA-MAC protocol uses a general antenna pattern model with sidelobes, and makes use of directional listening, directional transmission, and directional reception. Clearly, the DA-MAC protocol is tailed to enhance the spatial reuse.

4.4. Saturation Throughput

Recall that spatial diversity achieves diversity gain in fading channels, whereas directional antennas yield spatial reuse and antenna gain if there exists a LOS component. It is natural to expect that in the case with i.i.d. fading, the MAC protocol using spatial diversity can yield better performance, while in the case with LOS only, the MAC protocol corresponding to directional antennas would achieve better performance. In the following, we investigate the throughput performance for these two cases. As in [16], we focus on the

saturation throughput, which is defined as the maximum load that the system can carry in suturation conditions. That is, each user always has packets in its buffer waiting for transmission. We assume perfect channel sensing in the ad hoc systems with RTS/CTS mechanism, and thus collisions only occur on the RTS frames. For the sake of clarity, we first recapitulate the saturation throughput for omnidirectional antennas presented in [16]. We then derive the saturation throughput for ad hoc networks with directional antennas and spatial diversity, respectively.

4.4.1. Preliminary: The Omnidirectional Antenna Case

Along the line of [16], we assume the "collision" (loss) probability of each packet p to be a constant. By analyzing the Markov chain model for CSMA/CA, the probability τ that a station transmits in randomly chosen time slot is given by

$$\tau = \frac{2(1-2p)}{(1-2p)(W+1)+pW(1-(2p)^m)}, \tag{4.5}$$

where p is the collision probability, W is the minimum backoff window in terms of backoff slots, and m is the maximum backoff stage. In a basic service set (BSS) of IEEE802.11 with K users, each with an omnidirectional antenna, the probability p can be expressed as

$$p = 1-(1-\tau)^{K-1}. \tag{4.6}$$

The above two equations can be viewed a non-linear system, and can be solved by using numerical methods.

To find the saturation throughput, Bianchi [16] examined the states of the system in a randomly chosen time slot: 1) the channel is empty, 2) the channel experiences a successful transmission, and 3) the channel has a collision. Let P_{tr} be the probability that there is at

least one transmission in the considered slot time. Since K mobile stations contend on the same channel and each transmission with probability τ, we have

$$p_{tr} = 1 - (1 - \tau)^K. \tag{4.7}$$

The conditional successful probability p_s is given by

$$p_s = \frac{K\tau(1-\tau)^{K-1}}{p_{tr}} = \frac{K\tau(1-\tau)^{K-1}}{1-(1-\tau)^K}. \tag{4.8}$$

As a result, the channel is empty with probability $(1-p_{tr})$, contains a successful transmission with probability $p_{tr}p_s$, and has a collision with probability $p_{tr}(1-p_s)$. Then, the system saturation throughput can be calculated as

$$Th = \frac{p_{tr}p_s\mathbb{E}[L]}{(1-p_{tr})\epsilon + p_{tr}p_sT_s + p_{tr}(1-p_s)T_c}, \tag{4.9}$$

where $\mathbb{E}[L]$ is the average packet payload size, ϵ is the duration of a backoff slot, i.e., the minimum time needed for transmission detection, T_s is the average time of a successful transmission, and T_c is the average duration of a collision. In the system with RTS/CTS handshakes, the T_s and T_c are given as

$$T_s = RTS+SIFS+\delta+CTS+SIFS+\delta+OH+\mathbb{E}[T_p]+SIFS+\delta+ACK+DIFS+\delta, \tag{4.10}$$

$$T_c = RTS + DIFS + \delta, \tag{4.11}$$

where δ is the propagation delay, OH is the overhead including both MAC and PHY headers, and $\mathbb{E}[T_p]$ is the average transmission duration for payload.

4.4.2. Saturation Throughput for MIMO Ad Hoc Networks

Recall that in the IEEE802.11 standards, all users within one BSS can communicate directly with each others. Thus, if any two users transmit at the same time, a collision

would occur. By making use of this property, the saturation throughput can be obtained by examining the system states of a BSS. In an ad hoc system with directional antennas, however, simultaneous transmissions are feasible. Therefore, the methods in [16] cannot be applied directly to calculate the saturation throughput for the MIMO case. Instead, we focus on the average throughput per user. We assume a homogeneous ad hoc network, in which the events experienced by one user are statistically the same as those of other users; worth pointing out is that "statistically" here refers to long-term statistics. We say that such a user is a *typical user*. To this end, we examine the events experienced by a typical user, and derive the corresponding saturation throughput per user.

In the following, we model the events experienced by the typical user (say S_k) into five states: 1) S_k does not transmit and detects the channel empty; 2) S_k does not transmit and overhears one RTS packet from one of the neighboring users, as if it "views" that user has a successful transmission; 3) S_k does not transmit and overhears a collision among the transmissions of other users; 4) S_k has a successful transmission; and 5) the transmission of S_k collides with that of the others. Let $\{p_i, i = 1, \ldots, 5\}$ denote the probabilities corresponding to the above events. Then, the average throughput of a typical user under the saturation condition can be expressed as

$$Th_u = \frac{p_4 \mathbb{E}[L]}{\sum_{i=1}^{5} p_i T_i}, \tag{4.12}$$

where T_i denotes the duration of state i. Moreover, we have $T_1 = \epsilon$, $T_2 = T_4 = T_s$, and $T_3 = T_5 = T_c$.

Although the states of one user depend on those of the others, each user has a statistically identical (renewal) period $\sum_{i=1}^{5} p_i T_i$, and hence the same average throughput. Moreover, the successful transmitted packets of one user does not overlap with others. Therefore, the total average throughput in the area with K users can be expressed as $K \cdot Th_u$. We

note that the proposed analysis method can yield the same result given in (4.9), when it is applied to the omnidirectional case in [16].

It is interesting to re-examine the omnidirectional antenna case using the proposed method. Note that in a BSS, each user can hear all other users. The state probabilities are given as

$$p_1 = (1-\tau)(1-\tau)^{K-1} \tag{4.13}$$

$$p_2 = (1-\tau)(K-1)\tau(1-\tau)^{K-2} \tag{4.14}$$

$$p_3 = (1-\tau)\big[1-(1-\tau)^{K-1}-(K-1)\tau(1-\tau)^{K-2}\big] \tag{4.15}$$

$$p_4 = \tau(1-\tau)^{K-1} \tag{4.16}$$

$$p_5 = \tau\big(1-(1-\tau)^{K-1}\big). \tag{4.17}$$

Moreover, we have $T_1 = \epsilon$, $T_2 = T_4 = T_s$, and $T_3 = T_5 = T_c$. Then, we can calculate the average throughput of one user by plugging $\{p_i, T_i\}, i = 1, \cdots, 5$ into (4.12), i.e.,

$$Th_u = \frac{\tau(1-\tau)^{K-1}\mathbb{E}[L]}{(1-\tau)^K\epsilon + K\tau(1-\tau)^{K-1}T_s + [1-(1-\tau)^K - K\tau(1-\tau)^{K-1}]T_c}. \tag{4.18}$$

Since each user statistically has the same performance, the system saturation throughput is given by

$$\begin{aligned} Th &= \sum_{i=1}^{K} Th_i \\ &= \frac{K\tau(1-\tau)^{K-1}\mathbb{E}[L]}{(1-\tau)^K\epsilon + K\tau(1-\tau)^{K-1}T_s + [1-(1-\tau)^K - K\tau(1-\tau)^{K-1}]T_c} \\ &= \frac{p_{tr}p_s\mathbb{E}[L]}{(1-p_{tr})\epsilon + p_{tr}p_sT_s + p_{tr}(1-p_s)T_c}. \end{aligned} \tag{4.19}$$

That is, the proposed analysis method yields the same results in [16].

A. Saturation Throughput: The Spatial Diversity Case

In this section, we derive the saturation throughput for the MIMO ad hoc networks with spatial diversity. We first establish the relationship between τ and p. Worth pointing out is that p denotes the probability that the RTS packet cannot be received correctly. Since fading can also cause the loss of RTS packet, in the spatial diversity case, p consists of the probability incurred by both collision and fading. Let p_c denote the loss probability that RTS is not faded at the destination node but corrupted by other contention signals, and p_f denote the packet loss probabilities corresponding to fading.

Assume there are K_a users within the coverage of each node, in a homogeneous ad hoc network. K_a can be approximated as $K_a = \pi A^2 \rho$, where A is the average coverage range, and ρ is the node density. Suppose that the distance r between any two nodes obeys a random distribution with pdf $f_r(r)$. Then, the average packet loss probability due to fading can be expressed as

$$p_f = \int_r p_f(r) f_r(r) dr, \tag{4.20}$$

where $p_f(r)$ is the loss probability due to fading, corresponding to a given distance r. The probability of the packet loss due to the collision can then be calculated as

$$\begin{aligned} p_c &= (1-p_f)\Big[1-(1-\tau)\Big((1-\tau)^{K_a-2} + \binom{K_a-2}{1}(1-\tau)^{K_a-3}\tau p_f + \cdots + (\tau p_f)^{K_a-2}\Big)\Big] \\ &= (1-p_f)\Big[1-(1-\tau)(1-\tau+\tau p_f)^{K_a-2}\Big]. \end{aligned} \tag{4.21}$$

Therefore, the loss probability of the RTS packet is given by

$$p = p_c + p_f = (1-p_f)\Big[1-(1-\tau)(1-\tau+\tau p_f)^{K_a-2}\Big] + p_f. \tag{4.22}$$

Combining (4.5) and (4.22), τ and p can be obtained by numerical methods. The state probabilities can be outlined as follows.

p_1**:** S_k is listening, and it detects the channel to be empty:

$$p_1 = (1-\tau)(1-\tau+\tau p_f)^{K_a-1}. \tag{4.23}$$

p_2**:** S_k is listening, and hears a handshaking packet from one of its neighbors:

$$p_2 = (1-\tau)(K_a-1)(1-p_f)\tau(1-\tau+\tau p_f)^{K_a-2}. \tag{4.24}$$

p_3**:** S_k is listening, and detects a "collision" among the transmissions of its neighbors.

$$p_3 = (1-\tau)[1-(1-\tau+\tau p_f)^{K_a-1} - (K_a-1)(1-p_f)\tau(1-\tau+\tau p_f)^{K_a-2}] \tag{4.25}$$

p_4**:** S_k transmits its RTS packet and the transmission is successful. Note that if the RTS packet is faded, there is no corresponding CTS transmission.

$$p_4 = (1-p_f)\tau(1-\tau)(1-\tau+\tau p_f)^{K_a-2}. \tag{4.26}$$

p_5**:** The RTS packet of S_k cannot be received correctly due to collision or fading:

$$p_5 = \tau\Big[1-(1-p_f)(1-\tau)(1-\tau+\tau p_f)^{K_a-2}\Big]. \tag{4.27}$$

We assume that the data rate is R. In general, R is time-varying due to the time-varying fading, and can be expressed as a function of the instantaneous signal-to-noise ratio (SNR), denoted γ. Thus, the average transmission duration is given as

$$\mathbb{E}[T_p] = \int_L \int_\gamma \frac{L}{R(\gamma)} g_{\mathrm{L}}(L) g_\gamma(\gamma) d\gamma dL, \tag{4.28}$$

where g_{L} and g_γ denote the probability density function for the payload L and SNR γ, respectively.

In summary, if the wireless channels corresponding to different antennas experience i.i.d. fading, the spatial diversity can be utilized. We have the following result on the saturation throughput per user.

Proposition 4.4.1 *In an ad hoc network with spatial diversity, the saturation throughput per user is given by*

$$Th_u(\rho) = \frac{p_4 \mathbb{E}[L]}{\epsilon p_1 + T_s(p_2 + p_4) + T_c(p_3 + p_5)}, \tag{4.29}$$

where $p_i, i = 1, \ldots, 5$ *are given in (4.23)–(4.27).*

We note that similar analysis can be carried out for the omnidirectional antenna case in time-varying fading channels. In particular, the saturation throughput has the same form as in (4.29), but with different values of parameters.

B. Saturation Throughput: The Directional Antenna Case

We now consider the case with directional antennas. Suppose that idea beamforming is achieved, and the directional antenna with beamwidth Δ is used from both transmission and reception. Let K_ρ denote the number of mobile stations within the coverage area of one station. In an ad hoc network with the coverage range d and the node density ρ, K_ρ can be approximated as

$$K_\rho = \frac{\Delta}{2\pi} \pi d^2 \rho. \tag{4.30}$$

Note that the coverage range d is determined by both directional transmission gain and directional reception gain, as depicted in Fig. 4.7. Moreover, we assume that node S_k knows the address and direction of its destination node, but does not have knowledge about the behavior of the other nodes. Statistically speaking, node S_k just "sees" its neighbors transmit in each direction with equal probability.

Next, we examine the DCF for directional antennas. Assuming the (RTS) packet collision probability is p, we note that the Markov chain of the backoff counter has the same state transition diagram as that for omnidirectional antennas, and thus τ and p have the

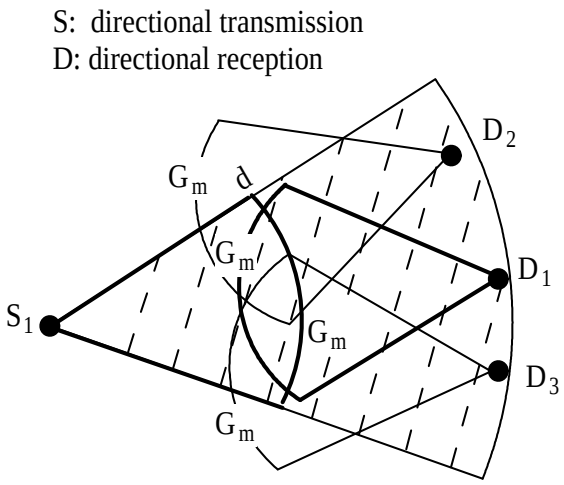

Figure 4.7. A diagram of coverage area in ad hoc networks with directional antennas

same relationship as in (4.5). Moreover, in the context above, if S_i is within the reception coverage of some other user, S_i would transmit in the direction to that user with probability $\frac{\Delta}{2\pi}\tau$. Then, the collision probability p can be expressed as

$$p = 1 - (1-\tau)(1-\frac{\Delta}{2\pi}\tau)^{K_\rho - 2}. \tag{4.31}$$

Therefore, τ and p can be derived by solving the equations (4.5) and (4.31).

In the following, we derive the saturation throughput for the ad hoc network with directional antennas by characterizing the state probabilities. Note that the probabilities p_1, p_2, and p_3 describe the states overheard by the S_k using its directional reception antenna, and p_4 and p_5 denote the states that the RTS packet of S_k is transmitted.

p_1: S_k is in the directional listening mode, and within its coverage area, all neighboring nodes do not transmit in the direction to S_k:

$$p_1 = (1-\tau)(1-\frac{\Delta}{2\pi}\tau)^{K_\rho - 1}. \tag{4.32}$$

p_2: S_k is in the directional listening mode, and overhears a handshaking packet from one

of its neighbors in the direction to S_k:

$$p_2 = (1-\tau)(K_\rho - 1)\frac{\Delta}{2\pi}\tau(1-\frac{\Delta}{2\pi}\tau)^{K_\rho - 2}. \tag{4.33}$$

p_3**:** S_k is in the directional listening mode, and detects a collision among its neighboring nodes:

$$p_3 = (1-\tau)\Big[1-(1-\frac{\Delta}{2\pi}\tau)^{K_\rho - 1} - (K_\rho - 1)\frac{\Delta}{2\pi}\tau(1-\frac{\Delta}{2\pi}\tau)^{K_\rho - 2}\Big]. \tag{4.34}$$

p_4**:** S_k transmits its RTS packet and the transmission is successful:

$$p_4 = \tau(1-\tau)(1-\frac{\Delta}{2\pi}\tau)^{K_\rho - 2}. \tag{4.35}$$

p_5**:** The RTS packet of S_k collides with the transmission of its neighbors:

$$p_5 = \tau\Big[1-(1-\tau)(1-\frac{\Delta}{2\pi}\tau)^{K_\rho - 2}\Big]. \tag{4.36}$$

Moreover, the durations T_s and T_c have the forms shown in (4.10) and (4.11), while the average transmission duration is expressed as

$$\mathbb{E}[T_p] = \frac{\mathbb{E}[L]}{R}, \tag{4.37}$$

where R is the transmission data rate.

In summary, if the channel has a strong LOS, directional antennas can be used to enhance the performance of ad hoc networks. We have the following result on the saturation throughput per user.

Proposition 4.4.2 *In an ad hoc network with directional antennas, the saturation throughput per user is given by*

$$Th_u(\rho) = \frac{p_4\mathbb{E}[L]}{\epsilon p_1 + T_s(p_2+p_4) + T_c(p_3+p_5)}, \tag{4.38}$$

where $p_i, i = 1,\ldots,5$ *are given in (4.32)–(4.36).*

Table 4.1. System parameters in ad hoc networks

Propagation Delay	1 μs
SIFS	10 μs
DIFS	50 μs
Backoff Slot	20 μs
MAC Header	272 bits
PHY Header	192 bits
RTS	160 bits + PHY Header
CTS	112 bits + PHY Header
ACK	112 bits + PHY Header
Payload	8184 bits
Min Backoff Window	32 slots
Max Backoff Stage	3
Antenna Elements	4

Table 4.2. SNR vs. Data Rate

Threshold (dB)	0	3	5.5	8.5
Data Rate (Mbps)	1	2	5.5	11

4.4.3. Numerical Examples

In this section, we illustrate the performance of the above MAC protocols, via numerical examples. We focus on a typical user, and its surrounding circle area with a radius of converge range, namely a *typical area*. We choose the ad hoc network with omnidirectional antennas as a baseline. Specifically, we define a throughput gain as $\frac{Th_m - Th_0}{Th_0}$, where Th_m and Th_0 denote the saturation throughput per user corresponding to multiple antennas and omnidirectional antennas, respectively. The common parameters used in both directional antenna case and spatial diversity case, are summarized in Table 4.1 (see also, [1]).

First, we examine the performance of ad hoc networks with spatial diversity in fading channels (with no LOS). To start with, we present the saturation throughput using the proposed analytical models, where the fading of each link is statistically identical. We

Table 4.3. Saturation throughput vs. number of users: the single antenna case (for the i.i.d. fading channel)

Number of users	10	15	20	30
Th_u (Mbps)	0.138	0.0927	0.0696	0.0465
K_aTh_u (Mbps)	1.38	1.390	1.392	1.395

assume that the user are uniformly distributed in the plane. The path loss factor is 2.5. The coverage range is 200m. The average SNR on the boundary for the omnidirectional antenna case is 0 dB, and multirate transmission is used. Using the practical parameters from D-Link, the information-bearing data rate and the SNR (after the processing of spatial diversity) have the relationship as in Table 4.2. For comparison, we also present the saturation throughput of ad hoc networks with single antennas in Table 4.3 As expected, we can see that each user has a lower throughput, when the number of users (denoted K_a) involved in contention increases, whereas the total throughput of the one typical coverage area tends to saturate. Suppose that 4-element antenna arrays are used. Table 4.4 depicts the saturation throughput of MIMO ad hoc networks with spatial diversity. Similarly, we observe that the throughput per user user decreases as the number of users of a typical area increases. One interesting observation is that in contrast to the case for single antennas, K_aTh_u corresponding to spatial diversity case decreases slightly as K_a increases. Our intuition is that for transceivers with single antennas, the fading plays a very important role in determining the loss probability and transmission probability. But, in MIMO ad hoc networks with spatial diversity, the link quality is greatly improved, and thus the transmission probability and loss probability is governed by the contention. It is already known that in the channels without time-varying fading, the more contention, the lower the system suturation throughput (see [16]). Therefore, as would be expected, in the spatial diversity

Table 4.4. Saturation throughput vs. number of users: the spatial diversity case (for the i.i.d. fading channel)

Number of users	10	15	20	30
Th_u(Mbps)	0.330	0.219	0.163	0.107
$K_a Th_u$(Mbps)	3.30	3.28	3.26	3.21

Table 4.5. Throughput gain vs. number of users (for the i.i.d. fading channel)

Number of users	10	15	20	30
Throughput gain	1.40	1.36	1.34	1.30

case, the saturation throughput of a typical area decreases when more contention users are involved.

To evaluate the performance gain from spatial diversity, we define the throughput gain as $\left(Th_u^s - Th_u^0\right)/Th_u^0$, where Th_u^s and Th_u^0 are saturation throughput per user corresponding to the spatial diversity case and single antenna case, respectively. Table 4.5 shows that spatial diversity can enhance the system performance in fading channels. Intuitively, due to spatial diversity, the reliability of each link is improved, leading to a higher probability of high-data-rate transmissions. As a result, the system throughput are improved. Moreover, we observe that the throughput gain decreases as the the number of users increases. Our intuition is that with multiple antennas, the link is more reliable. The packet loss due to fading is "overwhelmed" by the packet loss due to (contention) collision. Thus, when more users are involved in contention, the packet loss probability in the spatial diversity case may increase much faster than that in the single antenna case, resulting in a lower throughput gain.

Next, we demonstrate the performance of MIMO ad hoc networks with spatial diversity, in more practical scenarios. Particularly, we run simulations using an event-driven network

Table 4.6. Simulation parameters for GloMoSim

Tx Power	20 dBm
Path-loss Factor	2.5
Rx Sensitivity	-91 dBm
Rx Threshold	-88 dBm
Rx SINR Threshold	0 dB

Table 4.7. Saturation throughput vs. number of users (for the i.i.d. fading channel)

(Single antenna case)

Number of users	10	15	20	30
Th_u(Mbps)	0.0941	0.0660	0.0515	0.0357
$K_a Th_u$(Mbps)	0.941	0.99	1.03	1.07

(Spatial diversity case)

Number of users	10	15	20	30
Th_u(Mbps)	0.222	0.1407	0.1035	0.066
$K_a Th_u$(Mbps)	2.22	2.11	2.07	1.98

simulator — GloMoSim [97]. We have implemented space-time block codes $\mathcal{H}_4$ (see [91] for details) in this simulator. The simulation parameters are listed in Table 4.6. The nodes are uniformly distributed in an area of (250m, 250m). Table 4.7 presents the saturation throughput for the single antenna case and the spatial diversity case. Table 4.8 depicts throughput gains with respect to the number of users in the area. We observe that the conclusions derived from above ideal cases can be carried out to the practical cases using GloMoSim. It illustrates that the spatial diversity can improve the performance of ad hoc networks greatly in practical circumstances.

In the following, we demonstrate the utility of directional antennas in the ad hoc

Table 4.8. Throughput gain vs. number of users (for the i.i.d. fading channel)

Number of users	10	15	20	30
Throughput gain	1.36	1.13	1.01	0.85

networks. Worth pointing out is that with directional antennas, the nodes can achieve power gains. That is, using the same transmission power, each node can have a greater coverage range. This can have impact on both connectivity and routing. For instance, the power gain can be used to implement longer hop transmission/routing, and yield further improvement (such impact on upper layers is beyond the scope of this chapter). Thus, in this example, we allow the nodes with antenna arrays to tune the transmission power, thereby having the same coverage range as that with omnidirectional antennas [54]. We assume that 4-element antenna arrays are used and each antenna array has a directional antenna pattern with $\Delta = \pi/2$ approximately. Then, if there are 20 nodes in a typical (circle) area, on average 5 nodes are within the coverage area of a directional antenna. Since the channel has LOS only and is fixed over time, we assume a fixed transmission rate 1Mbps. Table 4.9 depicts the saturation throughput per user with respect to the number of nodes in a typical surrounding area. We see that as the number of users increases, the saturation throughput per user decreases. That is, the higher the node density, the lower saturation throughput each user has. It is because that when the node density increases, more nodes are involved in the channel contention. As a result, each node achieves a lower throughput. Table 4.10 describes the throughput gain with respect to the number of nodes in a typical area. We observe that with 4-element directional antennas, the ad hoc networks can yield a throughput gain around 10-fold. Our intuition is that by using directional antennas at both transmitters and receivers, the smart antenna technique can increase the spatial reuse, thereby improving the system throughput significantly. We also observe that the throughput gain increases slightly with the number of users. Our intuition is that when the number of users in the typical area increase, due to the spatial reuse, the users within the coverage area of each directional antenna can increase more slowly than

Table 4.9. Saturation throughput per user vs. number of users: the directional antenna case (for the LOS channel model)

(Omnidirectional antenna case)

Number of users	20	28	36	44	60
Th_u (Mbps)	0.0491	0.0350	0.0271	0.0221	0.016

(Directional antenna case)

Number of users	20	28	36	44	60
Th_u (Mbps)	0.4896	0.3928	0.3279	0.2808	0.2191

that in the omnidirectional antenna case. In this sense, the spatial reuse "dominates" over contention. That is to say, the saturation throughput of each user with omnidirectional antennas may decrease much faster than the users with directional antennas. As a result, ad hoc networks with directional antennas can achieve greater gains in higher user density regimes.

Worth pointing out is that the above numerical results are derived by the analytical methods in Section 4.4.2, based on the ideal directional antenna model: $\Delta = \pi/2$ and the antenna gain for sidelobes is 0. In practice, the beamwidth is determined by the antenna pattern design algorithms [10]. Given the number of antennas, the beamwidth (of the mainlobe) Δ, the mainlobe antenna gain, and the sidelobe antenna gain are correlated. Roughly speaking, there is always a tradeoff between the beamwidth and the antenna gain. The narrower the beamwidth, the smaller the gain difference between the mainlobe and the sidelobes; and vice versa. To make the interference from the sidelobes negligible, the antenna gain difference should be large, dictating a wider beam, (possibly $\Delta > \pi/2$ for 4-element antennas). In such cases, the system would achieve smaller throughput gains than in the ideal case above.

It should be cautioned that the above results for the spatial diversity case and the directional antenna case are for different wireless channel models (so they are not compa-

Table 4.10. Throughput gain vs. number of users (for the LOS channel model)

Number of users	20	28	36	44	60
Throughput gain	9.97	11.2	12.1	12.7	13.7

rable). When there is i.i.d. time-varying fading, the spatial diversity can combat fading and improve the link quality, whereas the directional antennas would not work well. In contrast, when there is a LOS, the channel is more reliable. Thus, the directional antennas can exploit spatial reuse to improve the system throughput.

4.5. Conclusions

In this chapter, we explore the utility of multiple-antenna techniques for MAC design in ad hoc networks. When the channels corresponding to different antennas are not correlated, we address ad hoc networks with spatial diversity. We examine the impact of spatial diversity on the MAC design, and devise the corresponding MAC protocol, namely SD-MAC. In contrast, when the channel has a LOS component, we employ directional antennas to improve the performance of ad hoc networks. We demonstrate the utility of directional listening and incorporate it into the MAC protocol. The proposed MAC protocol, namely DA-MAC, takes into account directional listening, directional transmission, directional reception, and a general directional antenna model with sidelobes. We also develop analytical methods in characterizing the saturation throughput for MIMO ad hoc networks. The proposed analytical methods takes into account the MIMO techniques, fading, and contention, and can be applied directly to the multi-hop ad hoc networks with spatial diversity in fading channels, or with directional antennas in the channels of LOS. Furthermore, we evaluate the MIMO ad hoc networks via both theoretical methods and GloMoSim simulations. The

throughput results show that the spatial diversity technique and smart antenna technique can enhance the performance of ad hoc networks significantly.

There are many interesting problems deserving further investigation. For example, we observe that it is difficult to achieve the spatial multiplexing gain in the interference-limited scenarios. But, if one MAC protocol could mitigate the interference, spatial multiplexing might yield gains in ad hoc network. Therefore, it is of interest to explore such a protocol, and/or study the system performance when more antennas are used. Moreover, the proposal of directional antennas is based on the assumption of a strong LOS component. But, when the LOS component is insignificant, the antenna pattern becomes inaccurate in representing the spatial energy distribution. That is to say, in this case, the spatial footprint of radio energy can be quite different in specific directions. It remains open what is an applicable threshold that can be used to distinguish the environments where directional antennas can work or not?

CHAPTER 5

Joint Design of MIMO MAC and Routing

In Chapter 4, we explore MIMO MAC design in ad hoc networks. It should be cautioned that an isolated cross-layer strategy may yield unintended system performance, when such a strategy interacts with protocols in other layers. For instance, in [49], the authors show that rate adaptive MAC working with minimum hop routing may lead to poorer performance than fixed high-rate IEEE802.11 MAC with minimum hop routing. This points to that a good cross-layer scheme should take into account the interactions across multiple layers. Then, it is natural to ask: how would the system perform when the MIMO MAC interacts with routing? To answer this question, we extend our cross-layer study in Chapter 4 to joint consideration of MAC and routing. In particular, building on our previous study in MAC design, we investigate the impact of MIMO MAC on routing, and characterize the optimal hop length by making use of the information from PHY and MAC layers.

In a multi-hop network, the "end-to-end" transport delay is a key performance metric [52]. Roughly speaking, the transport delay consists of the waiting time and the transmission delay. Consider a multi-hop network, where every user uses a given transmission power, and multi-rate adaptation is conducted based on the channel conditions. Let d denote the hop length, T_d denote the corresponding one-hop delay, and D denote the end-to-end distance a message would travel. Consider a large network where each node can find a relay node

with a lop length close to d. Thus, the total delay T_{tot}, can be approximated as [52]

$$T_{tot} = T_d \frac{D}{d}, \tag{5.1}$$

where $T_d = f(\text{hop length, rate adaptation, contention})$. Given rate adaption and multi-access strategies, the design of the hop length is of great importance to minimize the transport delay.

We note that in multi-hop networks, the interactions between MAC and the hop length design are evident in many aspects. In particular, there is always a tradeoff between the hop length and the transmission rate. Roughly, shorter hops often imply stronger signal strength, and thus higher transmission rates, and vice versa. Moreover, the contention also has a significant impact on the throughput. Indeed, in an ad hoc network using CSMA/CA, the behavior of each user is affected by its neighboring nodes, while the neighboring nodes can also be affected by their neighboring nodes. Thus, in principle, any node can be affected by all other nodes in the entire network. In this chapter, we quantify these effects and characterize the optimal hop length to minimize the end-to-end delay.

For related works, in [30], the authors examine the gain of diversity combining in ad hoc networks using on-demand routing protocols, such as AODV and DSR. In [93], the impact of SNR thresholds on choosing routing paths is studied. Also, much attention has been paid to find optimal hop lengths for multihop networks using ALOHA (e.g., [51], [37], [107]). Worth pointing out is that in ALOHA systems, each user transmits packets in a pre-determined constant probability, whereas in ad hoc networks with CSMA/CA, carrier sensing is used by each user to regulate its transmission and to mitigate the collision. As mentioned above, in such an ad hoc network, the contention of each user affects the entire network, making the optimal design more challenging.

5.1. System Models

We study MIMO ad hoc networks where each node is equipped multiple antennas, and communicates with a given transmission power. Suppose that the relay nodes are chosen by the routing strategy. We assume that SD-MAC (i.e., a MAC protocol taking into account spatial diversity) is used. In what follows, we first present the wireless channel model and give a brief introduction to SD-MAC.

5.1.1. Channel Models

In a wireless communication environment, if the antenna elements are spaced far apart, the spatial channels corresponding to different antenna elements experience more or less independent fading. Throughout this chapter, we assume that the spatial channels experience i.i.d. fading due to distance-related attenuation and Rayleigh fading.

5.1.2. SD-MAC

In Chapter 4, SD-MAC is developed based on the RTS/CTS mechanism in IEEE802.11. Compared with IEEE802.11, the SD-MAC exhibits the following new features: 1) Space-time codes are used for four-way handshaking; 2) For carrier sensing, if the average interference across antenna elements is higher than the threshold, the channel is determined as busy, and the node has to defer its transmission; 3) The reception node estimates the channel condition based on the RTS signals, and feeds the channel state information back to the source node via CTS; and 4) The transmission node adapts data rate for data transmission.

5.2. The End-to-End Delay

Next, we intend to solve the following optimization problem:

$$d^* = \arg\min \left(T_d \frac{D}{d}\right). \tag{5.2}$$

Note that given a fixed transmission power and communication techniques, the coverage range is determined. In practice, the one-hop length cannot be greater than the coverage range. The optimal value d^* can provide insights on how to make use of the gain from the MIMO techniques. For instance, if d^* corresponding to the MIMO ad hoc networks is greater than that of ad hoc networks using single antennas, the gain from MIMO channels can be used for longer hop routing.

5.2.1. The Optimal Hop Length

We note that in many practical cases, the routing is optimized at large time scales. For tractability, we consider a homogenous network with heavy traffic, where all nodes have packets for transmission at all times (a.k.a. saturation condition). In such a scenario, we assume that optimization problem is done for the average transport delay. We first derive the one-hop average throughput of each user by using the method proposed in Chapter 4. Along the line of [16], we assume the "collision" (loss) probability of each packet p to be constant. By analyzing the Markov chain model for CSMA/CA [16], the probability τ that a station transmits in randomly chosen time slot is given by

$$\tau = \frac{2(1-2p)}{(1-2p)(W+1) + pW(1-(2p)^m)}, \tag{5.3}$$

where p is the collision probability, W is the minimum backoff window in terms of backoff slots, and m is the maximum backoff stage.

We next establish the relationship between τ and p in ad hoc networks using spatial diversity. Worth pointing out is that p herein denotes the probability that the RTS packet cannot be received correctly. In the case with spatial diversity, p consists of the probability incurred by both collision and fading. Let p_c and p_f denote the RTS loss probabilities corresponding to collision and fading, respectively. Assume that there are K_a users within the coverage of each node. Let ρ denote the node density. Then, K_a can be approximated as $K_a = \pi A^2 \rho$, where A is the average coverage range.

Needless to say, the packet loss probability due to fading is related to the distance between the nodes. Let $p_f(r)$ denote the packet loss probability due to fading, corresponding to a distance r. For the signal of the desired link, since the hop length is d, the corresponding packet loss probability q_s is given by $p_f(d)$. Next, we calculate pack loss probability due to fading for the signal from contention nodes. Suppose the contention nodes in the neighborhood are uniformly distributed, the probability density function (pdf) of r can to be expressed as

$$f_r(r) = \frac{2r}{A^2}, 0 < r < A, \tag{5.4}$$

Thus, the average packet loss probability due to fading, corresponding to the signals from each neighboring contention node can be given as

$$q_n = \int_0^A p_f(r) f_r(r) dr. \tag{5.5}$$

The probability the RTS packet loss due to the collision can be calculated as

$$\begin{aligned} p_c &= (1-q_s)\Big[1-(1-\tau)\Big((1-\tau)^{K_a-2} + \binom{K_a-2}{1}(1-\tau)^{K_a-2}\tau q_n + \cdots + (\tau q_n)^{K_a-2}\Big)\Big] \\ &= (1-q_s)\Big[1-(1-\tau)(1-\tau+\tau q_n)^{K_a-2}\Big]. \end{aligned} \tag{5.6}$$

Therefore, the loss probability of the RTS packet is given by

$$p = p_c + q_s = (1 - q_s)\Big[1 - (1 - \tau + \tau q_n)^{K_a - 1}\Big] + q_s. \tag{5.7}$$

Combining (5.3) and (5.7), τ and p can be obtained by numerical methods.

Then, we investigate the states of a typical user. Along the line of Chapter 4, we model the events experienced by a typical user (say S_k) into five states. Note that in this case, the distance of the desired the relay link d is deterministic.

p_1**:** S_k is listening, and it detects the channel to be empty:

$$p_1 = (1 - \tau)(1 - \tau + \tau q_s)(1 - \tau + \tau q_n)^{K_a - 2}. \tag{5.8}$$

p_2**:** S_k is listening, and hears a handshaking packet from one of its neighbors:

$$p_2 = (1-\tau)\big[(1-q_s)\tau(1-\tau+\tau q_n)^{K_a-2} + (1-\tau+\tau q_s)(K_a-2)(1-q_n)\tau(1-\tau+\tau q_n)^{K_a-3}\big]. \tag{5.9}$$

p_3**:** S_k is listening, and detects a "collision" among the transmissions of its neighbors.

$$\begin{aligned} p_3 =& (1-\tau)\big\{1 - (1-\tau+\tau q_s)(1-\tau+\tau q_n)^{K_a-2} - \big[(1-q_s)\tau(1-\tau+\tau q_n)^{K_a-2} \\ &+ (1-\tau+\tau q_s)(K_a-2)(1-q_n)\tau(1-\tau+\tau q_n)^{K_a-3}\big]\big\} \\ =& (1-\tau)\big\{1 - \big[\tau(1-\tau+\tau q_n)^{K_a-2} \\ &+ (1-\tau+\tau q_s)(K_a-2)(1-q_n)\tau(1-\tau+\tau q_n)^{K_a-3}\big]\big\} \end{aligned} \tag{5.10}$$

p_4**:** S_k transmits its RTS packet and the transmission is successful. Note that if the RTS packet is faded, there is no corresponding CTS transmission.

$$p_4 = (1 - q_s)\tau(1 - \tau)(1 - \tau + \tau q_n)^{K_a - 2}. \tag{5.11}$$

p_5**:** The RTS packet of S_k cannot be received correctly due to collision or fading:

$$p_5 = \tau\Big[1 - (1-q_s)(1-\tau)(1-\tau+\tau q_n)^{K_a-2}\Big]. \tag{5.12}$$

Assume the payload size of one MAC packet is L. The one-hop average throughput can be expressed as

$$U(d) = \frac{p_4 L}{\epsilon p_1 + T_s(p_2+p_4) + T_c(p_3+p_5)}, \tag{5.13}$$

where ϵ is the duration of a backoff slot, i.e., the minimum time needed for transmission detection, T_s is the average time of a successful transmission, and T_c is the average duration of a collision. In the system with RTS/CTS handshakes, the T_s and T_c are given as

$$\begin{aligned} T_s =& RTS + SIFS + \delta + CTS + SIFS + \delta + OH \\ &+ \mathbb{E}[T_p] + SIFS + \delta + ACK + DIFS + \delta, \end{aligned} \tag{5.14}$$

$$T_c = RTS + DIFS + \delta, \tag{5.15}$$

where δ is the propagation delay, OH is the overhead including both MAC and PHY headers, and $\mathbb{E}[T_p]$ is the average transmission duration for payload. Note that $\mathbb{E}[T_p]$ is a function of hop length, and hence T_s and U.

Now, we are ready to solve the optimization problem in (5.2). For simplicity, consider that there is a M/M/1 queue at each node (see also [15]). The traffic arrival rate of node k is given by

$$\lambda_k = \sum_i \beta_{ik} x_i \tag{5.16}$$

where x_i is the average transmission rate of node i, and β_{ik} denotes the fraction of packets of node i that go to node k. Note that under the saturation conditions, since each node always has packets in its buffer, the service rate becomes the bottleneck. Recall that in a homogeneous network, each node experiences the same statistics. Then, the arrival rate of

one node can be expressed as

$$\lambda = \sum_i \beta_{ik}\mu \tag{5.17}$$

where μ denotes the average throughput of one user. It follows that the utilization factor $\rho = \lambda/\mu$ is a constant. Thus, the average one-hop delay T_d is given by

$$T_d = \frac{1/\mu}{1-\rho} = aT_h, \tag{5.18}$$

where $T_h = 1/\mu$ is one-hop transmission delay, and $a = 1/(1-\rho)$ is constant. As a result, the optimization problem boils down to the following simplified one:

$$d^* = \arg\min \big(T_h \frac{D}{d}\big). \tag{5.19}$$

Let B denote the size of one packet. The one-hop delay of one network packet has the form of

$$T_h(d) = \frac{B}{U(d)}. \tag{5.20}$$

Differentiating $T_h(d)\frac{D}{d}$ with respect to d, we can find that the optimal hop length d^* satisfies

$$\frac{\partial T_h(d)}{\partial d} = \frac{T_h(d)}{d}. \tag{5.21}$$

Therefore, we can get the optimal solution d^* by using numerical methods. Similar to the giant stepping idea in [52], typically the optimal solution is achieved at the point of tangency, as shown in Fig. 5.1, where the curve depicts a trade-off profile between the hop length and the delay.

5.2.2. Numerical Examples

Now, we present some numerical examples to illustrate how to characterize the optimal hop length. We also study the impact of MIMO techniques, rate adaptation strategies, and contention on the hop length.

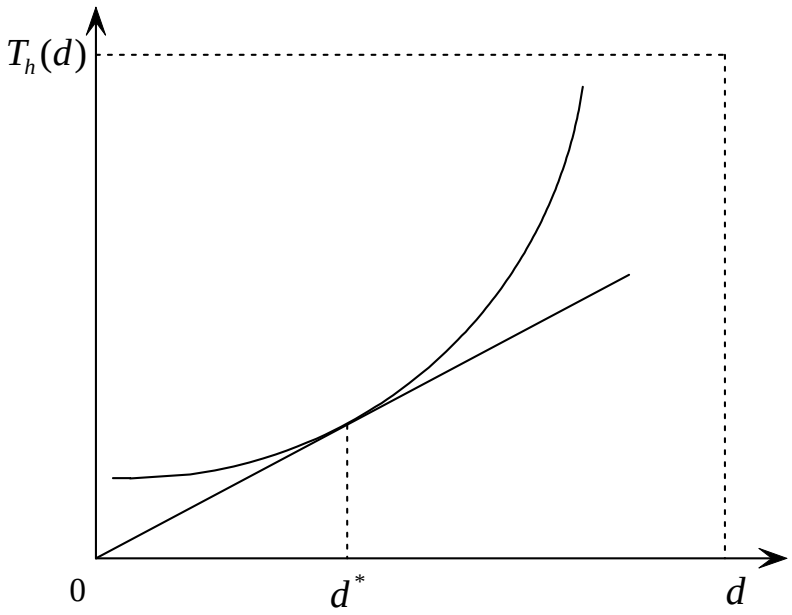

Figure 5.1. Optimal hop length for minimizing delay

The parameters used in the following examples are the same as in Table 4.1 in Chapter 4. We assume that the total transmission power of each node is 20dBm and the path loss factor is 2.5, resulting in a coverage range 200m for transceivers with signal antennas. The rate adaptation strategy uses the practical parameters from "D-Link". The information-bearing data rate and the SNR (after the processing of spatial diversity) have the relationship as in Table 4.2.

For the sake of comparison, we first examine ad hoc networks with single antennas. Fig. 5.2 depicts the one-hop transmission delay of each CBR packet with respect to the hop length, where $K_a = 10$. We observe that the minimum-hop routing using maximal hop length together with rate adaptation, or the minimum-hop routing using short-hop with 11Mbps data rate transmission cannot achieve the best performance. The optimal hop length minimizing the transmission delay can be found to be 100m, using the method in Section 5.2.1. The corresponding average transmission data rate is 5.66Mbps.

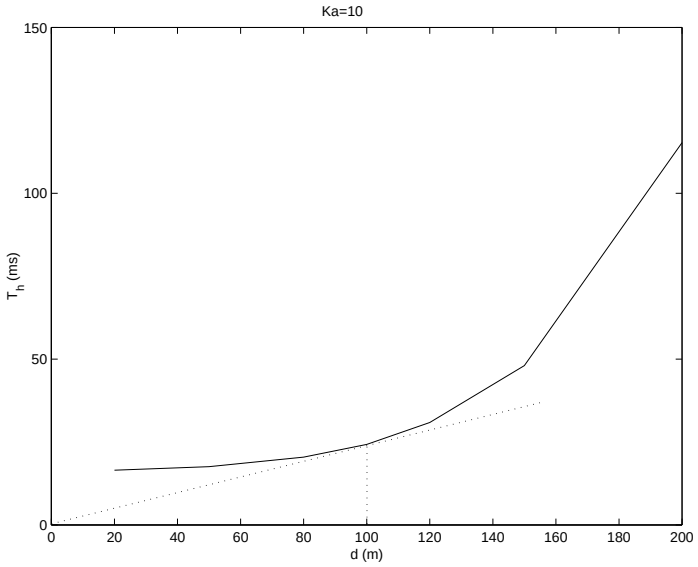

Figure 5.2. One-hop delay vs. hop length (the single antenna case)

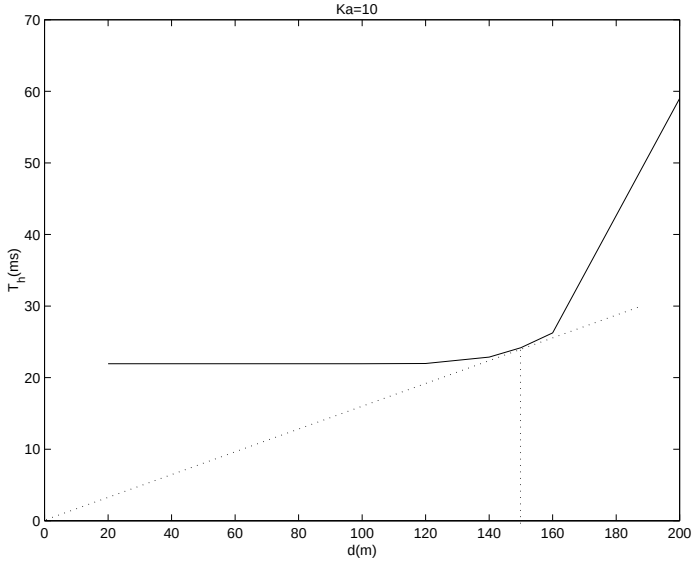

Figure 5.3. One-hop delay vs. hop length (the spatial diversity case)

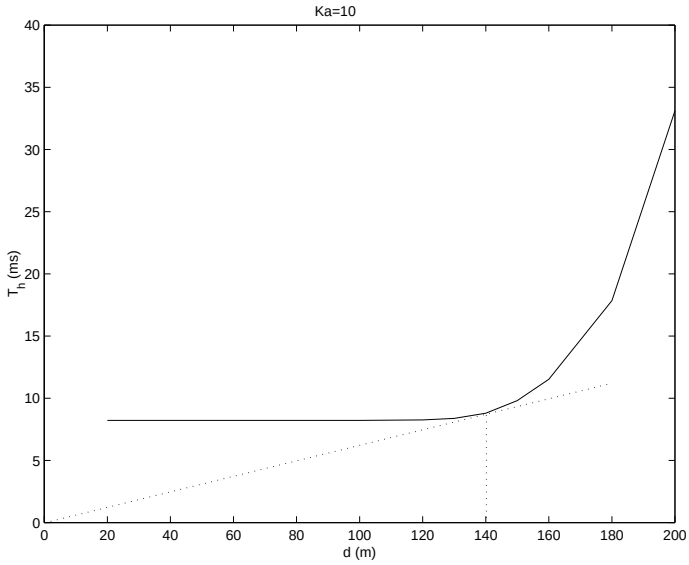

Figure 5.4. One-hop delay vs. hop length (spatial diversity with OAR)

Next, we examine the performance of ad hoc networks with spatial diversity in i.i.d. Rayleigh fading channels. Suppose that 4-element antenna arrays are used and rate adaptation strategy still follows Table 4.2. In Fig. 5.3, we observe that the optimal lop-length is much larger than the one in the single antenna case, as would be expected. Intuitively speaking, this is because the spatial diversity can improve the link quality greatly. That is to say, for the same data rate, longer hops can be used for the spatial diversity case. Moreover, we note that the reduction in delay is not substantial in the short-hop region, because the transmission strategy does not make full use of the improved link quality offered by spatial diversity. According to the MAC protocol, the RTS, CTS, ACK, and overhead of DATA are transmitted in the basic data rate 1Mbps. Under such a strategy, the overhead dominates the improvement from higher rate for data transmission. We conclude that the gain from spatial diversity should be used to achieve longer hop lengths.

From the above example, we notice that the rate adaptation scheme can have a significant impact on the routing. In the following, we examine the performance of different rate

Table 5.1. SNR vs. Coefficient α

Threshold (dB)	0	3	5.5	8.5
Coefficient α	0.1	0.12	0.23	0.33

adaptation schemes. In particular, we first apply the opportunistic auto rate (OAR) scheme to SD-MAC. Suppose that the duration of MAC payload is fixed to be $T_{\text{OAR}} = 8184\mu$s. It follows that the payload of one MAC packet is $T_{\text{OAR}}R$, where R is the transmission rate. As expected, the optimal hop length is achieved at the point of tangency (see Fig. 5.4). In this case, the optimal hop length is 140m. Also, we can see one-hop delay is reduced significantly. It is because the gain from spatial diversity is also smartly used for achieving higher data rate. Next, we use a rate adaption scheme that allows higher and more continuous data rates. Specifically, we assume that the transmission data rate can be expressed as (see also, [98])

$$R(t) = \alpha C(t), 0 < \alpha < 1 \tag{5.22}$$

where $C(t)$ is the channel capacity:

$$C(t) = \mathbf{B}\log_2\big(1 + \text{SNR}(t)\big), \tag{5.23}$$

and $\mathbf{B}$ denotes the bandwidth. Using the parameters in Table 4.2 from D-link, we capture the relationship between the practical data rate and the channel capacity, and the coefficient α is given in Table 5.1. In Fig. 5.5, we can see that this scheme combined with OAR can reduce the transmission delay significantly. Our intuition is that when the rate adaptation schemes allow higher data rates, more flexibility is provided for choosing the hop length and the data rate, leading to better performance. Moreover, we observe that the optimal performance is achieved at the hop length $d^* = 125$m, which is shorter than before.

As mentioned before, the contention in ad hoc networks would affect the throughput of

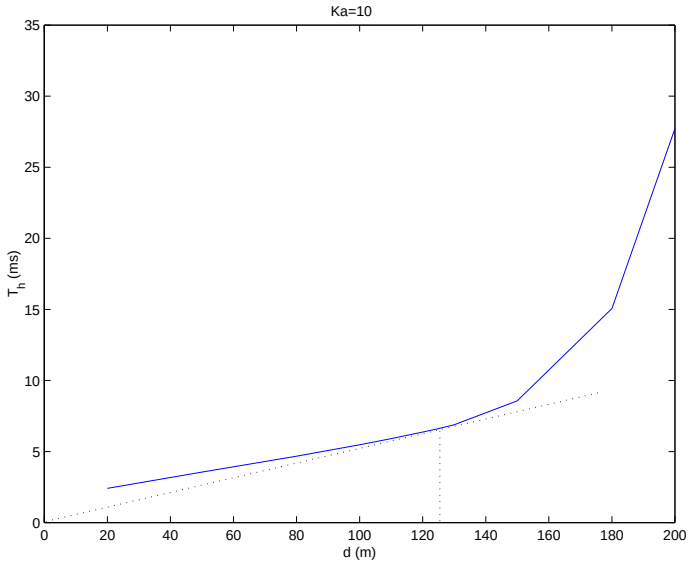

Figure 5.5. One-hop delay vs. hop length (spatial diversity with OAR and higher rates)

each user. We now investigate the impact of contention. We use the same rate adaptation scheme in the above. Fig. 5.6 depicts the optimal hop length in the cases with different node densities, (where $K_a = \pi A^2 \rho$). An interesting observation is that the optimal length is not sensitive to the node density. We also examine the impact of the node density on the optimal hop length, under different rate adaptation schemes; and the same observation on the insensitivity also carries over to those cases. Our intuition is that the one-hop delay is approximately proportional to the node density (as shown in Fig. 5.6), which leads to the same optimal hop length. The insensitivity implies that the optimal hop length be used to optimize the design of path metrics, because such independence makes it easy to predict the performance of ad hoc systems experiencing different node densities, and thus can provide simplicity to the routing design.

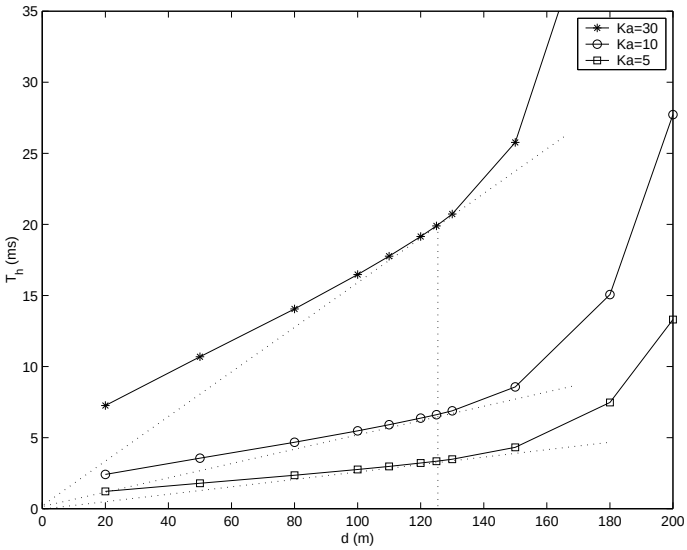

Figure 5.6. One-hop delay vs. hop length (for cases with different node densities)

5.3. Routing with Optimal Hop Lengths

In the following, we examine the performance for ad hoc networks using different hop length. We first outline a routing protocol that makes use of predetermined hop length. Then, we illustrate the gain of using optimal hop length via GloMoSim simulations.

5.3.1. A Routing Algorithm Based on Distance Deviation

To investigate the impact of MIMO techniques, we devise a routing scheme that exploits hop length information. Note that in minimum hop routing, the number of hops is used as a metric, and the routing protocol optimizes the routing tables by choosing the pathes with the smallest possible metric. Along a similar line, we use total distance deviation as a performance metric, i.e.,

$$\Delta_D = \sum_i |d_i - d|, \tag{5.24}$$

Table 5.2. Average end-to-end delay vs. hop length (the single antenna case)

Hop length	50	100	150	200
Delay (s)	2.56	0.57	0.90	3.02

Table 5.3. Average end-to-end delay vs. hop length (the spatial diversity case)

Hop length	50	100	150	200
Delay (s)	0.46	0.13	0.11	0.48

where d is the predetermined hop length, and d_i is the length of the ith hop. Then, a Bellman-Ford algorithm can be used to optimize the routing by minimizing this distance deviation metric.

5.3.2. Numerical Examples

Next, we examine the routing of ad hoc networks with different predetermined hop lengths. Particularly, we evaluate the average end-to-end delay and system throughput performance via GloMoSim. For simplicity, we place 100 nodes in a (500m, 500m) grid plane, and the distance of every two next nodes is 50m. Four 15-minute CBR connections are started simultaneously for far-away nodes with the same total distance.

For comparison, we first examine the average end-to-end delay for ad hoc networks with single antennas/spatial diversity. In Table 5.2, we can see in the ad hoc networks with single antennas, a hop length 100m leads to the best performance. As expected, we observe in Table 5.3 that the ad hoc networks with spatial diversity achieve smaller end-to-end delay. The intuition is that with spatial diversity, the ad hoc network would experience smaller packet loss rate, therefore the retransmission of RTS packet is reduced. Also, the improved SNR by spatial diversity would lead to a higher transmission data rate. As a result, the

Table 5.4. Average system throughput vs. hop length (the single antenna case)

Hop length	50	100	150	200
Throughput(kbps)	260	487	412	174

Table 5.5. Average system throughput vs. hop length (the spatial diversity case)

Hop length	50	100	150	200
Throughput (kbps)	464	634	798	348

end-to-end delay is reduced significantly. Moreover, in contrast to ad hoc networks with single antennas, in the case for ad hoc networks with spatial diversity, a longer hop length (i.e., 150m) yields the best performance. This result coincides with our theoretical analysis.

We also examine the system throughput for ad hoc networks with different predetermined hop lengths. The results corresponding to the single antenna case and spatial diversity case are presented in Table 5.4 and Table 5.5, respectively. We observe that the ad hoc network with spatial diversity achieves much higher (end-to-end) system throughput. Also, as expected, the optimal hop length yields better throughput performance in both of the cases.

From the examples above, we conclude that interaction between MIMO MAC and routing has an important impact on the network performance. There exists optimal hop lengths that can maximize the gain from the MIMO techniques. More specifically, the gain from spatial diversity, can be used not only to increase the transmission data rate, but also to enlarge the hop length. Such an optimal point can be found by using the proposed approach above.

It should also be cautioned that the optimal hop length is derived from the case study. Although the proposed method can be applied to any homogeneous ad hoc networks, the ab-

solute value of the optimal hop length depends on the network topology, the rate adaptation strategy, the node density, and the MAC protocol.

5.4. Conclusions

In this chapter, we study joint design of MIMO MAC and routing. We first examine the impact of MIMO MAC on routing. We identify the factors that affect the transmission delay in MIMO ad hoc networks, and quantify their impacts. Building on this, we characterize the optimal hop length to minimize the end-to-end delay. The impacts of different rate adaption and MAC schemes on the optimal hop length are also investigated. We observe that to minimize the delay, the gain from spatial diversity can be used not only to increase the data rate, but also to enlarge the hop length. Moreover, we illustrate the existence of the optimal hop length via GloMoSim simulations. Our results show that the proposed algorithm is effective in determining the optimal hop length.

CHAPTER 6

Conclusions

Fuelled by the great success of Internet and wireless telephony communications, there has been a dramatic demand for wireless data access. To fulfill such extensive demands for wireless services, it is of great importance to develop new methods in modeling, optimization, and analysis to improve the spectrum efficiency. We note that within the OSI layered architecture, it is possible to yield significant gains, if the system optimizes the performance by making use of the interaction across protocol layers. Thus motivated, in this thesis, we focus on cross-layer design for resource allocation in wireless data networks; and a main objective is to improve the system performance by incorporating the information from the physical layer and the network layer into the design of resource management.

First, we study data communications in the downlink of CDMA systems. We exploit a predictive temporal structure of the multi-access interference (MAI) for adaptive resource allocation, particularly rate control and admission control. Specifically, we establish that the MAI process in CDMA data networks is "self-similar". The MAI self-similarity indicates that there exists a nontrivial predictive MAI structure at larger time scales, which enables more accurate interference prediction. Therefore, we devise a multiple time-scale interference predictor. Rate adaptation is carried out based on the predicted MAI level. Our numerical results show that this rate control scheme achieves better performance than

that based on the packet-level MAI prediction only. Building on the rate adaptation, we then devise a joint rate control and admission control scheme. Our results reveal that this admission control scheme may be very useful for bursty data applications.

Then, we study medium access control in opportunistic communication systems. We first propose a traffic-aided smooth admission control (SAC) scheme that aims to guarantee throughput provisioning. Simply put, in the SAC scheme, the admission decision is "spread" over a trial period, by increasing gradually the amount of the time resource allocated to incoming users. Specifically, using the modified weighted proportional fair (WPF) scheduling, we devise a QoS driven weight adaptation algorithm, and the weights assigned to new users are increased in a guarded manner. Then, an admission decision is made based on the measured throughput within a time-out window. Our results show that the proposed SAC scheme works well in opportunistic communication systems. Next, we explore the possibility of reducing the completion time by incorporating traffic information into opportunistic scheduling. In particular, we first establish convexity properties for opportunistic scheduling with the file size information. Then, we develop new traffic aided opportunistic scheduling (TAOS) schemes by making use of file size information and channel variation in a unified manner. We also derive lower bounds and upper bounds on the total completion time. Our results show that the proposed TAOS schemes can yield significant reduction in the total completion time.

Next, we turn our attention to ad hoc networks using multiple antennas. Specially, we study MIMO ad hoc networks in heavy-loaded regimes. In particular, we investigate the utility of spatial diversity for MAC design, when the spatial channels experience independent fading. We develop the corresponding MAC protocol, namely SD-MAC. In contrast, when the wireless channel has a strong line of sight, we exploit smart antennas to improve spatial

reuse in ad hoc networks. We propose to use directional listening to resolve the hidden terminal problem incurred by the asymmetry in antenna gain, and demonstrate its practicality. Building on this, we develop a MAC protocol using directional antennas, namely DA-MAC. The proposed DA-MAC takes into account a general directional antenna model with sidelobes, and makes use of directional listening, directional transmission, and directional reception. Moreover, we develop analytical methods for characterizing the saturation throughput. The proposed methods take into account the impact of MIMO techniques, fading and contention. We then analyze saturation throughput for ad hoc networks with these two multiple-antenna techniques, via the analytical methods and GloMoSim simulations. Our results show that the cross-layer design using smart antennas and spatial diversity can yield significant gains in ad hoc networks.

Finally, we study joint optimization for MAC design and routing in MIMO ad hoc networks. More specifically, we focus on the end-to-end delay of a homogeneous ad hoc network. We characterize the impact of the hop length, rate adaptation, and contention levels on the delay quantitatively. Building on this, we present an approach to find the optimal hop length, in the sense of minimizing the end-to-end delay. Our results show that there exists an optimal solution to the delay, and the gain from spatial diversity can be used to improve both transmission rates and hop lengths. The impact of rate adaptation and contention is also examined.

In this thesis, we have presented our main results on the cross-layer design for resource allocation in wireless data networks. (Most of the results in this thesis have been published or submitted.) Specifically, the results on resource allocation in CDMA wireless data systems in Chapter 2, have been published in [39], [113], [114]. The smooth admission control in Chapter 3, together with relay-aided opportunistic scheduling, has been published in [40],

[38]. The results on traffic-aided opportunistic scheduling schemes in Chapter 3 have been published partially in [41], [42], [43]. The results on MIMO ad hoc networks in Chapter 4 and 5 have also been submitted for publication.

In summary, this thesis research has made some promising steps towards a cross-layer design framework for resource allocation in wireless data networks. The proposed schemes make use of the information across different protocol layers in a more unified manner, and can improve the spectrum efficiency significantly for wireless data communications. Also, the schemes presented in this thesis are relatively easy to implement, and are potentially useful for practical systems.

REFERENCES

[1] "IEEE standard for wireless LAN medium access control (MAC) and physical layer (PHY) specifications," Nov. 1997.

[2] R. J. Adler, R. E. Feldman, and M. S. Taqqu, *A practical Guide to heavy Tails: Statistical Techniques and Applications*. Boston: Birkhauser, 1998.

[3] R. Agrawal, A. Bedekar, R. J. La, R. Pazhyannur, and V. Subramanian, "Class and channel condition based scheduler for EDGE/GPRS," in *Modeling and Design of Wireless Networks, Proceeding of SPIE*, pp. 59–68, 2001.

[4] S. Alamouti, "A simple transmit diversity technique for wireless communications," *IEEE Journal on Selected Area in Communications*, vol. 16, pp. 1451–1458, Oct. 1998.

[5] P. Ameigeiras, "Performance evaluation of packet scheduling with qos guarantees in hsdpa." http://cpk.auc.dk/ persa/Thursday_seminars.html, Feb. 2003.

[6] M. Andersin, Z. Rosberg, and J. Zander, "Soft and safe admission control in cellular networks," *IEEE/ACM Transactions on Networking*, vol. 5, pp. 255–265, Apr. 1997.

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

[8] D. Ayyagari and A. Ephremides, "Power control based admission algorithms for maximing throughput in DS-CDMA networks with multimedia traffic," in *WCNC'99*, 1999.

[9] H. Balakrishnan, V. Padmanabhan, S. Seshan, and R. H. Katz, "A comparison of mechanisms for improving TCP performance over wireless links," *IEEE/ACM Transactions on Networking*, Dec. 1997.

[10] C. Balanis, *Antenna Theory Analysis and Design.* New York: John Wiley & Sons Inc., 1997.

[11] N. Bambos, S. C. Chen, and G. J. Pottie, "Channel access algorithms with active link protection for wireless communication networks with power control," *IEEE/ACM Trans. Networking*, vol. 5, pp. 583–597, Oct. 2000.

[12] L. Bao and J. Garcia-Luna-Aceves, "Transmission scheduling in ad hoc networks with directional antennas," in *Proc. IEEE/ACM MobiCom 2002*, Sept. 2002.

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

[14] J. Beran, *Statistics for Long-Memory Processes.* Chapman & Hall/CRC, 1994.

[15] D. Bertsekas and R. Gallager, *Data Networks.* Prentice Hall, 2000.

[16] G. Bianchi, "Performance analysis of the IEEE802.11 distributed coordination function," *IEEE Journal on Selected Area in Communications*, vol. 18, pp. 535–547, Mar. 2000.

[17] P. Billingsley, *Convergence of Probability Measures.* John Wiley & Sons, Inc., 1968.

[18] H. Bölcskei, "Fundamental performance tradeoffs in coherent MIMO signaling," *Private Communication*, 2003.

[19] S. Borst, "User-level performance of channel-aware scheduling algotirthms in wireless data networks," in *Proc. IEEE INFOCOM'03*, 2003.

[20] S. Borst and P. Whiting, "Dynamic rate control algorithms for HDR throughput optimization," in *Proc. IEEE INFOCOM'01*, pp. 976–985, 2001.

[21] S. Choi and K. G. Shin, "An uplink CDMA system architecture with diverse QoS guarantees for heterogeneous traffic," *IEEE/ACM Transactions on Networking*, vol. 7, pp. 616–628, Oct. 1999.

[22] R. R. Choudhary, X. Yang, R. Ramanathan, and N. H. Vaidya, "Using directional antennas for media access control in ad-hoc networks," in *Proceedings of the IEEE/ACM MobiCom Conference*, 2002.

[23] C. Comaniciu, N. Mandayam, D. Famolari, and P. Agrawal, "Wireless access to the World Wide Web in an integrated CDMA system," preprint.

[24] M. E. Crovella and A. Bestavros, "Self-similarity in World Wide Web traffic: Evidence and possible causes," *IEEE/ACM Transactions on Networking*, vol. 5, pp. 835–846, Dec. 1997.

[25] M. E. Crovella, R. Frangioso, and M. Harchol-Balter, "Connection scheduling in Web servers," in *USENIX Symposium on Internet Technologies and Systems*, Oct. 1999.

[26] H. A. David, *Order Statistics*. John Wiley & Sons Inc., 2nd ed., 1981.

[27] S. Floyd, "TCP and explicit congestion notification," *ACM Computer Communication Review*, vol. 24, pp. 10–23, Oct. 1994.

[28] G. J. Foschini and M. J. Gans, "On limits of wireless communications in a fading environment when using multiple antennas," *Wireless Personal Communications*, vol. 6, no. 3, pp. 311–335, Mar. 1998.

[29] S. Furman and M. Gerla, "The design of a spatial diversity model to mitigate narrowband and broadband interference in DSSS ad hoc networks," in *Proc. ICC 2003*, May 2003.

[30] S. Furman, J. Martin, and R. Bagrodia, "A comparative study on the effects of spatial diversity in ad hoc networks using on-demand routing protocols," in *Proceedings of IEEE ICPPW'02*, Aug. 2002.

[31] D. Gesbert, M. Shafi, D. shan Shiu, P. J. Smith, and A. Naguib, "From theory to practice: An overview of MIMO spacetime coded wireless systems," *IEEE Journal on Selected Area in Communications*, vol. 21, pp. 281–302, Apr. 2003.

[32] K. S. Gilhousen, I. M. Jacobs, R. Padovani, A. J. Viterbi, L. A. Weaver Jr., and C. E. Wheatley III, "On the capacity of a cellular CDMA system," *IEEE Transactions on Vehicular Technology*, vol. 40, pp. 303–312, May 1991.

[33] M. Gudmundson, "Correlation model for shadow fading in mobile radio systems," *Electronics Letters*, vol. 27, pp. 2145–2146, 1991.

[34] M. Harchol-Balter, N. Bansal, B. Schroeder, and M. Agrawal, "Size-based scheduling to improve Web performance," *ACM Transactions on Computer Systems*, vol. 21, no. 2, pp. 207–223, May 2003.

[35] B. M. Hochwald, T. L. Marzetta, and V. Tarokh, "Multi-antenna channel hardening and its implications for rate feedback and scheduling," *IEEE Transactions on Information Theory*, 2004.

[36] M. L. Honig and J. B. Kim, "Resource allocation for packet data transmission in DS-CDMA," in *Proc. 33th Allerton Conf.*, 1995.

[37] T. Hou and V. Li, "Transmission range control in multihop packet radio networks," *IEEE Transaction on Communications*, pp. 38–44, Jan. 1986.

[38] M. Hu and J. Zhang, "Opportunistic multi-access: Multiuser diversity, relay-aided opportunistic scheduling, and traffic-aided smooth admission control," *Mobile Networks and Applications (MONET).* to appear.

[39] M. Hu and J. Zhang, "Rate adaptation for bursty data transmission in CDMA networks," in *Proc. IEEE 35th Asilomar Conference*, (Monterey, CA), pp. 1718–1722, Nov. 2001.

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

[41] M. Hu, J. Zhang, and J. Sadowsky, "Size-aided opportunistic scheduling in wireless networks," in *Proc. IEEE Globecom 2003*, (San Francisco, CA), pp. 538–542, Dec. 2003.

[42] M. Hu, J. Zhang, and J. Sadowsky, "Traffic aided opportunistic scheduling for wireless networks: Algorithms and performance bounds," in *IEEE Infocom 2004*, (Hongkong), Mar. 2004.

[43] M. Hu, J. Zhang, and J. Sadowsky, "Traffic aided opportunistic scheduling for wireless networks: Algorithms and performance bounds," *Computer Networks Journal.* to appear.

[44] S. Jafar and A. Goldsmith, "Optimal rate and power adaptation for multirate CDMA," in *VTC Fall 2000*, 2000.

[45] W. C. Jakes, *Microwave Mobile Communication.* Piscataway, NJ: IEEE Press, second ed., 1994.

[46] A. Jalali, R. Padovani, and R. Pankaj, "Data throughput of CDMA-HDR a high efficiency-high data rate personal communication wireless system," in *IEEE Vehicular Technology Conference Proceedings*, vol. 3, pp. 1854–1858, 2000.

[47] N. Joshi, S. R. Kadaba, S. Patel, and G. S. Sundaram, "Downlink scheduling in CDMA data networks," in *Proc. IEEE/ACM MobiCom 2000*, pp. 179–190, 2000.

[48] D. Karger, C. Stein, and J. Wein, "Scheduling algorithms," in *Handbook of Algorithms and Theory of Computation* (M. J. Atallah, ed.), CRC Press, 1997.

[49] V. Kawadia and P. R. Kumar, "A cautionary perspective on cross layer design," *preprint*, 2003.

[50] F. Kelly, "Charging and rate control for elastic traffic," *European Transactions on Telecommunications*, vol. 8, pp. 33–37, 1997.

[51] L. Kleinrock and J. Silvester, "Optimum transmission radii for packet radio networks or why six is a magic number," in *IEEE National Telecommunications Conference*, Dec. 1978.

[52] L. Kleinrock, "On some principles of nomadic computing and multi-access communications," *IEEE Communications Magazine*, pp. 46–50, July 2000.

[53] R. Knopp and P. Humlet, "Information capacity and power control in single cell multiuser communications," in *Proc. IEEE ICC 95*, vol. 1, pp. 331–335, June 1995.

[54] T. Korakis, G. Jakllari, and L. Tassiulas, "A MAC protocol for full exploitation of directional antennas in ad-hoc wireless networks," in *Proceedings of the IEEE/ACM MobiHoc Conference*, pp. 98–107, June 2003.

[55] D. Levine, I. Akyildiz, and M. Naghshineh, "A resource estimation and call admission algorithm for wireless multimedia networks using the shadow cluster concept," *IEEE/ACM Transactions on Networking*, vol. 5, pp. 1–12, 1997.

[56] J. C. Liberti and T. S. Rappaport, *Smart Antennas for Wireless Communications.* Prentice Hall, 1999.

[57] T. Liu and J. Silvester, "Joint admission/congestion control for wireless CDMA systems supporting integrated services," *IEEE Journal on Selected Areas in Communications*, vol. 12, pp. 845–857, Aug. 1998.

[58] X. Liu, E. K. Chong, and N. B. Shroff, "A framework for opportunistic scheduling in wireless networks," *Computer Networks*, vol. 41, no. 4, pp. 451–474, Mar. 2003.

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

[60] D. Mitra and J. A. Morrison, "A distributed power control algorithm for bursty transmissions on cellular, spread spectrum wireless networks," in *Proc. 5th WINLAB Workshop on Third Generation Wireless Information Networks* (J. M. Holtzman, ed.), pp. 201–212, Kluwer Academic Publishers, 1996.

[61] A. Muqattash and M. Krunz, "Power controlled dual channel (PCDC) medium access protocol for wireless ad hoc networks," in *Proc. IEEE INFOCOM'03*, Apr. 2003.

[62] S. Nanda, K. Balachandran, and S. Kumar, "Adaptation techniques in wireless packet data services," *IEEE Communications Magzine*, pp. 54–64, Jan. 2000.

[63] T. Nandagopal, T.-E. Kim, X. Gao, and V. Bharghavan, "Achieving mac layer fairness in wireless packet networks," in *ACM Mobicom'00*, (Boston, MA), Aug. 2000.

[64] A. Narula, M. D. Trott, and G. W. Wornell, "Performance limits of coded diversity methods for transmitter antenna arrays," *IEEE Trans. Inform. Theory*, vol. 45, pp. 2418–2433, Nov. 1999.

[65] S. Oh and K. Wasserman, "Adaptive resource allocation in power contrained CDMA mobile networks," in *WCNC'99*, 1999.

[66] S.-J. Oh and K. Wasserman, "Dynamic spreading gain control in multi-service CDMA networks," *IEEE Journal on Selected Areas in Communications*, vol. 17, pp. 918–927, 1999.

[67] T. Ojanperä and R. Prasad, *Wideband CDMA for Third Generation Mobile Communications*. Boston: Artech House, 1998.

[68] S. A. M. Ostring, H. R. Sirisena, and I. Hudson, "Rate control of elastic connections competing with long-range dependent network traffic," *IEEE Transactions on Communications*, vol. 48, pp. 1092–1111, 2001.

[69] K. Park and W. Willinger, *Self-Similar network Traffic and Performance Evaluation*. John Wiley & Sons, Inc., 2000.

[70] S. Parkvall, E. Dahlman, P. Frenger, P. Beming, and M. Persson, "The high speed packet data evolution of WCDMA," in *Proc. IEEE VTC Spring*, pp. 2287–2291, 2001.

[71] V. Paxson, "Empirically derived analytic models of wide-area TCP connections," *IEEE/ACM Transactions on Networking*, vol. 20, no. 4, pp. 316–336, Aug. 1994.

[72] R. L. Peterson, R. E. Ziemer, and D. E. Borth, *Introduction to Spread Spectrum Communications*. Prentice Hall International, Inc., 1995.

[73] S. Ramakrishna and J. M. Holtzman, "A scheme for throughput maximization in a dual-class CDMA system," *IEEE Journal on Selected Areas in Communications*, vol. 16, pp. 830–844, 1998.

[74] T. S. Rappaport, *Wireless Communications: Principles and Practice*. New Jersey: Prentice Hall, 1996.

[75] R. H. Riedi and W. Willinger, "Toward an improved understanding of network traffic dynamics," in *Self-Similar network Traffic and Performance Evaluation* (K. Park and W. Willinger, eds.), pp. 507–530, John Wiley & Sons, Inc., 2000.

[76] H. L. Royden, *Real Analysis*. Prentice Hall, Inc., third ed., 1988.

[77] B. Sadeghi, V. Kanodia, A. Sabharwal, and E. Knightly., "Opportunistic media access for multirate ad hoc networks," in *Proceedings of the IEEE/ACM Mobicom Conference*, Sept. 2002.

[78] A. Sampath and J. M. Holtzman, "Access control of data in integerated voice/data CDMA systems: Benifits and tradeoffs," *IEEE Journal on Selected Area in Communications*, vol. 15, pp. 1511–1526, 1997.

[79] M. Sánchez, T. Giles, and J. Zander, "CSMA/CA with beam forming antennas in multi-hop packet radio," in *Proceedings of the Swedish Workshop on Wireless Ad-hoc Networks*, (Stockholm, Johannesbergs Slott), Mar. 2001.

[80] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas Propagation*, vol. 34, pp. 276–280, 1986.

[81] L. E. Schrage and L. W. Miller, "The queue M/G/1 with the shortest remaining processing time discipline," *Operations Research*, vol. 14, pp. 670–684, 1966.

[82] S. Shakkottai, R. Srikant, and A. L. Stolyar, "Pathwise optimality and state space collapse for the exponential rule," in *Proceedings of IEEE Symposium on Information Theory*, July 2002.

[83] Z. Shao and U. Madhow, "Scheduling heavy-tailed traffic over the wireless Internet," in *Proc. IEEE Vehicular Technology Conference*, vol. 2, pp. 1158–1162, Sept. 2002.

[84] D. Shen and C. Ji, "Admission control of heterogeneous traffic for third generation CDMA network," in *Proceedings of IEEE INFOCOM*, (Israel), 2000.

[85] A. Shiryayev, *Probability*. Springer-Verlag, 1984.

[86] L. Song and N. Mandayam, "Hierarchical SIR and rate control on the forward link for CDMA data users under delay and error constraints," *IEEE Journal on Selected Areas in Communications*, pp. 1871–1882, Oct. 2001.

[87] G. L. Stüber, *Principles of Mobile Communication*. Kluwer Academic Publishers, 2nd ed., 2001.

[88] K. Sundaresan and R. Sivakumar, "On the medium access control problem in adhoc networks with smart antennas," in *Posters of the IEEE/ACM MobiHoc Conference*, June 2003.

[89] M. S. Taqqu and V. Teverovsky, "On estimating the intensity of long-range dependence in finite and infinite variance time series," in *A Practical Guide to Heavy Tails: Staticstical Techniques and Applications* (R. J. Adler, R. E. Feldman, and M. S. Taqqu, eds.), pp. 177–217, Birkhäuser, 1998.

[90] M. S. Taqqu, W. Willinger, and R. Sherman, "Proof of a fundamental result in self-similar traffic modeling," *Computer Communications Reviews*, vol. 27, no. 2, pp. 5–23, 1997.

[91] V. Tarokh, N. Seshadri, and A. R. Calderbank, "Space-time codes for high data rate wireless communication: Performance criterion and code construction," *IEEE Transactions on Information Theory*, vol. 44, no. 2, pp. 744–765, Mar. 1998.

[92] I. E. Telatar, "Capacity of multi-antenna gaussian channels," *European Transactions on Telecommunications*, vol. 10, pp. 585–595, Dec. 1999.

[93] S. Toumpis, "Capacity of wireless ad hoc networks and design principles." http://venus.ftw.tuwien.ac.at/ftw/events/telekommunikationsforum/SS2003, Apr. 2003.

[94] D. Tse, "Multiuser diversity in wireless networks." http://degas.eecs.berkeley.edu/~dtse/pub.html, Apr. 2001.

[95] B. S. Tsybakov, "File transmission over wireless fast fading downlink," *IEEE Transctions on Information Theory*, vol. 48, pp. 2323–2337, Aug. 2002.

[96] T. Tuan and K. Park, "Multiple time scale congestion control for self-similar network traffic," *Performance Evaluation*, vol. 36, pp. 359–386, Aug. 1999.

[97] UCLA, "http://pcl.cs.ucla.edu/projects/glomosim/," 2003.

[98] E. Uysal-Biyikoglu, B. Prabhakar, and A. E. Gamal, "Energy-efficient packet transmission over a wireless link," *IEEE Transactions on Networking*, no. 4, pp. 487–499, Aug. 2002.

[99] V. V. Veeravalli, "The coding-spreading tradeoffs in CDMA systems," in *Proc. 37rd Allerton Conf.*, Sept. 1999.

[100] S. Verdú, *Multiuser Detection*. Cambridge University Press, 1998.

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

[102] A. J. Viterbi, *CDMA—Principles of Spread Spectrum Communications*. Addison–Wesley, 1995.

[103] H. Wang and N. B. Mandayam, "Opportunistic file transfers over fading channels under energy and delay constraints," *Preprint*, 2002.

[104] W. Whitt, *Stochastic-Process Limits*. New York: Springer-Verlag, 2002.

[105] W. Willinger, V. Paxson, and M. S. Taqqu, "Self-similarity and heavy tails: Structual modeling of network traffic," in *A Practical Guide to Heavy Tails: Staticstical Techniques and Applications* (R. J. Adler, R. E. Feldman, and M. S. Taqqu, eds.), pp. 27–53, Birkhäuser, 1998.

[106] W. Willinger, M. S. Taqqu, R. Sherman, and D. V. Wilson, "Self-similarity through high-variability: Statistical analysis of Ethernet LAN traffic at the source level," *IEEE/ACM Transactions on Networking*, vol. 5, pp. 71–86, Feb. 1997.

[107] V. Wong and C. Leung, "Transmission strategies in multihop mobile packet radio networks," in *Canadian Conference on Electrical and Computer Engineering*, pp. 1004–1008, Sept. 1993.

[108] M. Xiao, N. B. Shroff, and E. K. P. Chong, "Distributed connection admission control for power-controlled cellular wireless systems," in *Proc. 37rd Allerton Conf.*, Sept. 1999.

[109] L. Xu, X. Shen, and J. W. Mark, "Dynamic bandwidth allocation with fair scheduling for WCDMA systems," *IEEE Wireless Communications*, vol. 9, pp. 26–32, Apr. 2002.

[110] W. Yang and E. Geraniotis, "Admission policies for integrated voice and data traffic in CDMA packet radio networks," *IEEE Journal on Selected Areas in Communications*, vol. 12, pp. 654–664, May 1994.

[111] J. Zhang, E. K. P. Chong, and I. Kontoyiannis, "Unified spatial diversity combining and power allocation schemes for CDMA systems," *IEEE Journal on Selected Areas in Communications*, pp. 1276–1288, 2001.

[112] J. Zhang, E. K. P. Chong, and D. N. C. Tse, "Output MAI distributions of linear MMSE multiuser receivers in DS-CDMA systems," *IEEE Transactions on Information Theory*, vol. 47, pp. 1128–1144, Mar. 2001.

[113] J. Zhang, M. Hu, and N. B. Shroff, "Bursty data over CDMA: MAI self similarity, rate control and admission control," in *Proc. IEEE INFOCOM'02*, (New York), pp. 391–399, 2002.

[114] J. Zhang, M. Hu, and N. B. Shroff, "Bursty data over CDMA: MAI self similarity, rate control and admission control," *Computer Networks Journal*, pp. 779–795, 2003.

[115] J. Zhang and T. Konstantopoulos, "Self-similarity of multi-access interference processes in multimedia CDMA networks," in *Proceedings of International Symposium on Information Theory*, (Lausanne, Switzerland), p. 47, 2002.

[116] J. Zhang and T. Konstantopoulos, "Multi-access interference process is self-similar in multimedia CDMA cellular networks," *IEEE Transactions on Information Theory*. to apprear.

[117] L. Zheng and D. N. Tse, "Diversity and multiplexing: a fundamental tradeoff in multiple-antenna channels," *IEEE Trans. Inform. Theory*, vol. 49, pp. 1073–1096, May 2003.