QoS-DRIVEN ADAPTIVE RESOURCE ALLOCATION FOR MOBILE WIRELESS COMMUNICATIONS AND NETWORKS

A Dissertation

by

JIA TANG

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

December 2006

Major Subject: Computer Engineering

QoS-DRIVEN ADAPTIVE RESOURCE ALLOCATION FOR MOBILE WIRELESS COMMUNICATIONS AND NETWORKS

A Dissertation

by

JIA TANG

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Approved by:

Chair of Committee,	Xi Zhang
Committee Members,	Costas N. Georghiades
	A. L. Narasimha Reddy
	Jianer Chen
Head of Department,	Costas N. Georghiades

December 2006

Major Subject: Computer Engineering

ABSTRACT

QoS-Driven Adaptive Resource Allocation for Mobile Wireless Communications and Networks. (December 2006)

Jia Tang, B.S. Xi'an Jiaotong University, Xi'an, China

Chair of Advisory Committee: Dr. Xi Zhang

Quality-of-service (QoS) guarantees will play a critically important role in future mobile wireless networks. In this dissertation, we study a set of QoS-driven resource allocation problems for mobile wireless communications and networks.

In the first part of this dissertation, we investigate resource allocation schemes for statistical QoS provisioning. The schemes aim at maximizing the system/network throughput subject to a given queuing delay constraint. To achieve this goal, we integrate the information theory with the concept of *effective capacity* and develop a unified framework for resource allocation. Applying the above framework, we consider a number of system infrastructures, including single channel, parallel channel, cellular, and cooperative relay systems and networks, respectively. In addition, we also investigate the impact of imperfect channel-state information (CSI) on QoS provisioning. The resource allocation problems can be solved efficiently by the convex optimization approach, where closed-form allocation policies are obtained for different application scenarios.

Our analyses reveal an important fact that there exists a fundamental tradeoff between throughput and QoS provisioning. In particular, when the delay constraint becomes loose, the optimal resource allocation policy converges to the *water-filling* scheme, where ergodic capacity can be achieved. On the other hand, when the QoS constraint gets stringent, the optimal policy converges to the *channel inver-*

sion scheme under which the system operates at a constant rate and the zero-outage capacity can be achieved.

In the second part of this dissertation, we study adaptive antenna selection for multiple-input-multiple-output (MIMO) communication systems. System resources such as subcarriers, antennas and power are allocated dynamically to minimize the symbol-error rate (SER), which is the key QoS metric at the physical layer. We propose a selection diversity scheme for MIMO multicarrier direct-sequence code-division-multiple-access (MC DS-CDMA) systems and analyze the error performance of the system when considering CSI feedback delay and feedback errors. Moreover, we propose a joint antenna selection and power allocation scheme for space-time block code (STBC) systems. The error performance is derived when taking the CSI feedback delay into account. Our numerical results show that when feedback delay comes into play, a tradeoff between performance and robustness can be achieved by dynamically allocating power across transmit antennas.

To my family

ACKNOWLEDGMENTS

First of all, I would like to express my deep appreciation to my Ph.D. advisor Professor Xi Zhang for his guidance and contribution to this research. Without his advice, deep insight, and the offer of funding support, this dissertation would not have been possible. His commitment to high-quality research work through hardwork impressed me most and will motivate me for my future career. He gives students freedom to conduct research while providing necessary help. Furthermore, he is willing to share his knowledge and career experience with me and encourage me whenever needed. He is not only an adviser and a professor, but also a friend.

I would also like to thank Professor Costas N. Georghiades, Professor A. L. Narasimha Reddy and Professor Jianer Chen for serving on my dissertation committee and for offering me valuable advice on my research. Special thanks are given to Professor Costas N. Georghiades, who not only gave me educational guidance on modulation theory and information theory, but also offered generous support throughout my Ph.D. studies. I am also grateful to Professor Jean-François Chamberland for his generous and persistent help on my research, especially on stochastic process and large deviation theory. Moreover, I have learned a lot from Professor Krishna R. Narayanan and from Professor Tie Liu both on and off classes. I also want to thank Professor Chanan Singh, Professor Garng M. Huang, and Professor Weiping Shi for their encouragement and kind help in my career development.

My colleagues and friends have made my life at TAMU enjoyable and memorable. I especially thank Qinghe Du for his friendship, insightful comments and discussion on my research, from XJTU to TAMU. I want to take this opportunity to thank my friends at TAMU, Hang Su, Lingjia Liu, Lin Xie, Wei-Yu Chen, Qiang Li, Yik-Chung Wu, Jing Jiang, Yalin Jin, Yunli Xiong, Qiang Hu, Jin Yang, Jimin Lin, Xudong Xu, Dan Ye, Tao Yang, and Peichao Zhang, and Peimin Zhu, among others. I would also like to thank my other friends in the U.S. and China, who helped and encouraged me before and after I came to the U.S., Tan Gao, Xin Liu, Yuanliang Meng, Shuang Hou, Lanbin Ren, Minghui Dong, Shi Pu, Xin Zhou, Yabo Li, Jiangxuan Du, Yonghua Lin, Zhao Du, Yunwen Su, Yan Han, Xiaohai Wang, Lei Tao, Jiangfeng Di Wu, Yijia Yao, Yu Han, Lie Cao, Jinghua Zhang, Lei Lei, Xiaojun Wu, Aigang Feng, Wenjie Wang, Ke Deng, Jianguo Zhang, Minhao Tie, Chao Du, Professor Ruiqing Du, Professor Dongyin Cui, Professor Yong Qi, among others. With their support, my life at TAMU during the last three years has been colorful.

In particular, I would like to thank my advisor in China, Professor Qinye Yin, for his guidance and support. Without the graduate study on wireless communications and signal processing at Xi'an Jiaotong University under his supervision, it would have been impossible for me to pursue my Ph.D. at TAMU.

Finally, but most importantly, this dissertation is dedicated to all of my family members, my parents, my grandparents, my uncles and aunts, my brothers and sisters, and my wife, for their infinite love and support.

The funding for this research was supported in part by the National Science Foundation CAREER Award under Grant ECS-0348694, and in part by the Fouraker Graduate Research Fellowship Award from Department of Electrical and Computer Engineering, Texas A&M University.

TABLE OF CONTENTS

LIST OF TABLES

TABLE Page

LIST OF FIGURES

FIGURE Page

FIGURE Page

FIGURE Page

FIGURE Page

CHAPTER I

INTRODUCTION

A. Background of the Research

1. Resource Allocation for Statistical QoS Guarantees

a. Motivations

Quality-of-service (QoS) guarantees play a critically important role in future mobile wireless networks. Depending on their distinct QoS requirements, differentiated mobile users are expected to tolerate different levels of delay for their service satisfactions. For instance, non-real-time services such as data disseminations aim at maximizing the throughput with a loose delay constraint. In contrast, for real-time services like multimedia video conference, the key QoS metric is to ensure a stringent delay-bound, rather than to achieve high spectral efficiency. There also exist some services falling in between, e.g., paging and interactive web surfing, which are delay-sensitive but whose delay QoS requirements are not as stringent as those of real-time applications. The diverse mobile users impose totally different and sometimes even conflicting delay QoS constraints, which impose great challenges to the design of future mobile wireless networks.

Unlike its wired counterparts, supporting diverse delay QoS in wireless environment is much more challenging since the wireless channel has a significant impact on network performance. In particular, a deterministic delay-bound QoS guarantee over wireless networks is practically infeasible due to the time-varying nature of fading channels. Alternatively, a more practical solution is to provide the *statistical*

The journal model is *IEEE Transactions on Automatic Control.*

QoS guarantees [1], where we guarantee the given delay-bound with a small violation probability.

Over the wireless environment, the most scarce radio resources are power and spectral bandwidth. In response, a great deal of research has been devoted to the resource allocation problems for wireless communications and networks [2–5]. The framework used to evaluate these techniques is mainly based on information theory [6, 7], using the concept of either *ergodic capacity* [8, 9] or *outage capacity* [10, 11]. The ergodic capacity maximizes the average spectral efficiency with an infinite long delay. The outage capacity, on the other hand, maintains a constant rate transmission with a certain outage probability. From the point-of-view of delay QoS, such an information-theoretic framework maximizes the system throughput either without any delay constraint (i.e., ergodic capacity), or with a stringent delay constraint (i.e., outage capacity). These two extremes may not refine enough for the user's satisfactions, where a wide range of delay constraints may be requested for different applications. Consequently, to provide diverse QoS guarantees, it is necessary to take the QoS metrics into account when applying the prevalent information theory.

Thanks to the dual concepts of *effective bandwidth* and *effective capacity*, we obtain a powerful approach to evaluate the statistical QoS performance from the networking perspective. The effective-bandwidth theory has been extensively studied in the early 90's with the emphasis on wired asynchronous transfer mode (ATM) networks [1,12–19]. This theory enables us to analyze network statistics such as queue distributions, buffer overflow probabilities, and delay-bound violation probabilities, which are important for statistical QoS guarantees. In [20–23], Wu and Negi proposed an interesting concept termed effective capacity, which turns out to be the dual of the effective bandwidth. The effective-capacity approach is particularly convenient for analyzing the statistical QoS performance of wireless transmissions where the service process is driven by the time-varying wireless channel. In this dissertation, we integrate information theory with effective capacity and propose QoS-driven resource allocation. The scheme aims at optimally allocating resources such as power and rate subject to a given queuing delay constraint, and providing diverse QoS guarantees for the wireless communications and networks.

b. Related Works

As mentioned above, one of the foundations for our research is the theory of statistical QoS guarantees, originating from the concept of *effective bandwidth* [1,12–15,18,19,24]. However, there exist very few works that integrated statistical QoS guarantees with physical layer implementations of wireless communications. To the best of our knowledge, the first attempt to relate statistical QoS with physical layer infrastructure was by [16,17], where the focus of the research is two-state Markov arrival (ON-OFF traffic) and two-state Markov service process (ON-OFF channel), which is significantly different from our problem setting. The effective capacity proposed by Wu and Negi [20–23] becomes the starting point of our research. However, all the above mentioned works [16,17,20–23] did not focus on resource allocation.

At the physical layer, on the other hand, there do exist a large number of papers discussing delay-constrained resource allocation problems [3,5,25–33].

1. In papers [3,5,25,26] and references herein, the focus is *outage capacity* (also termed *delay-limited capacity*) analyses, which is the most popular framework for delay-constrained analysis at the physical layer. As will be discussed in detail in the dissertation, these works can be considered as a special case of our proposed QoS-driven resource allocation policies as the QoS constraint becomes stringent. However, it is important to note that the information theoretic results

are initial works of our research.

2. When the queuing effect comes into play, most of the previous works at the physical layer focus on *average queuing delay* (i.e. the marginal delay statistics) [27–33], using *Little's theorem*, which is significantly different from our work, where the focus is delay-bound violation probabilities. In particular, *none of the above work is based on (1) large deviation principle (LDP); (2) fluid queuing model; (3) parallel channel model which is general and applicable to multiple-input-multiple-output (MIMO) and orthogonal-frequency-division-multiplexing (OFDM) systems.* In [30,31], in addition to the average queuing delay analysis, the authors also discussed power allocation problem to guarantee absolute delay and statistical delay bound. However, in these cases, the authors in [30,31] assume an additive white Gaussian noise (AWGN) channel, instead of fading channel. The approach for solving the problem is either based on dynamic programming or iterative algorithms. The complexity of the resource allocation grows exponentially fast as number of channel states increases. Furthermore, no *general* closed-form power allocation policies are obtained. However, it is also worth noting that based on average queuing delay analysis, the authors in [30] obtained similar observation with ours, but the research approach and focuses are very different.

In summary, the research conducted in this dissertation is significantly different from the existing works. The key idea of this research is to “borrow” the concept of statistical QoS (through effective bandwidth and effective capacity) at network layer into physical layer problem, which has not been previously done by the researchers from both upper-layers and physical layer research communities.

2. Adaptive Antenna Selection

a. Motivations

In general, the MIMO techniques can be classified into two categories:

1. Improving the reliability by spatial diversity.
2. Enhancing the throughput by spatial multiplexing.

In the second part of this dissertation, we focus on the first category, i.e., spatial diversity based MIMO system, and develop schemes to allocate the physical layer resources such as power, subcarriers, and antennas, to improve the symbol-error rate (SER) performance of wireless transmissions. Note that bit-error rate (BER) or SER is the key QoS metric evaluated at the physical layer.

The spatial diversity based MIMO technique, making use of multiple antennas at transmitter and/or receiver, is an effective approach to combat the time-varying fading channels. There exists a large number of promising transmit/receive diversity schemes. For example, when the channel-state information (CSI) is available at both sides of the wireless link, maximal-ratio transmission (MRT, also known as transmit beamforming) and maximal-ratio combining (MRC) are known as the optimal transmit- and receive-diversity schemes [34], respectively. When the CSI is not available at the transmitter side, space-time block code (STBC) is a powerful approach to achieve transmit diversity [35, 36]. The antenna selection diversity (SD) at either the transmitter or receiver side, on the other hand, emerges as a good tradeoff between the performance and complexity [37], and thus received a great deal of research attentions [38–49].

b. Related Works

The SD technique can be considered as a special case of the more general hybrid-selection (H-S) schemes [43–49], where a subset of the available antennas are selected at transmitter and/or receiver. The H-S technique over independent identical fading environment has been studied in [38–47], followed by some more complicated models such as the unequal fading and power models [48] and the correlated fading models [49]. In [44], the authors developed the novel approach termed as *virtual branch* to analyze the performance of H-S scheme, which transforms the mutually *dependent* order statistics to *independent* virtual branches and thus significantly simplify the analyses. In [47], the authors further studied the impact of CSI estimation error on H-S scheme.

Most previous works in this area mainly focused on H-S employed at the *receiver* side only [39, 43, 44, 47–49] such that the CSI *feedback* from the receiver to the transmitter *is not needed* to make the antenna selection decision. In contrast, when the selection is applied at the transmitter side, the CSI feedback is necessary and thus the imperfectness of CSI feedback will impair the performance of the H-S scheme. While the powerful *virtual branch* technique [44] particularly focuses on the *order statistics*, this technique is hard to be extended to our imperfect CSI feedback analyses, specifically, for the delayed-CSI feedback analysis, because our delayed-feedback analysis involves the *induced order statistics* [50]. In [38, 42], the authors investigated the impact of feedback delay on SD scheme. Also, the authors in [46] studied the impact of feedback error on STBC-based antenna-selection scheme proposed in [45]. However, no closed-form SER expressions were obtained in [38, 42, 46]. In this dissertation, out main target is to derive closed-form expressions of SER for a set of H-S schemes when considering imperfect CSI feedback.

B. Contributions of the Dissertation

In the first part of this dissertation:

1. We propose a media access control (MAC)- physical (PHY) cross-layer model for wireless QoS guarantees. This model is shown to be simple and accurate that can efficiently characterize the interaction between PHY layer infrastructure and MAC layer queuing delay-QoS requirements.

2. We propose QoS-driven resource allocation schemes for single-input-single-output (SISO), MIMO, OFDM, and cooperative relay communication systems and networks. The scheme maximizes the system throughput subject to a given queuing delay constraint. Our analyses reveal an important fact that there exists a fundamental tradeoff between throughput and QoS provisioning. In particular, when the delay constraint becomes loose, the optimal resource allocation policy converges to the well-known *water-filling* scheme, where ergodic capacity can be achieved. On the other hand, when the QoS constraint gets stringent, the optimal policy converges to the *channel inversion* scheme under which the system operates at a constant rate and the zero-outage capacity of the system can be achieved. Thus, from a queuing delay point-of-view,[1] both ergodic capacity (i.e., infinite delay) and delay-limited capacity (i.e., zero delay) can be considered as special cases of QoS-driven resource allocation. Moreover, we also consider the impact of channel estimation error on the performance of QoS provisioning, and propose the corresponding optimal resource allocation policies.

In the second part of this dissertation:

[1] The information theory mainly considers coding delay, instead of queuing delay.

1. We investigate selection diversity for MIMO multicarrier (MC) direct-sequence (DS) code-division-multiple-access (CDMA) systems and analyze the error performance of the system when considering feedback delay and feedback errors. We also propose a joint antenna selection and power allocation scheme for Alamouti STBC systems and analyze the error performance of the scheme when taking the feedback delay into account.

C. Outline of the Dissertation

The reminder of this dissertation is organized as follows.

In Chapter II, we propose a physical (PHY)-media access control (MAC) cross-layer system model for QoS guarantees. We show how to utilize the concept of effective capacity as a bridge for cross-layer design and optimization. Moreover, the fundamental properties of the effective capacity are introduced, which also serve as a basis for our resource allocation derived in the following chapters. Chapter II is in part a reprint of the material in the papers [51, 52].

In Chapter III, we investigate the QoS-driven resource allocation for single channel communication systems. By integrating information theory with effective capacity, we derive the optimal resource allocation policy subject to a given queuing delay constraint. Our results reveal an important tradeoff between throughput and QoS. We also study the resource allocation policy for the more practical MQAM systems over finite-state Markov chain (FSMC)-modeled channel. The results provide a guidance on how time correlation of the channel can impact the QoS provisioning. Chapter III is in part a reprint of the material in the papers [53, 54].

As a natural extension of Chapter III, Chapter IV discusses QoS-driven resource allocation of multichannel communication systems for QoS guarantees. The multichannel models either diversity-based systems or multiplexing-based systems, which play a fundamental role in current and future wireless communications. Consistent with our previous work, the results demonstrate that, when the delay constraint is loose, the optimal resource allocation approaches water-filling. On the other hand, when the delay constraint is stringent, the optimal resource allocation approaches "channel inversion".[2] Furthermore, our simulation results show that the MIMO infrastructure significantly outperforms SISO system in terms of QoS provisioning. Chapter IV is in part a reprint of the material in the paper [55–57].

In Chapters III and IV, we assume that the CSI feedback is perfect. In Chapter V, we further consider the impact of channel estimation error on QoS guarantees. We derive the power allocation algorithms subject to either *total* power constraint or *average* power constraint, or both constraints. Under stringent delay requirement, we provide necessary and sufficient conditions for the convergence of the average power, which shows that in the presence of channel estimation errors, the average power always diverges. Furthermore, a positive zero-outage capacity is proved to be unattainable. Alternatively, we explicitly obtain the power allocation scheme to minimize the outage probability. Chapter V is in part a reprint of the material in the paper [58, 59].

In Chapter VI, we apply the proposed resource allocation scheme for cellular wireless networks. Our proposed scheme dynamically assigns power-levels and time-slots for heterogeneous real-time mobile users to satisfy the variation of statistical delay-bound QoS requirements. We derive the admission-control and power/time-slot allocation algorithms, guaranteeing the statistical delay-bound for heterogeneous mobile users. When designing such an algorithm, we study the impact of physical-layer

[2] The channel inversion is in a wide sense, see Chapter IV for details.

issues such as adaptive power-control and CSI feedback delay on the QoS provisioning performance. Through numerical and simulation results, we observe that the adaptive power adaptation has a significant impact on statistical QoS-guarantees. In addition, the analyses indicate that our proposed resource-allocation algorithms are shown to be able to efficiently support the diverse QoS requirements for various real-time mobile users over different wireless channels. Also, in an in-door mobile environment, our proposed algorithm is shown to be robust to the CSI feedback delay. Chapter VI is in part a reprint of the material in the paper [60].

In Chapter VII, we apply our effective capacity based approach to cooperative relay networks. We focus on simple half-duplex relay protocols, namely, amplify-and-forward (AF) and decode-and-forward (DF), and develop the associated dynamic resource-allocation algorithms. The resulting resource allocation policy in turn provides a guideline on how to design the relay protocol that can efficiently support stringent QoS constraints. For DF relay networks, we also study a fixed power allocation scheme and investigate its performance. The simulations and numerical results verify that our proposed resource allocation can efficiently support diverse QoS requirements over wireless relay networks. Moreover, both AF and DF relays show significant superiorities over direct transmissions when the delay constraints are stringent. On the other hand, our results demonstrate the importance of deploying the dynamic resource allocation for stringent delay QoS guarantees. Chapter VII is in part a reprint of the material in the paper [61].

In Chapter VIII, we turn our attention to the adaptive antenna selection. We propose the scheme to integrate SD with MC DS-CDMA wireless networks. Applying the proposed SD scheme, the transmitter jointly selects the *optimal subcarrier-and-antenna pair* to decrease the peak-to-average power ratio (PAPR), which is one of the main problems inherently associated with MC DS-CDMA communications. Over the frequency-selective Nakagami-m fading channels, we develop the unified framework to analyze the SER of the scheme implemented in different types of wireless networks, while dealing with the *perfect* and *imperfect* CSI feedbacks, respectively. Our analyses show that in a wide variation of the feedback imperfectness, our proposed SD scheme has significant advantages over the conventional schemes for both downlink cellular networks and ad hoc wireless networks. However, our analytical findings indicate that SD scheme cannot always outperform the conventional schemes even when the perfect CSI feedbacks are available. Chapter VIII is in part a reprint of the material in the paper [37, 62].

In Chapter IX, we investigate the Alamouti scheme with joint antenna selection and power allocation over flat-fading Rayleigh channels. Based on the CSI feedbacks, the transmitter selects the optimal two antennas out of all possible antennas to transmit data using STBC. Then, the transmitter adaptively allocates transmit power among the selected antennas to minimize the SER. We derive the SER as either the closed-form expression or the single-fold finite integral when assuming perfect and delayed CSI feedbacks, respectively. Our results show that when the CSI feedback is perfect, the optimal power allocation is to assign all power to the single optimal antenna, such that the SD-STBC reduces to the simpler SD scheme. On the other hand, when taking the CSI feedback delay into account, the SD-STBC scheme with dynamic power allocation ensures the better SER performance than the conventional SD scheme and SD-STBC scheme with equal power allocation. Chapter IX is in part a reprint of the material in the paper [63].

In Chapter X, we summarize the dissertation and point out future research directions.

CHAPTER II

CROSS-LAYER MODELING BY EFFECTIVE CAPACITY

A. Introduction

The explosive development of wireless services such as wireless Internet, mobile computing, and cellular telephoning motivates an unprecedented revolution in wireless networks. This also presents great challenges to system designers since the time-varying fading channel has the significant impact on supporting the quality-of-service (QoS) requirements for heterogeneous mobile users. A large number of effective schemes are developed at physical layer to overcome the impact of wireless fading channels [2, 3]. Among them, the multiple-input-multiple-output (MIMO) infrastructure [8, 35, 36] and adaptive-modulation-and-coding (AMC) scheme [27, 64] are promising techniques that have received significant research attention in recent years.

There have been a great deal of research efforts on applying both MIMO and AMC to improve the spectral-efficiency at physical layer. However, the problems on how to efficiently employ the unique nature of such techniques for enhancing upper-layer protocol design, and what is the impact of these physical-layer's new techniques on supporting the diverse upper-layer QoS requirements, have been neither well understood, nor thoroughly studied. Consequently, it becomes increasingly important to develop the *cross-layer* system model to integrate the QoS provisioning algorithms/protocols at higher network-protocol layers with MIMO and AMC implemented at physical layer. In this chapter, our focus is on designing the cross-

©2006 IEEE. Part of the material presented in this chapter is reprinted with permission from "Cross-layer-based modeling for quality of service guarantees in mobile wireless networks" by X. Zhang, J. Tang, H.-H. Chen, S. Ci, and M. Guizani, published in *IEEE Communications Magazine*, vol. 44, no. 1, pp. 100–106, Jan. 2006.

layer model that can characterize the interactions across physical-layer and data-link-layer, and mapping the physical-layer parameters to data-link-layer's real-time multimedia *delay-bound* QoS requirements. Based on our proposed cross-layer model, the advanced mechanisms such as adaptive resource-allocation, admission control, and packet scheduling schemes can be developed to guarantee the diverse QoS requirements for, e.g., the third-generation (3G) and beyond 3G cellular networks and wireless local-area networks (WLAN).

There have been a variety of research works focusing on wireless system modeling in both physical-layer and data-link-layer. In [65], Wang and Moayeri proposed using the finite-state Markov chains (FSMC) to model the wireless fading channel. This model has then been extensively studied for both Rayleigh and Nakagami-m fading channel (e.g., see [66–68] and references therein), and successfully applied in various scenarios to evaluate the QoS performance over wireless links [16, 17, 28, 69]. In [16, 17], Krunz and Kim modeled both arrival traffic (i.e., ON-OFF model) and channel service (i.e., Gilbert-Elliot model) as two-state Markov processes, and analyzed the delay-bound violation probability of point-to-point wireless transmissions. The model in [16, 17] did an excellent work in cross-layer-based wireless QoS guarantees. However, this model is not comprehensive enough to characterize the more realistic scenarios where both arrival and service processes are much more complicated. In [28, 69], Liu, Zhou, and Giannakis integrated the FSMC with AMC, and then jointly considered the physical-layer channel and data-link-layer queueing characteristics. Their model in [28, 69] was developed across physical-layer and data-link-layer, characterizing the impact of physical-layer variations on data-link-layer's QoS performance over wireless networks. However, the main QoS requirement addressed in [28, 69] is the *average delay* of the wireless transmission, which does not effectively support the real-time wireless multimedia services, where the key QoS metric is the *bounded delay*, as addressed in this dissertation.

In [21], Wu and Negi proposed a very interesting concept termed as "effective capacity". This concept turns out to be the *dual problem* of the so-called "effective bandwidth", which has been extensively studied in the early 90's with the emphasis on wired asynchronous transfer mode (ATM) networks [1,12–19]. The effective capacity and effective bandwidth enable us to analyze the statistical delay-bound violation probability, which is critically important for real-time multimedia wireless networks. Yet, there also exists limitations in [21]. First, the model assumes a constant arrival traffic, which is not too realistic for most practical wireless networks. Second, the authors employed an experimental-based method to measure the effective capacity (the procedure was called "link-layer channel estimation" in [21]), which is practically viable but lack of general analytical merits. Moreover, the estimation-based method in [21] further requires that the arrival traffic has the constant rate, which itself limits the applications of the effective capacity to the scenarios where the arrival processes are time-varying.

To overcome the aforementioned problems, in this chapter we propose a cross-layer approach to investigate the impact of physical-layer infrastructure on data-link-layer QoS performance in mobile wireless networks. At physical layer, in contrast to either the work in [16,17] that uses Gilbert-Elliot channel model or the study in [28,69] which considers the single-input-single-output (SISO) antenna system, we integrate the MIMO transmit/receive-diversity with AMC. The MIMO-diversity schemes include maximal-ratio transmission (MRT, also known as beamforming), space-time block coding (STBC), maximal-ratio combining (MRC), transmit/receive selection combining (SC) [37] (also see in Chapter VIII for details), as well as the integration of those schemes. We show in this chapter that different antenna-diversity and AMC-based service processes can be combined into a *unified* model with a set of parameters varying.

At data-link layer, our focus is on how the physical-layer infrastructure influences the real-time multimedia delay-bound QoS provisioning performance. Based on our FSMC model developed at physical-layer, we characterize the QoS provisioning performance at data-link-layer by applying and extending the effective capacity method [21]. We show how the effective capacity can function as a bridge that connects the physical-layer across with the data-link-layer. As compared to the *experimental* estimation-based approach developed in [21], we *analytically* derive the effective capacity expressions and characterize the fundamental properties of the effective capacity over wireless links. Our studies suggest that the QoS provisioning performance can be determined for *any stationary* arrival process, removing the *constant-arrival* assumption used in [21]. The numerical and simulation results obtained demonstrate that our approach can efficiently capture the interactions across different network-protocol-layers and accurately characterize the QoS provisioning performance.

The rest of this chapter is organized as follows. Section B describes the physical-layer system model. Section C investigates the effective capacity and its relationship with the cross-layer design. Section D develops the analytical framework to analyze the effective capacity of FSMC-based service process. Section E presents the numerical results on effective capacity and statistical QoS guarantees. Section F discusses the impact of power-control and feedback-delay on the effective capacity. The chapter concludes with Section G.

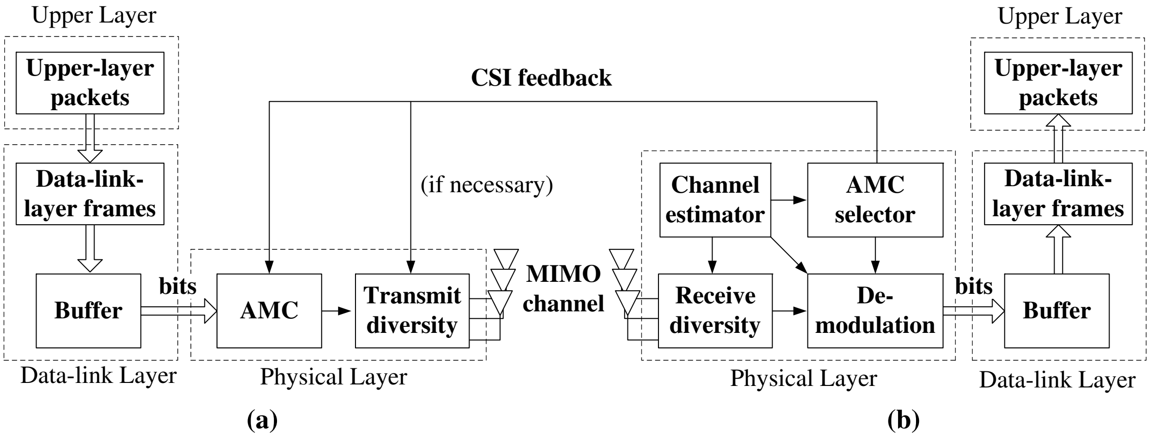


Fig. 1. The cross-layer system model. (a) Basestation transmitter. (b) Mobile wireless receiver.

B. The Physical-Layer System Model

The system model is shown in Fig. 1. In this chapter, we concentrate on a point-to-point wireless downlink with N_t antennas (Tx) at the basestation transmitter and N_r antennas (Rx) at the mobile receiver. As shown by Fig. 1, the upper-protocol-layer packets are first divided into a number of frames at data-link layer. The frames are stored at the transmitter-buffer and then split into bit-streams at physical layer, where the AMC and MIMO-diversity are employed, respectively. The reverse operations are executed at the receiver side. Also, the channel-state information (CSI) is estimated at the receiver and fed back to the transmitter for AMC and MIMO-diversity (if necessary, depending on the specific MIMO-diversity scheme used). The upper-protocol-layer packets have the same packet-size, which consists of N_p bits. Also, the frame at data-link layer has the same time-duration, which is denoted by T_f. Due to the employment of AMC, the number of bits per frame varies depending on the modulation-and-code modes selected. Therefore, each frame comprises a various portion of the packet. Furthermore, at the data-link layer, the system integrates forward-error control (FEC) with automatic retransmission request (ARQ) strategies, which will be detailed in Section 2.

We assume that the wireless channel is flat-fading with Nakagami-m distribution, which is independent identically distributed (i.i.d.) between each transmit/receive antenna-pair. Also, the channel is invariant within a frame's time-duration T_f, but varies from one frame to another. We use the Nakagami-m channel model because this model is very general and often best fits the land-mobile and indoor-mobile multipath propagations [34, 37]. As the fading parameter m varies, where $m \in [1/2, +\infty)$, the model spans a wide range of fading environments, including one-sided Gaussian fading channel ($m = 1/2$, the worst fading case), the Rayleigh fading channel ($m = 1$), the precise approximations of Rician and lognormal fading channels ($m > 1$), and the conventional Gaussian channel ($m = \infty$, no fading). We assume that the CSI is perfectly estimated at the receiver and reliably fed back to the transmitter. However, the CSI feedback can be delayed, which is particularly addressed in Section F. We also assume that the data-link-layer buffer-size is infinite.

1. MIMO Diversity

In general, the MIMO techniques can be classified into two categories:

1. Improving the reliability by spatial diversity.

2. Enhancing the throughput by spatial multiplexing.

Due to the hardware limitation and the power constraint of the mobile terminal (MT), the number of receive antennas at MT is limited to a small number. Therefore, the benefit of using spatial-multiplexing is restricted and at cost of complexity at MT. In contrast, it is practically more attractive to apply spatial-diversity, especially the basestation transmit-diversity technique, in mobile wireless networks. Thus, in this chapter we only focus on the first category, i.e., the spatial-diversity-based MIMO systems.

There exists a number of promising transmit/receive diversity schemes. For example, when the CSI is available at both sides of the wireless link, MRT and MRC are known as the optimal transmit- and receive-diversity schemes [34], respectively. When the CSI is not available at the transmitter side, STBC is a powerful approach to achieve transmit diversity [35, 36]. Moreover, the SC at either the transmitter or receiver side emerges as a good tradeoff between performance and complexity [37]. For a variety of MIMO-diversity schemes, we can show that the probability density function (pdf) of the combined signal-to-noise ratio (SNR), denoted by $p_\Gamma(\gamma)$, can be derived as the *unified* expression as follows:

$$\begin{aligned} p_\Gamma(\gamma) &= \frac{M}{\Gamma(mL)}\sum_{i=0}^{M-1}(-1)^i\binom{M-1}{i} \\ &\quad \cdot \exp\left(-(i+1)\frac{\beta m}{\overline{\gamma}}\gamma\right)\sum_{j=0}^{i(mL-1)}\xi_{ji}\left(\frac{\beta m}{\overline{\gamma}}\right)^{j+mL}\gamma^{j+mL-1} \qquad (2.1) \end{aligned}$$

where $\Gamma(\cdot)$ represents the Gamma function, $\overline{\gamma}$ denotes the average SNR of the combined signal, m denotes the fading parameter, ξ_{ji} is the multinomial expansion coefficients determined by $\xi_{ji} = \sum_{p=a}^{b}\xi_{p\,(i-1)}/[(j-p)!]$ with $a = \max\{0, j-(M-1)\}$, $b = \min\{j, (i-1)(M-1)\}$, $\xi_{j0} = \xi_{0i} = 1$, $\xi_{j1} = 1/(j!)$, and $\xi_{1i} = i$, and finally, the parameters M, L, and β are MIMO-diversity-scheme dependent, which are specified in Table I.[1] Note that in Eq. (2.1), M and L denote the *selection diversity* order and *combining diversity* order, respectively, and β only affects STBC scheme that reduces the variation of the channel. The total diversity order is determined by $M \times L$ (also

[1] Note that when the number of receive antennas $N_r > 1$, the pdf of the SNR using MRT/MRC can be derived as in [34, Section 9.15]. However, this pdf cannot be included in the framework shown in Eq. (2.1). Therefore, we only provide a performance upper-bound of that scheme in TABLE I.

Table I. Parameter Identifications for Unified MIMO Diversity.

MIMO Diversity Schemes	M	L	β
Tx-MRT/Rx-1	1	N_t	1
Tx-STBC/Rx-MRC	1	N_tN_r	N_t
Tx-SC/Rx-MRC	N_t	N_r	1
Tx-MRT/Rx-SC	N_r	N_t	1
Tx-STBC/Rx-SC	N_r	N_t	N_t
Tx-SC/Rx-SC	N_tN_r	1	1
Performance Upper-Bound	1	N_tN_r	1

equal to $N_t \times N_r$).

2. Adaptive Modulation and Coding

The AMC technique has emerged as one of the key solutions to increase the spectral-efficiency of wireless networks. In [64], the authors did pioneering work in this area. In [27], the authors studied adaptive modulation integrated with convolutional-code-based FEC strategy. Following the work of [27], the specific modulation-and-coding modes for the AMC scheme used in this chapter are illustrated in Fig. 2.

As shown in Fig. 2, we partition the entire SNR range by, e.g., $K = 7$ non-overlapping consecutive intervals, resulting in $K + 1$ boundary points denoted by $\{\Gamma_k\}_{k=1}^{K+1}$, where $\Gamma_1 < \Gamma_2 < \cdots < \Gamma_{K+1}$ with $\Gamma_1 = 0$ and $\Gamma_{K+1} = \infty$. The first mode (Mode 1) corresponds to the "outage" mode of the system, i.e., the transmitter does not transmit data in Mode 1. For the other 6 modes, with the code rates ranging from 1/2 to 3/4 and the constellations from BPSK to 64-QAM, the spectral-efficiency of the system, denoted by R_k, varies from 0.5 to 4.5 bits/sec/Hz. As the SNR increases, the system selects the AMC mode with higher spectral-efficiency to transmit data. As

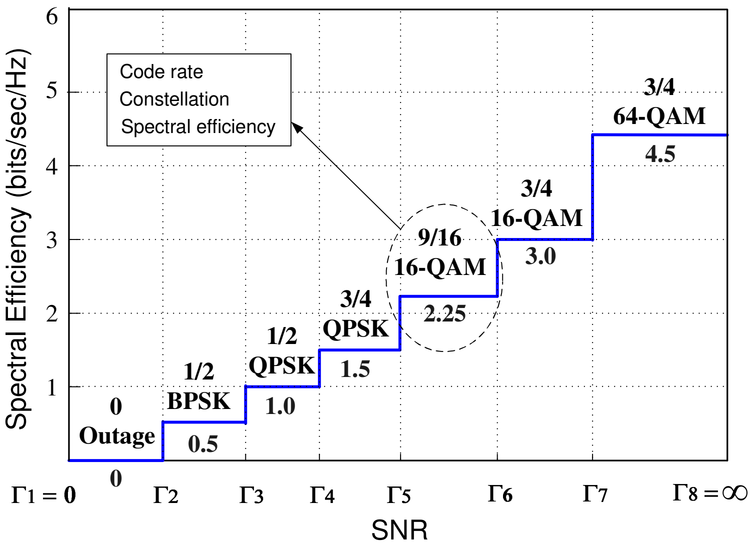


Fig. 2. AMC parameters according to ETSI HIPERLAN/2 standard.

the SNR gets worse, the system decreases the transmission-rate to adapt the degraded channel condition. In the worst case, the transmitter will stop transmitting data as in the "outage" mode of the system.

The packet-error rate (PER) when using the kth AMC mode ($k = 2, 3, ..., K$), denoted by $\mathrm{PER}_k(\gamma)$, can be approximated as follows [27, eq. (3)]:

$$\mathrm{PER}_k(\gamma) = \begin{cases} 1, & \text{if } 0 < \gamma < \gamma_k \\ a_k \exp(-g_k \gamma), & \text{if } \gamma \geq \gamma_k \end{cases} \tag{2.2}$$

where a_k, g_k, and γ_k are mode-dependent parameters, see [27, TABLE II] for details. Correspondingly, the AMC is in mode k if the SNR γ falls into the range of $\Gamma_k \leq \gamma < \Gamma_{k+1}$, where $k = 1, 2, ..., K$. Based on the pdf of the SNR in Eq. (2.1), the probability π_k, that the SNR falls into mode k is determined by

$$\pi_k = \int_{\Gamma_k}^{\Gamma_{k+1}} p_\Gamma(\gamma) d\gamma = \left[\frac{\gamma\left(mL, \frac{\beta m}{\bar{\gamma}}\Gamma_{k+1}\right)}{\Gamma(mL)} \right]^M - \left[\frac{\gamma\left(mL, \frac{\beta m}{\bar{\gamma}}\Gamma_k\right)}{\Gamma(mL)} \right]^M \tag{2.3}$$

where $k = 1, 2, ..., K$ and $\gamma(\cdot, \cdot)$ denotes the incomplete Gamma function. We select the boundaries such that $\Gamma_k \geq \gamma_k$ for all $k = 2, 3, ..., K$. Then, using Eq. (2.2), we obtain the average PER of mode k, denoted by $\overline{\mathrm{PER}}_k$, as follows:

$$\begin{aligned}
\overline{\mathrm{PER}}_k &= \frac{1}{\pi_k}\int_{\Gamma_k}^{\Gamma_{k+1}} a_k \exp(-g_k\gamma) p_\Gamma(\gamma) d\gamma \\
&= \frac{a_k M}{\pi_k \Gamma(mL)} \sum_{i=0}^{M-1} (-1)^i \binom{M-1}{i} \sum_{j=0}^{i(mL-1)} \xi_{ji} \left(\frac{\beta m}{b_k}\right)^{j+mL} \\
&\qquad \times \left[\gamma\left(j+mL, \frac{b_k \Gamma_{k+1}}{\overline{\gamma}}\right) - \gamma\left(j+mL, \frac{b_k \Gamma_k}{\overline{\gamma}}\right)\right] \qquad (2.4)
\end{aligned}$$

where $b_k = g_k\overline{\gamma} + (i+1)\beta m$. Thus, the average PER can be expressed as follows:

$$\mathrm{PER} = \frac{\sum_{k=2}^{K} R_k \pi_k \overline{\mathrm{PER}}_k}{\sum_{k=2}^{K} R_k \pi_k} \qquad (2.5)$$

where R_k denotes the spectral-efficiency of the kth mode. We can numerically obtain the boundaries $\{\Gamma_k\}_{k=2}^{K}$ such that the average PER satisfies the reliability QoS requirement. When taking the ARQ into account, the achieved spectral-efficiency of the kth mode, denoted by $\widetilde{R}_k$, can be expressed as

$$\widetilde{R}_k = R_k \left(1 - \overline{\mathrm{PER}}_k\right). \qquad (2.6)$$

3. Service Process Modeling Using FSMC

In this chapter, we employ the FSMC model to characterize the variations of the MIMO-diversity and AMC-based wireless-channel service process. The state of FSMC corresponds to the mode of AMC, where the effective transmission rate of the kth mode is $\widetilde{R}_k$. Let $p_{i,j}$ denote the transition probability of the FSMC from state i to state j. We assume a slow-fading channel model such that transition only happens between adjacent states [65–68]. Under such an assumption, we have $p_{ij} = 0$ for all

$|i-j|>1$. The adjacent transition probability can be approximated as [65]

$$\begin{cases} p_{k,k+1} \approx \frac{N_\Gamma(\Gamma_{k+1})T_f}{\pi_k}, \text{ where } k=1,2,...,K-1, \\ p_{k,k-1} \approx \frac{N_\Gamma(\Gamma_k)T_f}{\pi_k}, \quad \text{where } k=2,3,...,K \end{cases} \tag{2.7}$$

where $N_\Gamma(\gamma)$ is the level-crossing rate (LCR) calculated at SNR value of γ [68]. Then, the remaining transition probability can be derived as

$$\begin{cases} p_{1,1} = 1 - p_{1,2} \\ p_{K,K} = 1 - p_{K,K-1} \\ p_{k,k} = 1 - p_{k,k-1} - p_{k,k+1}, \ k=2,...,K-1. \end{cases} \tag{2.8}$$

Thus, applying Eqs. (2.7) and (2.8), we obtain the transition probability matrix of the FSMC, which is denoted by $\mathbf{P} = [p_{ij}]_{K\times K}$. Correspondingly, the stationary distribution of the FSMC, denoted by $\boldsymbol{\pi}$, is determined by $\boldsymbol{\pi} = [\pi_1, \pi_2, ..., \pi_K]$, where π_k is given by Eq. (2.3) for $k = 1, 2, ..., K$.

In order to obtain the transition probability matrix $\mathbf{P}$, it is necessary to find the LCR $N_\Gamma(\gamma)$ in Eq. (2.7). As derived in Appendix A, we obtain the *unified* closed-form expression of the LCR as follows:

$$\begin{aligned} N_\Gamma(\gamma) \;=\; & \frac{\sqrt{2\pi} f_d M}{\Gamma(mL)} \sum_{i=0}^{M-1} (-1)^i \binom{M-1}{i} \\ & \cdot \exp\left(-(i+1)\frac{\beta m \gamma}{\overline{\gamma}}\right) \sum_{j=0}^{i(mL-1)} \xi_{ji} \left(\frac{\beta m\gamma}{\overline{\gamma}}\right)^{j+mL-\frac{1}{2}} \end{aligned} \tag{2.9}$$

where f_d denotes the maximum Doppler frequency of the channel. Substituting Eq. (2.9) into Eq. (2.7), the transition matrix $\mathbf{P}$ is determined for different MIMO diversity schemes.

C. The Effective Capacity and Cross-Layer Designs

1. Statistical QoS Guarantees

During the early 90's, statistical QoS guarantees have been extensively studied in the contexts of effective bandwidth theory [1, 12–19]. The literature on effective bandwidth is abundant. The readers are referred to Chang [1] and Kelly *et. al.* [15] for a comprehensive review.

Based on large deviation principle (LDP), Chang in [1] showed that for a dynamic queueing system with stationary ergodic arrival and service processes, under sufficient conditions, the queue length process $Q(t)$ converges in distribution to a random variable $Q(\infty)$ such that

$$-\lim_{x\to\infty}\frac{\log\left(\Pr\{Q(\infty)>x\}\right)}{x}=\theta. \tag{2.10}$$

To be more specific, the above theorem states that the probability of the queue length exceeding a certain threshold x decays exponentially fast as the threshold x increases. Roughly speaking, Eq. (2.10) implies that

$$\Pr\{Q>x\}\approx e^{-\theta x},\ \text{for a large } x \tag{2.11}$$

where θ is a certain positive constant called "QoS exponent" [21]. For a small x, the following approximation is shown to be more accurate [21, 24]:

$$\Pr\{Q>x\}\approx \varepsilon e^{-\theta x} \tag{2.12}$$

where ε denotes the probability that the buffer is not empty, which can be approximated by the ratio of the average arrival-rate to the average service-rate [1, eq. (9.184)]. In other word, for a small x, the violation probability estimated by LDP is conservative, which serves as an upper-bound for the actual violation probability.

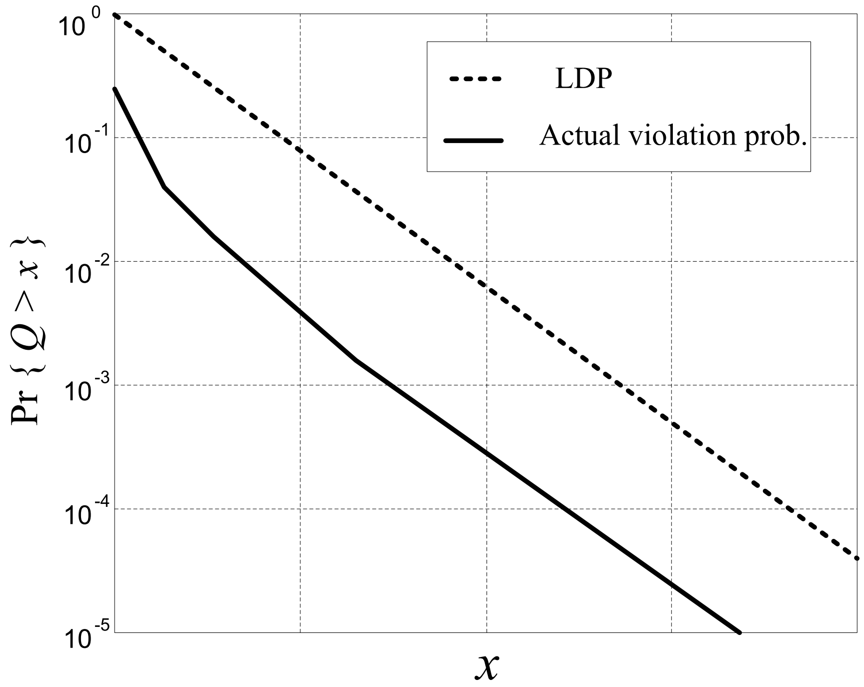


Fig. 3. A sketch of the violation probability comparisons between theoretic result (LDP) and actual result.

That is to say, when the delay constraint is small, the actual delay-bound violation probability should be smaller than the theoretic value derived from LDP. Fig. 3 shows a sketch of the violation probability comparisons between LDP-based theoretic result and the actual results. Therefore, although the LDP is an asymptotic result, it is still practically useful for real system design with finite (or small) x.

When delay-bound is the main QoS metric of interest (i.e., when the focus is on delay-bound violation probability), an expression similar to Eq. (2.12) can be obtained as

$$\Pr\{\text{Delay} > \tau_{\max}\} \approx \varepsilon e^{-\theta\delta\tau_{\max}} \tag{2.13}$$

where $\tau_{\max}$ denotes the delay-bound, and δ is jointly determined by both arrival and service processes, which will be detailed below.

Note that in the above, the parameter θ ($\theta > 0$) plays a critically important role for statistical QoS guarantees, which indicates the exponential decay rate of the QoS violation probabilities. A smaller θ corresponds to a slower decay rate, which implies that the system can only provide a *looser* QoS guarantee, while a larger θ leads to a faster decay rate, which means that a more *stringent* QoS requirement can be supported. In particular, when $\theta \to 0$, the system can tolerate an arbitrarily long delay. On the other hand, when $\theta \to \infty$, the system cannot tolerate any delay. Due to its close relationship with statistical QoS provisioning, θ is called the *QoS exponent* [20–23].

2. Effective Capacity and Cross-Layer Designs

Inspired by the effective bandwidth theory, Wu and Negi in [21] developed the concept of *effective capacity*, which is a *dual problem* of the original effective bandwidth. The effective capacity function, denoted by $\mathrm{E_C}(\theta)$, characterizes the attainable service-rate as a function of the QoS exponent θ. Specifically, in [21] the effective capacity $\mathrm{E_C}(\theta)$ is defined as the *constant* arrival-rate that the channel can support in order to guarantee a QoS requirement specified by θ. Analytically, the effective capacity can be formally defined as follows.

Let the sequence $\{R[i],\ i = 1, 2, ...\}$ denote a discrete-time stationary and ergodic stochastic service process and $S[t] \triangleq \sum_{i=1}^{t} R[i]$ be the partial sum of the service process. Assume that the Gärtner-Ellis limit of $S[t]$, expressed as $\Lambda_{\mathrm{C}}(\theta) = \lim_{t\to\infty}(1/t)\log\left(\mathbb{E}\left\{e^{\theta S[t]}\right\}\right)$ exists and is a convex function differentiable for all real θ [1, pp. 921]. Then, the effective capacity of the service process, denoted by $\mathrm{E_C}(\theta)$, where $\theta > 0$, is defined as [21, eq. (12)]

$$\mathrm{E_C}(\theta) \triangleq -\frac{\Lambda_{\mathrm{C}}(-\theta)}{\theta} = -\lim_{t\to\infty}\frac{1}{\theta t}\log\Big(\mathbb{E}\left[e^{-\theta S[t]}\right]\Big). \tag{2.14}$$

When the sequence $\{R[i],\ i=1,2,...\}$ is uncorrelated, it is clear that the effective capacity $\mathrm{E_C}(\theta)$ reduces to

$$\mathrm{E_C}(\theta) = -\frac{1}{\theta}\log\Big(\mathbb{E}\left[e^{-\theta R[i]}\right]\Big) = -\frac{1}{\theta}\log\Big(\mathbb{E}\left[e^{-\theta R[1]}\right]\Big). \tag{2.15}$$

The effective capacity expression Eq. (2.15) in uncorrelated case only depends on marginal statistics of a service process, which is much simpler than the general expression given by Eq. (2.14), where the higher order statistics of the service process are required. Since the block fading channel model generates an independent identically distributed (i.i.d.) service process, it can greatly simplify the effective capacity derivations.

Although the original concept of effective capacity is proposed based on *constant*-arrival assumption, it actually can be generalized to investigate the QoS performance of any *stationary* arrival process. Under such a condition, the arrival process should be represented by its effective bandwidth while the service process should be characterized by its effective capacity, respectively. Note that for a constant arrival-process, the corresponding effective-bandwidth is equal to its constant arrival-rate. Thus, the problem discussed in [21] can be considered as the special case of our more general scenario addressed in this chapter, where both arrival and service processes are time-varying. To help demonstrate the principles and identify the relationships between effective bandwidth and effective capacity, let us consider the case as illustrated in Fig. 4. Note that if the arrival-process has the constant rate, the corresponding effective-bandwidth is also a constant, with the value equal to its constant arrival-rate. Thus, the case in [21] can be considered as the special case of our more general scenario addressed in this chapter where both arrival and service processes are time-varying.

For any given arrival process and service process, we depict their effective-bandwidth function, denoted by $\mathrm{E_B}(\theta)$, and effective-capacity function, denoted by $\mathrm{E_C}(\theta)$, in Fig. 4, respectively. Let us define two limiting values as follows:

$$\begin{cases} \mu_A \triangleq \lim_{\theta\to 0} \mathrm{E_B}(\theta) \\ \mu_C \triangleq \lim_{\theta\to 0} \mathrm{E_C}(\theta). \end{cases} \tag{2.16}$$

The effective bandwidth theory demonstrates that μ_A is equal to the average arrival-rate of the traffic process [12, 19]. Also, we will show in the next section that μ_C is equal to the average service-rate of the service process. Therefore, using the approximation in [1, eq. (9.184)], the buffer non-empty probability ε in Eqs. (2.12) and (2.13) can be expressed as

$$\varepsilon \approx \frac{\mu_A}{\mu_C}. \tag{2.17}$$

The effective-bandwidth function $\mathrm{E_B}(\theta)$ intersects with the effective-capacity function $\mathrm{E_C}(\theta)$ at the point where the QoS-exponent is θ^* and the rate is δ.

In general, the delay-bound violation probability can be calculated in the following algorithm.

Algorithm 1. *Calculating the delay-bound violation probability by the following steps.*

S1: *According to the statistical characteristics of the arrival and service processes, find the effective-bandwidth function* $\mathrm{E_B}(\theta)$ *and effective-capacity function* $\mathrm{E_C}(\theta)$*. Determine the solution of the rate and QoS-exponent pair* (δ, θ^*) *such that* $\mathrm{E_B}(\theta^*) = \mathrm{E_C}(\theta^*) = \delta$*.*

S2: *Approximate the buffer non-empty probability* ε *by using Eq. (2.17).*

S3: *For any pre-determined delay-bound* $\tau_{\max}$ *and* (δ, θ^*) *obtained in* ***S2****, the delay-bound violation probability can be derived using Eqs. (2.13) and (2.17) as fol-*

lows:

$$\Pr\{\text{Delay} > \tau_{\max}\} \approx \varepsilon e^{-\theta^* \delta \tau_{\max}}. \tag{2.18}$$

From Fig. 4 we can gain insights about how the statistical QoS performance changes according to the service and arrival processes. As shown by Fig. 4, increasing the service-process bandwidth (as shown by the arrow at the lower position) results in higher effective capacity, which will lead to a larger QoS-exponent solution θ^*. This implies that the higher bandwidth service-process can support a more *stringent* QoS for a given arrival process. On the other hand, increasing the arrival-process bandwidth (as shown by the arrow at the upper position) makes the effective bandwidth increase, which generates a smaller QoS-exponent solution θ^* for a given service process. This implies that only a *looser* QoS can be guaranteed. When the bandwidth of the arrival process further increases such that $\mu_A > \mu_C$, there is no solution for $\theta^* > 0$ existing. Thus, the service process cannot support any QoS for the given arrival process, which is consistent with the queueing theory that if $\mu_A > \mu_C$, both queue length and the queueing delay will approach to infinity.

Inspired by the above analyses and observations, we propose to use the effective capacity as a bridge for the cross-layer modeling. The characterizations of the QoS performance guarantees are equivalent to investigating the dynamics of the effective capacity function, which turns out to be a simple and efficient cross-layer modeling approach. In [21], the authors employed an experimental-based method to measure the effective capacity. In fact, it is feasible to formulate the effective capacity problem in a more systematic manner. In the next section, we analytically investigate the effective capacity function for our FSMC-based wireless-channel service process.

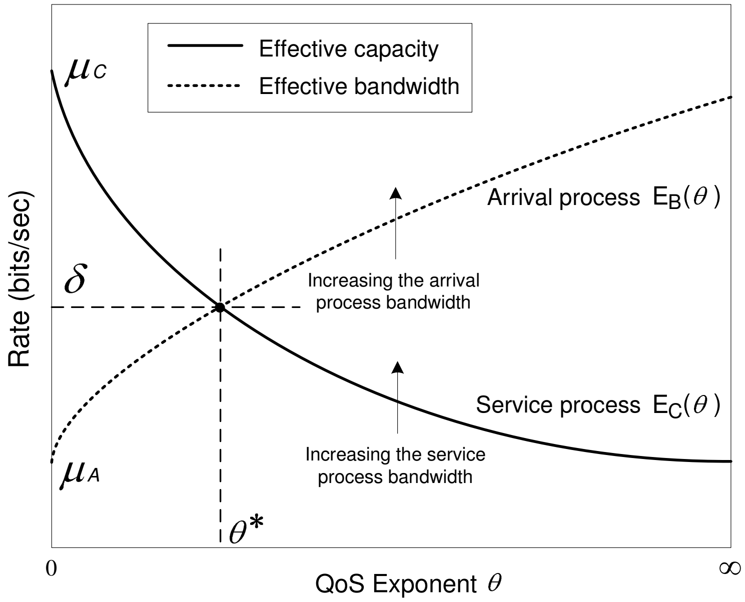


Fig. 4. The relationships between effective bandwidth and effective capacity as a function of the QoS exponent θ.

D. Effective Capacity for FSMC-Modeled Channel

1. Effective Capacity for FSMC-Modeled Channel

Based on our physical-layer FSMC model developed in Section 3, we have the following proposition.

Proposition 1. ***If*** *denoting the number of bits per frame transmitted at the state k of the FSMC-based service process by $\{\mu_k, k = 1, 2, ..., K\}$ and defining $\mathbf{\Phi}(\theta) \triangleq$ diag$\left\{e^{-\mu_1\theta}, e^{-\mu_2\theta}, ..., e^{-\mu_K\theta}\right\}$,* ***then*** *the effective capacity of the FSMC-based service process is determined by*

$$\mathrm{E_C}(\theta) = -\frac{1}{\theta}\log\Big(\rho\big\{\mathbf{P}\,\mathbf{\Phi}(\theta)\big\}\Big) \tag{2.19}$$

where $\mathbf{P}$ is the transition probability matrix determined by Eqs. (2.7) and (2.8), and $\rho\{\cdot\}$ denotes the spectral radius of the matrix.

Proof. The proof is provided in Appendix B.[2] □

Based on our system model in Section B, the transmission rate μ_k can be expressed as

$$\mu_k = \widetilde{R}_k T_f W, \quad \text{for all } k = 1, 2, ..., K \tag{2.20}$$

where W denotes the system spectral-bandwidth, and $\widetilde{R}_k$ is derived in Eq. (2.6) which takes the ARQ into consideration.

2. The Monotonic and Asymptotic Properties

We characterize the monotonic and asymptotic properties of the effective-capacity function by Proposition 2 that follows below.

Proposition 2. ***If*** *$\overline{\mu}$ and $\mu_{\min}$ denote the average and minimum number of bits per frame transmitted by the FSMC-based service process, respectively,* ***then*** *the following claims hold for the effective-capacity function $\mathrm{E_C}(\theta)$ of the FSMC-based service process:*

$$\underline{\textit{Claim 1.}} \quad \frac{d\mathrm{E_C}(\theta)}{d\theta} \le 0, \quad \textit{for all } \theta > 0. \tag{2.21}$$

$$\underline{\textit{Claim 2.}} \quad \sup_{\theta>0} \mathrm{E_C}(\theta) = \lim_{\theta\to 0} \mathrm{E_C}(\theta) = \overline{\mu}. \tag{2.22}$$

$$\underline{\textit{Claim 3.}} \quad \inf_{\theta>0} \mathrm{E_C}(\theta) = \lim_{\theta\to\infty} \mathrm{E_C}(\theta) = \mu_{\min}. \tag{2.23}$$

Proof. The proof is provided in Appendix C. □

[2] Note that the unit of the effective capacity in Eq. (2.19) is "bits per frame". To change the unit to "bits per second", the effective capacity in Eq. (2.19) should be normalized by the frame duration T_f.

Remark 1. Proposition 2 states that there is a tradeoff between the delay-QoS provisioning and the throughput. The effective capacity $\mathrm{E_C}(\theta)$ decreases from the average wireless-channel service-rate $\overline{\mu}$ to the minimum wireless-channel service-rate $\mu_{\min}$ as the delay-QoS requirement changes from the loose ($\theta \rightarrow 0$) to the stringent ($\theta \rightarrow \infty$) status, asymptotically.

3. The Scaling Property of the Effective Capacity

As mentioned before, our proposed cross-layer modeling can be used to design the adaptive resource-allocation algorithm for QoS guarantees. In such an algorithm, it is possible to assign various portions of the resources, e.g., bandwidth resource, time-slot resources, etc, to the mobile user in order to guarantee the user's QoS requirements. The following proposition provides a simple way to calculate the effective capacity for a number of service processes.

Proposition 3. ***If*** $\mathrm{E}_{\mathrm{C}_a}(\theta)$ *is the effective-capacity function of a service process* $R_a(t)$, ***then*** *the effective capacity of the service process* $R_b(t) = \chi R_a(t)$, *denoted by* $\mathrm{E}_{\mathrm{C}_b}(\theta)$, *is determined by*

$$\mathrm{E}_{\mathrm{C}_b}(\theta) = \chi \mathrm{E}_{\mathrm{C}_a}(\chi\theta) \tag{2.24}$$

where χ *is an arbitrary positive real-valued number.*

Proof. By definition, the effective capacity $\mathrm{E}_{\mathrm{C}_b}(\theta)$ can be expressed as

$$\begin{aligned} \mathrm{E}_{\mathrm{C}_b}(\theta) &= -\lim_{t\to\infty} \frac{1}{\theta t} \log\left(\mathbb{E}\left\{e^{-\theta \sum_{i=1}^{t} R_b(i)}\right\}\right) \\ &= -\lim_{t\to\infty} \frac{1}{\theta t} \log\left(\mathbb{E}\left\{e^{-\chi\theta \sum_{i=1}^{t} R_a(i)}\right\}\right) \\ &= -\chi \lim_{t\to\infty} \frac{1}{(\chi\theta) t} \log\left(\mathbb{E}\left\{e^{-(\chi\theta) \sum_{i=1}^{t} R_a(i)}\right\}\right) \\ &= \chi \mathrm{E}_{\mathrm{C}_a}(\chi\theta). \end{aligned} \tag{2.25}$$

The proof follows. □

Remark 2. Applying Proposition 3, the calculations of the effective capacity in some resource-allocation procedures can be significantly simplified. For example, in a dynamic time-division multiple access (TDMA) system, the mobile user can be adaptively assigned with different number of time-slots. Then, we only need to find the effective capacity of a single allocation scheme. The effective capacity of the other allocation schemes can be obtained directly by using Proposition 3.

E. Numerical Evaluations

We first evaluate the effective capacity by numerical solutions under different physical-layer diversity schemes and parameters, where we set the total system spectral-bandwidth $W = 100$ KHz, the upper-layer packet-size $N_p = 1080$ bits, and the data-link-layer frame time-duration $T_f = 2$ ms. Unless explicitly stated on the legend of the figures, the other system parameters are set as follows: the fading parameter $m = 1$, indicating the Rayleigh fading channel, the average SNR $\overline{\gamma} = 10$ dB, the Doppler frequency $f_d = 5$ Hz, and the average packet-error rate PER $= 10^{-3}$. To ease comparison with the spectral-efficiency, in the following discussions we plot the normalized effective capacity (*which is defined as the effective capacity divided by the spectral-bandwidth W and the frame duration T_f, and thus has the unit of "bits/sec/Hz"*).

Fig. 5 plots the effective capacity against the QoS exponent θ under different spatial diversity schemes. As shown in Fig. 5, the physical-layer antenna infrastructures have significant impact on the effective capacity. The effective capacities of MIMO (i.e., Tx-2/Rx-2) or multiple-input-single-output (MISO, i.e., Tx-2/Rx-1) systems are significantly larger than those of the SISO systems. Also, different diversity schemes

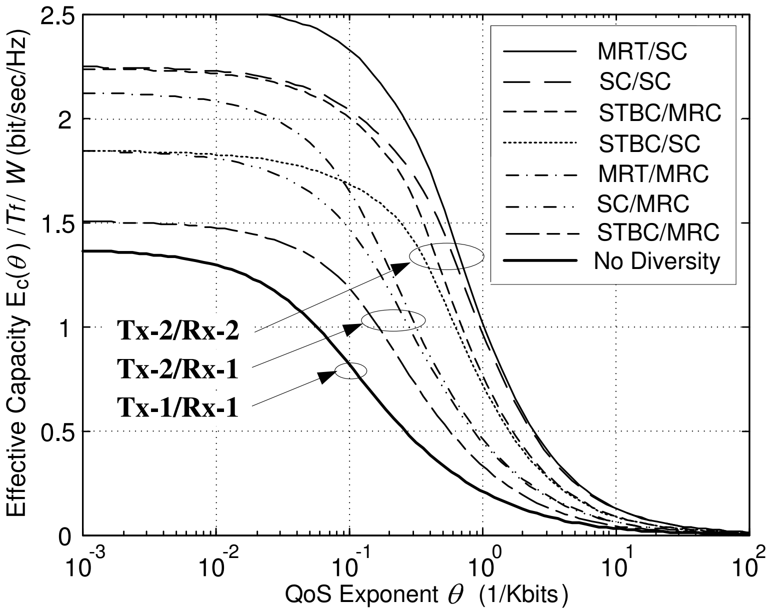


Fig. 5. The normalized effective capacity as a function of the QoS exponent θ under different spatial diversity schemes.

can achieve different effective capacities, depending on how much of the CSI information is utilized. An interesting observation is that when the QoS exponent θ is small, the effective capacity of some MIMO systems (e.g., STBC/SC) is lower than that of the MISO systems (e.g., MRT/MRC), which is because STBC/SC does not efficiently utilize the CSI while MRT/MRC fully utilizes the CSI. However, this situation changes as the QoS exponent θ increases. As shown in Fig. 5, for a large θ, the effective capacities of all MIMO systems are larger than those of the MISO systems. This implies that even under the condition that the MIMO system has lower spectral-efficiency than MISO system, it offers more significant advantages in supporting the stringent QoS requirement.

To further investigate the impact of antenna diversity on the effective capacity, Fig. 6 plots the effective-capacity gain (defined as the ratio of the effective-capacity with antenna-diversity-based systems to that with SISO-based systems) against the

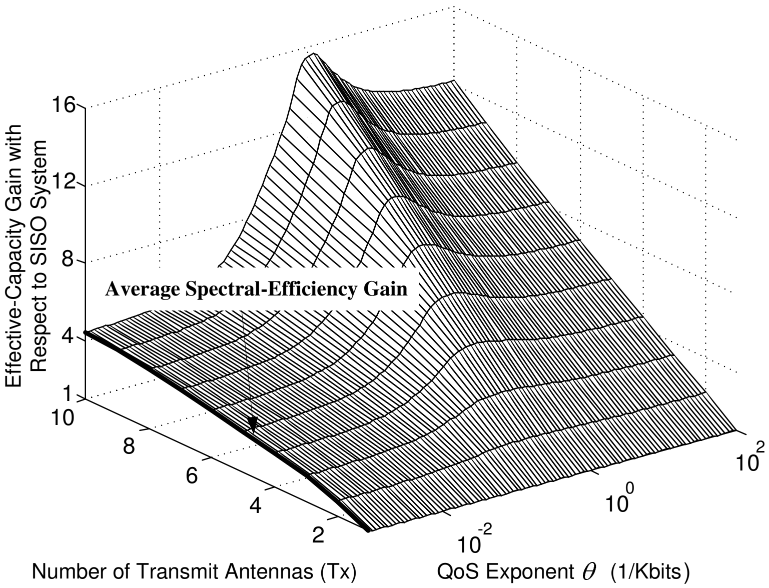
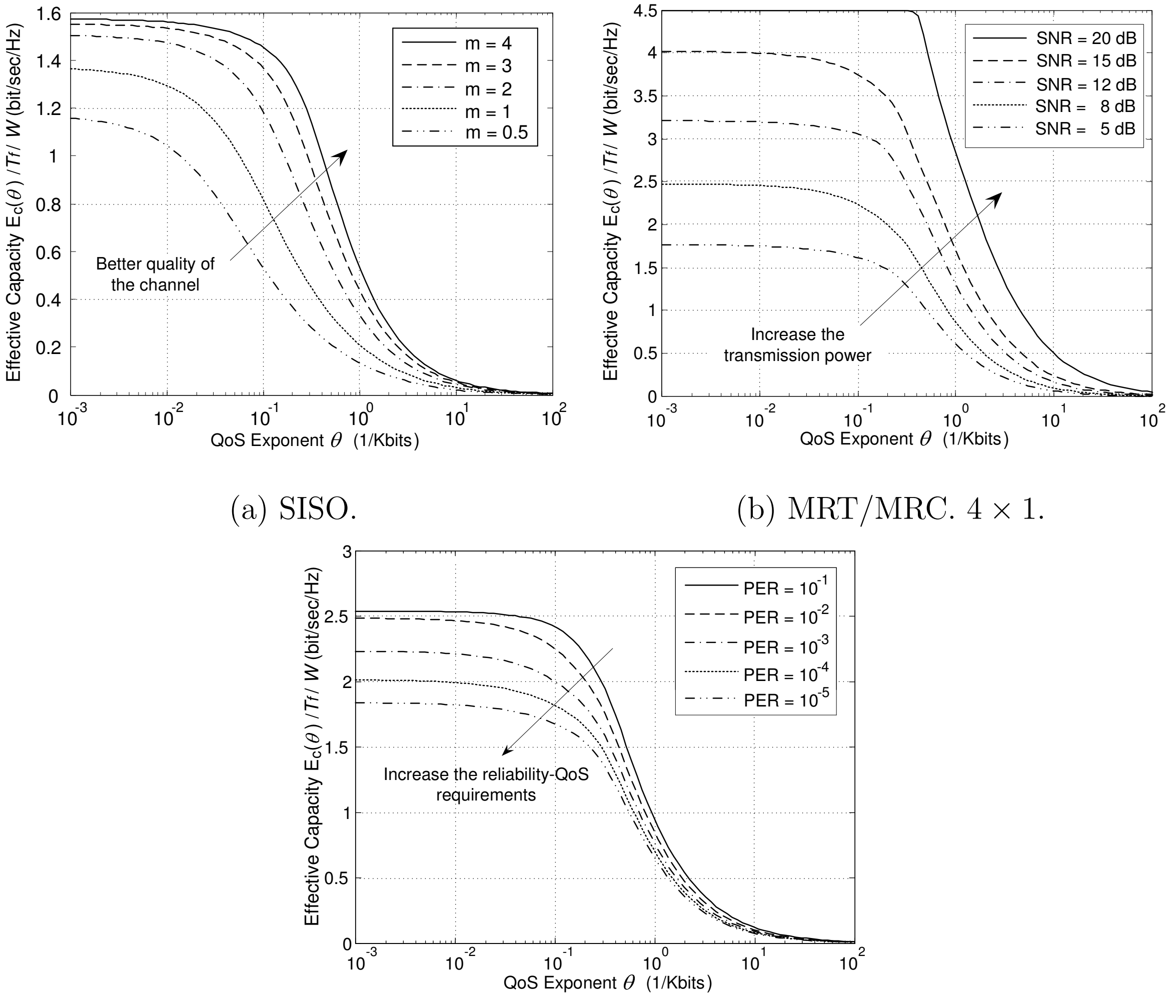


Fig. 6. The effective-capacity gain with respect to the SISO system. The MRT/MRC (Rx-1) is employed to fully utilize the CSI. The average SNR $\overline{\gamma} = 5$ dB.

number of transmit antenna N_t and the QoS exponent θ, where we employ MRT/MRC (Rx-1) to fully utilize the CSI. Notice from Proposition 2 that as the QoS-exponent θ approaches to 0, the effective-capacity converges to the *average service-rate*. Furthermore, the *average spectral-efficiency* is the average service-rate normalized by the spectral-bandwidth W. Therefore, the boldface line highlighted in Fig. 6 is actually the average spectral-efficiency gain achieved by antenna diversity. We observe from Fig. 6 that the effective-capacity gain with large θ is significantly higher than the average spectral-efficiency gain (indicated by the boldface line in Fig. 6 as $\theta \to 0$). Thus, Fig. 6 implies that *the superiority/gain of employing antenna diversity in terms of QoS-guarantees is even more significant than that in terms of the spectral-efficiency.*

Fig. 7(a) plots the effective capacity of SISO system against the QoS exponent θ with different channel distributions. From Fig. 7(a), we can observe that as the fading parameter m increases (the channel quality gets better), the effective capacity



(a) SISO.

(b) MRT/MRC. 4 × 1.

(c) STBC/MRC. 2 × 2.

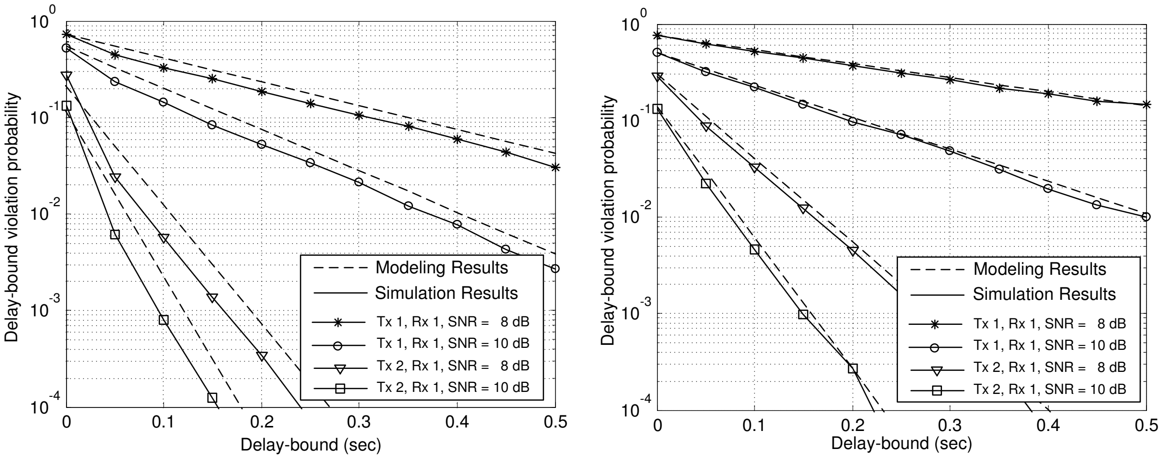
Fig. 7. The normalized effective capacity $\mathrm{E_c}(\theta)$ as a function of QoS exponent θ with different physical-layer diversity schemes and physical layer parameters.

increases correspondingly, which is expected since the stabler channel can support more stringent QoS. Fig. 7(b) plots the effective capacity against the QoS exponent θ when the SNR varies. As shown in Fig. 7(b), the better SNR of the wireless channel, or equivalently, increasing the transmission power, can improve the effective capacity. When the SNR $\overline{\gamma} = 20$ dB, we can see from Fig. 7(b) that the effective capacity gets saturated at spectral-efficiency of 4.5 bits/sec/Hz, which is because 4.5 bits/sec/Hz is the highest spectral-efficiency that can be obtained by the underlying AMC scheme (i.e., 64-QAM with code rate 3/4 as shown in Fig. 2). Fig. 7(c) depicts the effective capacity versus the QoS exponent θ with different reliability-QoS requirements of PER's, where the more stringent reliability-QoS results in the lower effective capacity. In summary, from Fig. 7 we can observe that the physical-layer variations have significant impact on the effective capacity, and thus on the QoS provisioning performance of wireless networks at higher-protocol-layers.

We also conduct simulations to verify the correctness and validity of our proposed cross-layer modeling technique and QoS provisioning performance. In the simulations, we generate two types of real-time services. The first type simulates the low speed audio service, where we model the arrival traffic by the well-known ON-OFF fluid model. The holding times in "ON" and "OFF" states are exponentially distributed with the mean equal to 8.9 ms and 8.4 ms, respectively. The "ON" state traffic is modeled as a constant rate of 16 Kbps. The system spectral-bandwidth for the audio service is set to $W = 10$ KHz. The second one simulates a high-speed video traffic flow. We employ a first-order auto-regressive (AR) process to simulate video traffic characteristics [70], the bit-rate of which can be expressed as

$$\nu(t) = a\nu(t-1) + bw \tag{2.26}$$

where $a = 0.8781$, $b = 0.1108$ [70] and w is a Gaussian random variable with the mean



(a) Audio traffic services.

(b) Video traffic services.

Fig. 8. The modeling and simulation results of the delay-bound violation probability for audio and video traffic services.

80 Kbps and standard deviation of 30 Kbps. The system spectral-bandwidth for the video service is set to $W = 100$ KHz. In the simulation, the transmitter employs STBC when the number of transmit antennas $N_t = 2$. The effective bandwidth of the audio traffic is derived according to [12], and the effective capacity of the video traffic is derived by using the approach proposed in [18], respectively.

Fig. 8(a) and 8(b) shows the QoS violation probability versus the delay-bound for audio and video traffic services, respectively. The delay-bound violation probability is derived using the approach described in Section 2. As expected, the delay-bound violation probabilities for both types of services *decay exponentially* as the delay-bound increases. When increasing the number of transmit antennas or increasing the transmit power, the QoS provisioning performance can be improved. As shown by both figures, our modeling results match the simulation results well, especially for video services. Thus, Fig. 8 confirms the correctness and accuracy of our cross-layer modeling.

F. Impact of Power Control and Feedback Delay

1. The Impact of Power Control

In previous sections, we employ the AMC scheme which uses the *constant* power. However, it is well known that the optimal power-control, i.e., water-filling-based scheme, can achieve higher spectral-efficiency than the constant-power schemes. A natural question is that if the water-filling-based power-control is also optimal in terms of QoS-guarantees? Surprisingly, our analyses indicate that this is not true.

In [64, eq. (5)], the authors provided the time-domain water-filling power-control strategy for un-coded adaptive QAM modulation. We integrate the approach in [64] with our adaptive-modulation-based FSMC and analytically derive the corresponding effective capacity, which is numerically plotted against θ as shown in Fig. 9. We employ un-coded adaptive modulation instead of coded scheme used in previous sections because the un-coded scheme is analytically convenient for water-filling-based power-control. We can see from Fig. 9 that the effective-capacity of the scheme with optimal power-control is larger than that of the constant-power scheme when the QoS-exponent θ is small, which is due to the fact that the water-filling scheme always has the better spectral-efficiency than that of the constant-power scheme. However, as θ increases, the effective-capacity of the scheme with optimal power-control is lower than the constant-power scheme. This implies that *the optimal power-control that maximizes the spectral-efficiency is not necessarily optimal for QoS guarantees.* The reason behind this counter-intuitive observation is because the water-filling scheme increases the variation (instability) of the service-rate, which is undesired in terms of QoS-guarantees.

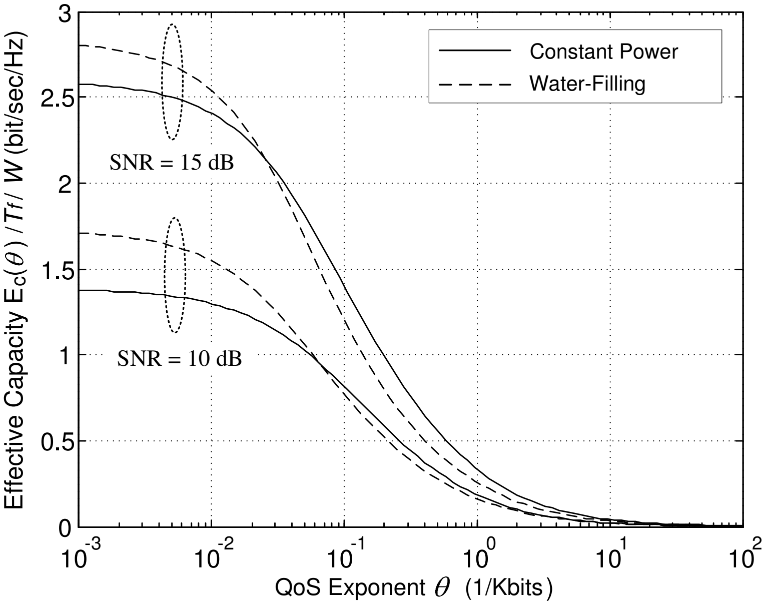


Fig. 9. The normalized effective capacity of the SISO system as a function of the QoS exponent θ with un-coded adaptive QAM modulation. The BER requirement is set to BER=10^{-3}.

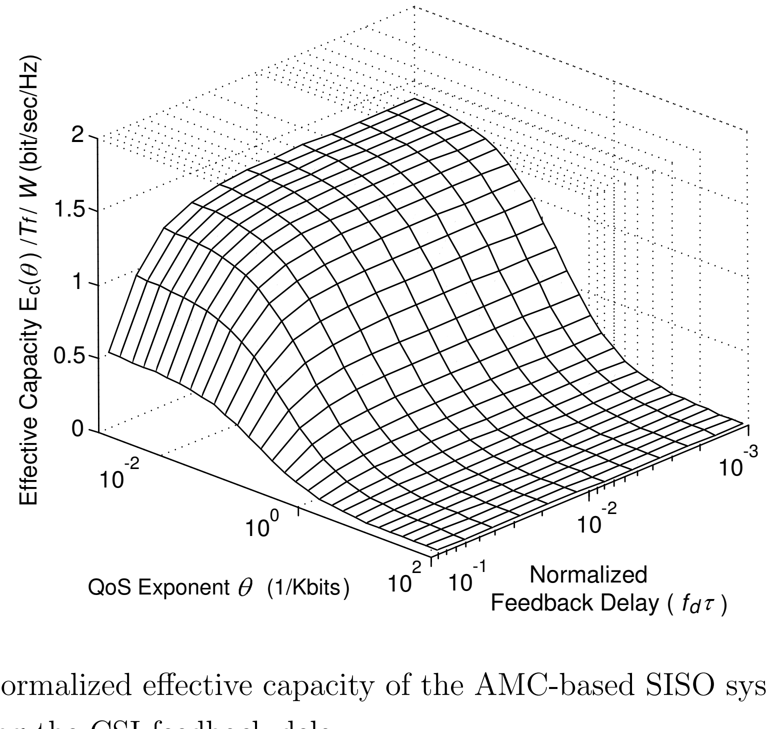


Fig. 10. The normalized effective capacity of the AMC-based SISO system when considering the CSI feedback delay.

2. The Impact of Feedback Delay

In previous sections, we assume that the CSI is *reliably* fed back to the transmitter without error and delay. However, in practical wireless networks, this assumption hardly holds. In particular, the CSI *feedback delay* is un-avoidable in most of the situations.

Denote the feedback delay by τ. In order to guarantee the reliability QoS requirement, the system needs to maintain the same PER or bit-error-rate (BER) as the case without feedback delay. As a result, the boundary points for the AMC should be re-calculated. In [64], the authors analyzed the impact of CSI feedback delay on BER performance for the adaptive modulation. In [37], we also investigated the feedback delay issue for SC/MRC scheme from BER perspective. Using the similar approach to [64] and [37], we derive analytical expressions for the effective capacity when considering CSI feedback delays, where the normalized effective capacity of the AMC-based SISO system is numerically plotted as a function of θ and $f_d\tau$ in Fig. 10. We can observe from Fig. 10 that as long as the normalized feedback delay, measured by $f_d\tau$, is within certain threshold (e.g., $f_d\tau \leq 10^{-2}$), the effective capacity is virtually unchanged with $f_d\tau$. When the normalized feedback delay further increases, the effective capacity decreases accordingly. Note that in our system model, we have $T_f \times f_d = 10^{-2}$. Thus, over the Rayleigh fading channel with Doppler frequency of $f_d = 5$ Hz, our system can tolerant CSI feedback-delay with approximately one frame's time-duration while still maintaining virtually the same statistical QoS performance.

G. Summary

We proposed the cross-layer design approach to study the interactions between physical-layer AMC and MIMO-diversity and higher-protocol-layer on the statistical QoS performance of the mobile wireless networks. We identified the critical relationships between effective bandwidth and effective capacity and analytically obtained the effective capacity function in our proposed system configurations. Our numerical results showed that the AMC and MIMO-diversity employed at physical-layer have significant impact on the statistical QoS performance at upper-protocol-layers. The proposed cross-layer modeling accurately characterize the influence of physical-layer infrastructure on statistical QoS performance at higher-protocol layers.

While in this chapter, we only investigate the single user QoS provisioning, our developed cross-layer modeling technique can be readily extended to the scenarios with multiple users sharing the wireless media in, e.g., dynamic TDMA-based wireless networks. More importantly, our developed cross-layer modeling technique also offers the practical and effective approach to develop the highly-efficient admission-control, packet scheduling, and adaptive resource-allocation schemes to guarantee the QoS for real-time multimedia traffics over mobile wireless networks.

With the power concept of effective capacity, we are able to deal with resource allocation problem for QoS guarantees. Remind that our original problem is maximizing the system throughput subject to a given delay-QoS constraint. Notice that the effective capacity can be considered as the maximum throughput under the constraint of QoS exponent θ. Therefore, by interpreting θ as the QoS constraint in our original problem, we can formulate an equivalent new problem, which is to maximize the effective capacity for a given θ. In the following chapters, we will focus on this new problem and design the corresponding power allocation algorithms under different system/network infrastructures. In the next chapter, we will first develop resource allocation policies for the single channel communication system.

CHAPTER III

RESOURCE ALLOCATION: SINGLE CHANNEL

A. Introduction

Quality-of-service (QoS) guarantees play a critically important role in future mobile wireless networks. Depending on their distinct QoS requirements, differentiated mobile users are expected to tolerate different levels of delay for their service satisfactions. For instance, non-real-time services such as data disseminations aim at maximizing the throughput with a loose delay constraint. In contrast, for real-time services like multimedia video conference, the key QoS metric is to ensure a stringent delay-bound, rather than to achieve high spectral efficiency. There also exist some services falling in between, e.g., paging and interactive web surfing, which are delay-sensitive but whose delay QoS requirements are not as stringent as those of real-time applications. The diverse mobile users impose totally different and sometimes even conflicting delay QoS constraints, which impose great challenges to the design of future mobile wireless networks.

Unlike its wired counterparts, supporting diverse delay QoS in wireless environment is much more challenging since the wireless channel has a significant impact on network performance. In particular, a deterministic delay-bound QoS guarantee over wireless networks is practically infeasible due to the time-varying nature of fading channels. Alternatively, a more practical solution is to provide the *statistical QoS guarantees* [1], where we guarantee the given delay-bound with a small violation probability.

Furthermore, for wireless communications, the most scarce radio resources are power and spectral bandwidth [71]. As a result, a great deal of research has been devoted to the techniques that can enhance the spectral efficiency of wireless systems [72]. The framework used to evaluate these techniques is mainly based on information theory, using the concept of Shannon capacity [6, 7, 73]. Among a large number of promising schemes, power and rate adaptation has been widely considered as one of the key solutions to improve the spectral efficiency. In [9, 64], the authors showed that the optimal power and rate control policy which maximizes spectral efficiency is the so-called *water-filling* algorithm. The water-filling scheme assigns more power when the channel is in good condition and less power when the channel becomes worse. In the case that the channel quality is below a certain threshold, no information is transmitted. On the other hand, a different idea of power and rate adaptation is the scheme referred as *total channel inversion* [9, 64], where the system assigns more power to combat with deep fading and less power for the good channel in order to maintain a constant signal-to-noise ratio (SNR), such that a constant rate service process can be obtained. Clearly, from the information-theoretic viewpoint, water-filling is better than total channel inversion since the former provides higher spectral efficiency. However, a natural question that follows is whether the former is also better than the latter in terms of QoS guarantees?

It is important to note that Shannon theory does not place any restrictions on complexity and delay [64]. Consequently, in order to answer the above question, it is necessary to take the QoS metrics into account when applying the prevalent information-theoretic results. Thanks to the dual concepts of *effective bandwidth* and *effective capacity*, we obtain a powerful approach to evaluate the statistical QoS performance from the networking perspective. Integrating information theory with the effective capacity, in this chapter we investigate the QoS-driven power and rate adaptation over wireless links in mobile wireless networks. The problem we are interested in is how to maximize the throughput subject to a given delay QoS constraint. We first focus on uncorrelated fading channels (also termed *block fading* or *quasi-static* fading channels) and investigate corresponding power and rate adaptation polices. Our analyses reveal an important fact that there exists a fundamental tradeoff between the throughput and the QoS provisioning. In particular, the higher throughput gain comes at the price of sacrificing more QoS provisioning, and vice versa. When the QoS constraint becomes loose, the optimal power-control law converges to the water-filling scheme, where Shannon (ergodic) capacity can be achieved. On the other hand, when the QoS constraint gets stringent, the optimal power-control law converges to the total channel inversion such that the system operates at a constant service rate. Motivated by the above observations, we then consider a more practical scenario where variable-power adaptive-modulation is applied over both uncorrelated and correlated fading channels. For simplicity, we use finite-state Markov chain (FSMC) to model the correlated channel processes. The FSMC-based channel model was previously proposed by Wang and Moayeri [65]. Then, this model has been extensively studied for both Rayleigh and Nakagami-m fading channel (e.g., see [66–68] and references therein). For both block fading and FSMC-correlated fading channels, we derive the corresponding power and rate adaptation policies. Our obtained results suggest that channel correlation has a significant impact on QoS-driven power and rate allocations. The higher the correlation is, the faster the power-control policy converges to the total channel inversion as the QoS constraint becomes more stringent. Finally, we conduct simulations to verify that although the FSMC-based channel model is not perfectly accurate, the power-control law derived from it can be applied to the more general Jake's channel model [39], which has been widely used and extensively studied in literatures.

The rest of the chapter is organized as follows. Section B describes our system model. Section C develops the optimal power and rate adaptation scheme that can

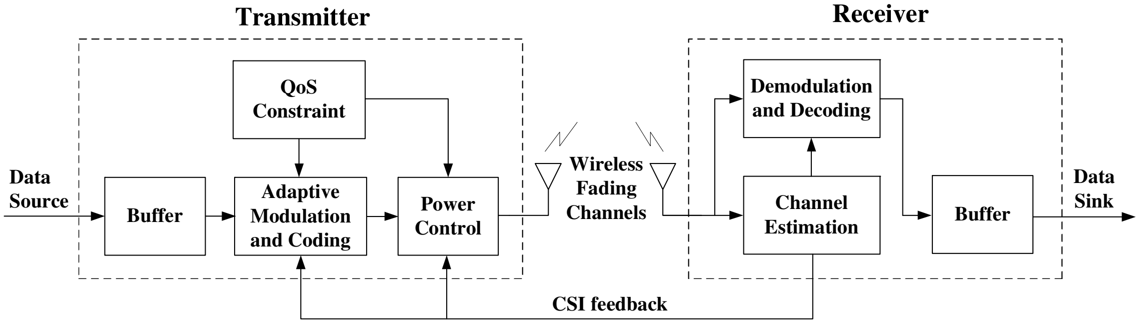


Fig. 11. The single channel system model.

maximize the effective capacity. Section D applies the above analyses to a more practical adaptive modulation-based scheme. Section E discusses the impact of channel correlation on the power and rate adaptations. Section F conducts simulations to evaluate the validity of our proposed adaptive schemes on the more general Jake's channel model. The chapter concludes with Section G.

B. System Model

The system model is illustrated in Fig. 11. We concentrate on the discrete-time system over a point-to-point wireless link between the transmitter and the receiver. Let us denote the system's total spectral bandwidth by B, the mean transmit power by $\overline{P}$, and the power density of the complex additive white Gaussian noise (AWGN) by $N_0/2$ per dimension, respectively. First, the upper-protocol-layer packets are divided into *frames* at the datalink layer, which forms the "data source" as shown in Fig. 11. We assume that the frames have the same time duration, which is denoted by T_f. The frames are stored at the transmit buffer and then split into bit-streams at the physical layer. Based on the QoS constraint and the channel-state information (CSI) fed back from the receiver, the adaptive modulation and power control are employed, respectively, at the transmitter. The reverse operations are executed at the receiver side. Finally, the frames are recovered at the "data sink" for further processing.

The discrete-time channel fading process is assumed to be stationary and ergodic, which is invariant within a frame's time-duration T_f, but varies from one frame to another. Moreover, the wireless channel is flat-fading with its envelope following Nakagami-m distribution.[1]

Denote the channel envelope process by $\{\alpha[i], i = 1, 2, ...\}$, where i is the time index of the frame. If we use constant power assignment, then the instantaneous transmit power, denoted by $P[i]$, is equal to $P[i] = \overline{P}$. The instantaneous received SNR, denoted by $\gamma[i]$, can be expressed as $\gamma[i] = \overline{P}\alpha^2[i]/(N_0 B)$, with its mean $\overline{\gamma} = \overline{P}\mathbb{E}\{\alpha^2[i]\}/(N_0 B)$, where $\mathbb{E}\{\cdot\}$ denotes the expectation. The probability density function (pdf) of $\gamma[i]$, denoted by $p_\Gamma(\gamma)$, can be expressed as [34]

$$p_\Gamma(\gamma) = \frac{\gamma^{m-1}}{\Gamma(m)} \left(\frac{m}{\overline{\gamma}}\right)^m \exp\left(-\frac{m}{\overline{\gamma}}\gamma\right), \ \gamma \geq 0 \tag{3.1}$$

where $\Gamma(\cdot)$ represents the Gamma function and m denotes the fading parameter of Nakagami-m distribution. Throughout this chapter, we assume that the CSI is perfectly estimated at the receiver and reliably fed back to the transmitter without delay. The discussions of the imperfect CSI are not the focus of this chapter.

Our original problem is maximizing the throughput subject to a given delay-QoS constraint. Notice that the effective capacity can be considered as the maximum throughput under the constraint of QoS exponent θ. Therefore, by interpreting θ as the QoS constraint in our original problem, we can formulate an equivalent new problem, which is to maximize the effective capacity for a given θ. In the following chapters, we will focus on this new problem and design the corresponding power

[1] The power and rate adaptation scheme discussed in this chapter can be applied to any other continuous channel distributions. We use Nakagami-m distribution in this chapter as a general example.

allocation algorithms.

C. Optimal Resource Allocation Policy

Conventionally, the power-control law can be expressed as a function of the instantaneous SNR $\gamma[i]$. However, our power-adaptation policy, denoted by $\mu\left(\theta, \gamma[i]\right)$, is a function of not only the instantaneous SNR $\gamma[i]$, but also the QoS exponent θ. Applying the power adaptation, the instantaneous transmit power becomes $P[i] = \mu\left(\theta, \gamma[i]\right)\overline{P}$. Note that the mean transmit power is upper-bounded by $\overline{P}$. Therefore, the power-control law needs to satisfy the mean power constraint:

$$\int_0^{\infty} \mu(\theta, \gamma) p_{\Gamma}(\gamma) d\gamma \leq 1, \text{ for all } \theta > 0. \tag{3.2}$$

In this section, we also make the following two assumptions.

A1: We first assume that the channel is block fading. We make such an assumption due to the following reasons. First, the effective capacity expression (2.15) in uncorrelated case only depends on marginal statistics of a service process, which is much simpler than the general expression given by (2.14), where the higher order statistics of the service process are required. Second, we will show in Section E that the resource allocation policy derived for block fading channel can be applied to correlated fading channel with a certain modifications.

A2: An underlying assumption is that the block duration is significantly shorter than the queuing delay, such that it is reasonable to only focus on queuing delay. This assumption is also feasible since in practical communication systems, the block duration is in the order to millisecond, but the queuing delay for real-time applications is in the order of tens of millisecond.

A3: We further assume that given the instantaneous SNR $\gamma[i]$ and the corresponding power-control law $\mu\left(\theta, \gamma[i]\right)$, the adaptive modulation and coding scheme can achieve the instantaneous capacity. Thus, the instantaneous service rate $R[i]$ of the frame i can be expressed as[2]

$$R[i] = T_f B \log_2\left(1 + \mu\left(\theta, \gamma[i]\right)\gamma[i]\right). \tag{3.3}$$

In the following discussions, we omit the discrete time-index i for simplicity. Using (2.15), (3.2), and (3.3), we can formally formulate our maximization problem as follows:

$$\mathrm{E}_{\mathrm{C}}^{\mathrm{opt}}(\theta) = \max_{\mu(\theta,\gamma):\int_0^\infty \mu(\theta,\gamma)p_\Gamma(\gamma)d\gamma=1} \left\{-\frac{1}{\theta}\log\left(\int_0^\infty e^{-\theta T_f B \log_2\left(1+\mu(\theta,\gamma)\gamma\right)} p_\Gamma(\gamma)d\gamma\right)\right\}. \tag{3.4}$$

where $\mathrm{E}_{\mathrm{C}}^{\mathrm{opt}}(\theta)$ denotes the maximum effective capacity achieved by the optimal policy. We derive the following theorem to characterize the optimal power and rate adaptation policy.

Theorem 1. *The optimal power-control policy, denoted by $\mu_{\mathrm{opt}}(\theta, \gamma)$, which maximizes the effective capacity given in (3.4), is determined by*

$$\mu_{\mathrm{opt}}(\theta, \gamma) = \begin{cases} \dfrac{1}{\gamma_0^{\frac{1}{\beta+1}}\gamma^{\frac{\beta}{\beta+1}}} - \dfrac{1}{\gamma}, & \gamma \geq \gamma_0 \\ 0, & \gamma < \gamma_0 \end{cases} \tag{3.5}$$

where we define $\beta \triangleq \theta T_f B/\log 2$ as the normalized *QoS exponent and γ_0 as the cutoff SNR threshold, which can be numerically obtained by meeting the mean power*

[2] Note that in our model, the unit for the service rate $R[i]$ and the effective capacity $\mathrm{E}_{\mathrm{C}}(\theta)$ is "bits per frame".

constraint:

$$\int_{\gamma_0}^{\infty}\left(\frac{1}{\gamma_0^{\frac{1}{\beta+1}}\gamma^{\frac{\beta}{\beta+1}}}-\frac{1}{\gamma}\right)p_{\Gamma}(\gamma)d\gamma=1. \tag{3.6}$$

Proof. The proof is provided in Appendix D. □

Theorem 1 gives the optimal power-control policy which maximizes the effective capacity. We can observe from (3.5) that as $\theta \to 0$, the optimal policy $\mu_{\text{opt}}(\theta,\gamma)$ converges to

$$\lim_{\theta\to 0}\mu_{\text{opt}}(\theta,\gamma)=\begin{cases}\dfrac{1}{\gamma_0}-\dfrac{1}{\gamma}, & \gamma\ge\gamma_0\\ 0, & \gamma<\gamma_0\end{cases} \tag{3.7}$$

which is just the water-filling formula in [64, eq. (5)]. Thus, our QoS-driven power and rate adaptation scheme reduces to the water-filling algorithm when the system can tolerate an arbitrarily long delay, which is expected since water-filling is well-known the optimal power allocation strategy without delay constraint. On the other hand, as the QoS exponent $\theta\to\infty$, the cutoff threshold $\gamma_0\to 0$ (note that $\gamma_0=\lambda/\beta$ as detailed in Appendix D). Therefore, the system does not enter the outage state almost surely. The optimal power control $\mu_{\text{opt}}(\theta,\gamma)$ converges to

$$\lim_{\theta\to\infty}\mu_{\text{opt}}(\theta,\gamma)=\frac{\sigma}{\gamma} \tag{3.8}$$

where $\sigma=(m-1)\overline{\gamma}/m$ for $m\ge 1$, which becomes the policy of the total channel inversion. Thus, for stringent delay QoS constraints, the optimal power control becomes the total channel inversion. Note that if the fading parameter $m<1$, implying that the fading is severer than Rayleigh, then no total channel inversion scheme exists since the transmit power is not enough to totally invert the channel. In this case, the

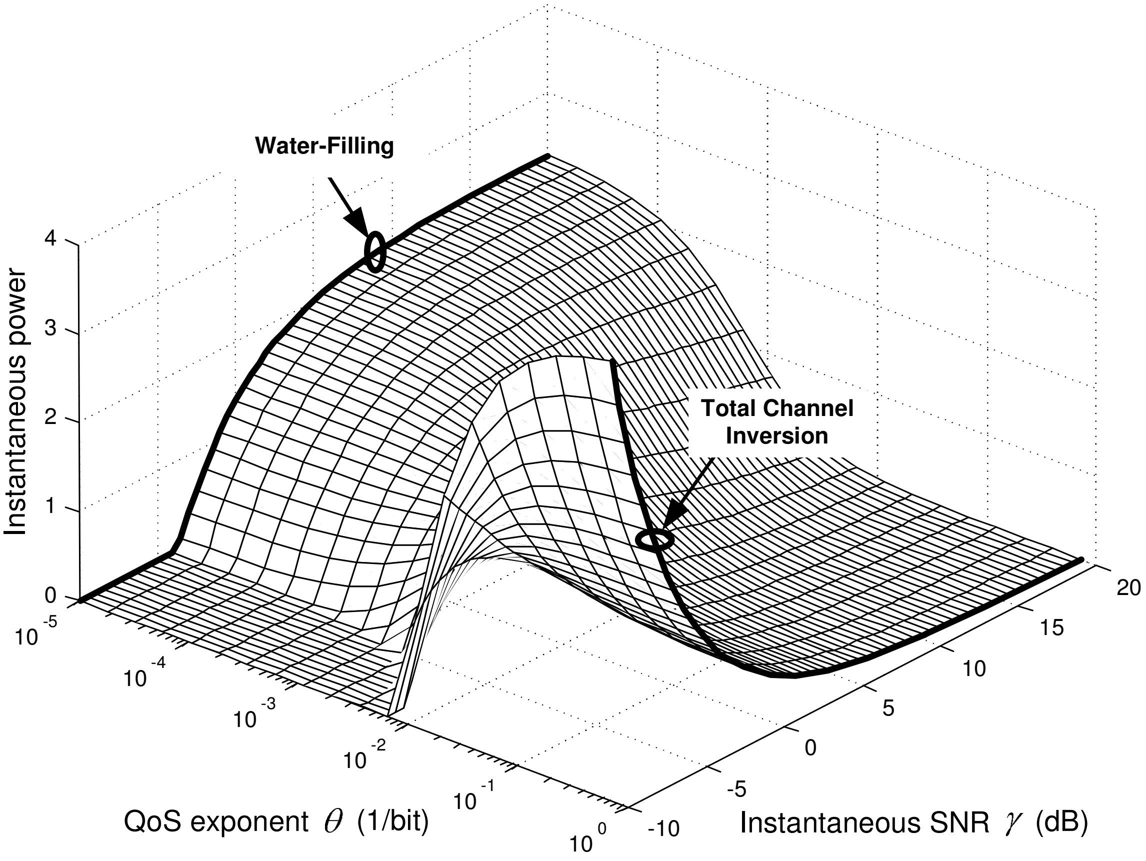
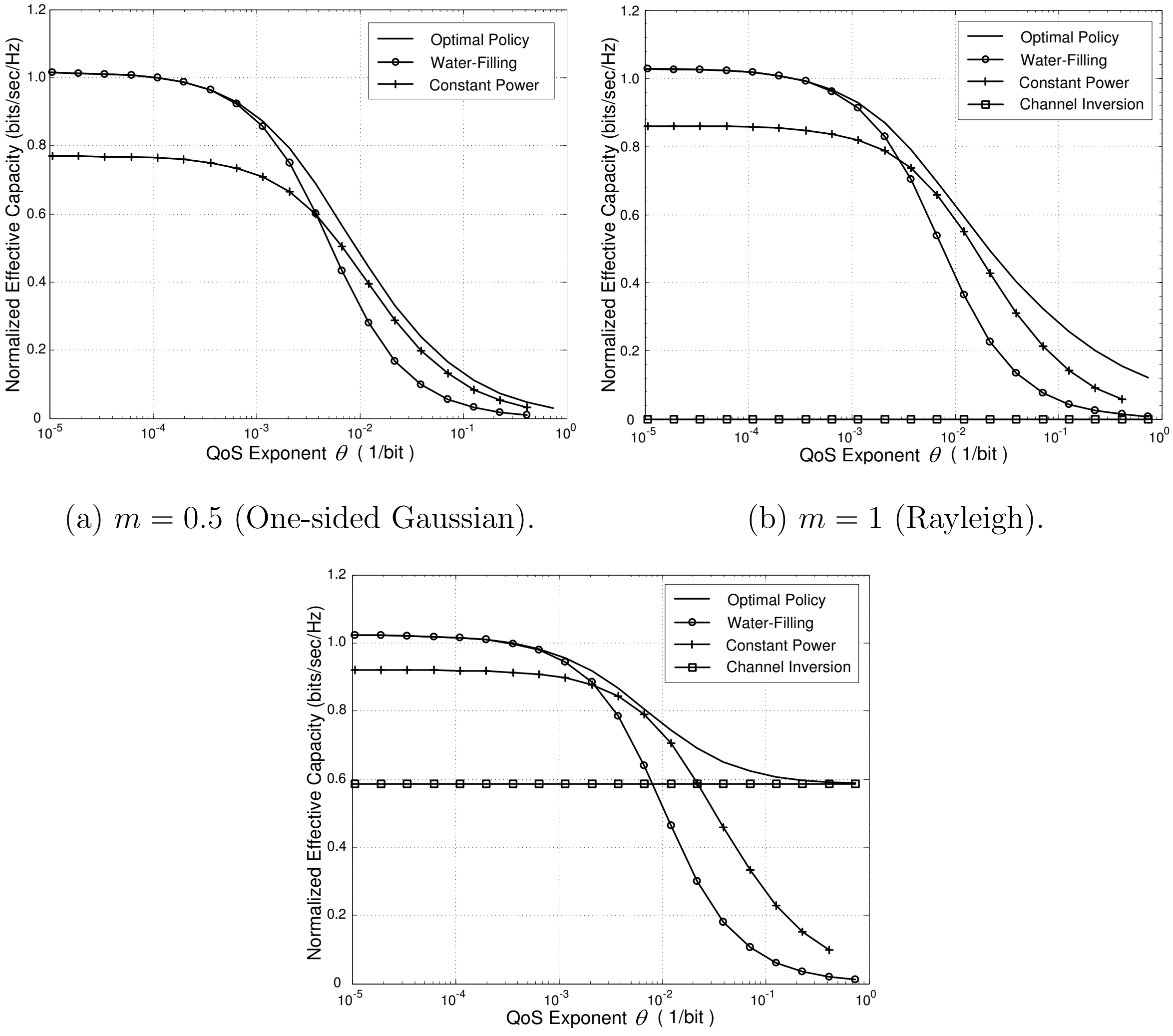


Fig. 12. The optimal power-adaptation policy. The fading parameter $m = 2$ and the average SNR $\overline{\gamma} = 0$ dB.

cutoff threshold γ_0 will converge to a small positive number as $\theta \to \infty$. Thus, the optimal power and rate adaptation policy becomes *truncated channel inversion* [64]. It is also worth noting that the optimal power-adaptation policy $\mu_{\text{opt}}(\theta, \gamma)$ depends on frame duration T_f and spectral bandwidth B through the parameter β, where a system with larger value of $T_f B$ can support more stringent QoS requirements.

In all the following numerical solutions or simulation results, which are presented in Figs. 12 – 19, we set the frame duration $T_f = 2$ ms and the spectral bandwidth $B = 10^5$ Hz. The other system parameters are detailed respectively in each of these figures. Using (3.5), we plot the instantaneous power assignments of the optimal



(a) $m = 0.5$ (One-sided Gaussian).

(b) $m = 1$ (Rayleigh).

(c) $m = 2$.

Fig. 13. The Shannon-capacity-based effective capacity under different resource allocation policies. The average SNR $\overline{\gamma} = 0$ dB.

power-adaptation policy in Fig. 12. We can observe from Fig. 12 that for small θ, the power control assigns more power to the better channel and less power to the worse channel. In contrast, for large θ, the power control assigns less power to the better channel, but more power to the worse channel. As the QoS exponent θ varies between $(0, \infty)$, reflecting different delay QoS constraints, the corresponding optimal power-adaptation policy swings between the water-filling and the total channel inversion schemes.

Given the optimal power and rate adaptation policy, we can derive the closed-form expression for the maximum effective capacity $\mathrm{E}_{\mathrm{C}}^{\mathrm{opt}}(\theta)$ as follows:

$$\begin{aligned} \mathrm{E}_{\mathrm{C}}^{\mathrm{opt}}(\theta) &= -\frac{1}{\theta}\log\left(\int_0^{\infty} e^{-\beta\log\left(1+\mu_{\mathrm{opt}}(\theta,\gamma)\gamma\right)} p_{\Gamma}(\gamma)d\gamma\right) \\ &= -\frac{1}{\theta}\Bigg\{\log\left(\left[\frac{m\gamma_0}{\overline{\gamma}}\right]^{\frac{\beta}{\beta+1}}\Gamma\left(m-\frac{\beta}{\beta+1},\frac{m\gamma_0}{\overline{\gamma}}\right)\right. \\ &\qquad \left. +\,\gamma\left(m,\frac{m}{\overline{\gamma}}\gamma_0\right)\right) - \log\Big(\Gamma(m)\Big)\Bigg\}. \end{aligned} \tag{3.9}$$

where $\gamma(\cdot,\cdot)$ and $\Gamma(\cdot,\cdot)$ denote the lower and upper incomplete Gamma functions, respectively.

For comparison purposes, we also derive the closed-form expressions of the effective capacity for other commonly used power-control policies, including the water-filling scheme, the constant power approach, and the total channel inversion. Omitting the derivation details, we obtain the closed-form expressions of the effective capacity for water-filling, denoted by $\mathrm{E}_{\mathrm{C}}^{\mathrm{WF}}(\theta)$, as follows:

$$\mathrm{E}_{\mathrm{C}}^{\mathrm{WF}}(\theta) = -\frac{1}{\theta}\Bigg\{\log\left(\left[\frac{m\gamma_0}{\overline{\gamma}}\right]^{\beta}\Gamma\left(m-\beta,\frac{m\gamma_0}{\overline{\gamma}}\right)+\gamma\left(m,\frac{m}{\overline{\gamma}}\gamma_0\right)\right)-\log\Big(\Gamma(m)\Big)\Bigg\}, \tag{3.10}$$

and the effective capacity for the constant power approach, denoted by $\mathrm{E}_{\mathrm{C}}^{\mathrm{const}}(\theta)$, as follows:

$$\begin{aligned} \mathrm{E}_{\mathrm{C}}^{\mathrm{const}}(\theta) &= -\frac{1}{\theta}\log\left(\frac{\Gamma(\beta-m)}{\Gamma(\beta)}\left(\frac{m}{\overline{\gamma}}\right)^{m} {}_1F_1\left(m; m-\beta+1; \frac{m}{\overline{\gamma}}\right)\right. \\ &\quad \left. +\frac{\Gamma(m-\beta)}{\Gamma(m)}\left(\frac{m}{\overline{\gamma}}\right)^{\beta} {}_1F_1\left(\beta; \beta-m+1; \frac{m}{\overline{\gamma}}\right)\right) \qquad (3.11) \end{aligned}$$

respectively, where ${}_1F_1\left(\cdot;\cdot;\cdot\right)$ denotes the confluent Hypergeometric function [74]. Finally, the effective capacity of total channel inversion is simply a constant equal to $T_f B \log_2(1+\sigma)$.

The normalized effective capacity (which is defined as the effective capacity divided by B and T_f, and thus has the unit of "bits/sec/Hz") comparisons between different power and rate adaptation schemes are shown in Fig. 13. As expected, our proposed optimal power and rate adaptation always achieves the maximum effective capacity among all control policies. The optimal scheme converges to the water-filling for small θ and to the total channel inversion for large θ (when the total channel inversion exists). Note that for one-sided Gaussian channel ($m = 0.5$) and Rayleigh channel ($m = 1$), even using the optimal policy, the effective capacity also converges to zero as the QoS exponent $\theta \to \infty$. However, this is the best that the power control can do to maximize the effective capacity. This implies that no matter how much power and spectral bandwidth resource are assigned and no matter how elegant coding/modulation is employed, if no other technique (e.g., diversity or multiplexing) helps to compensate for the fading effect, Nakagami-m channels with $m \leq 1$ cannot support stringent delay QoS requirement when θ is large, which is also coincident with the fact that the zero-outage capacity of Rayleigh fading channel is zero [10].

D. QoS-Driven Resource Allocation for MQAM

Based on Shannon theory and the concept of effective capacity, Section C discusses the resource allocation when using ideal channel codes. In this section, we study the scenario where the transmitter employs adaptive MQAM modulation.

1. Continuous MQAM

We first assume that there is no restriction on the constellation size of adaptive MQAM, which implies that the rate of the service process can be adapted continuously. In [64], Goldsmith and Chua showed that the continuous rate adaptive MQAM has a constant power loss as compared to the Shannon capacity, where the constant only depends on bit-error rate (BER) requirement. Specifically, for each given received SNR γ and power-control policy $\mu\left(\theta,\gamma\right)$, the corresponding constellation size, denoted by $M(\gamma)$, is determined by [64, eq. (20)]

$$M(\gamma) = 1 + K\mu\left(\theta,\gamma\right)\gamma \tag{3.12}$$

where K is defined as $K \triangleq -1.5/\log(5\text{BER})$, with BER denoting the required bit-error rate. Note that continuous rate MQAM is originally proposed to investigate the insight relationship between the Shannon capacity and the achievable spectral efficiency of MQAM modulation [64]. In practice, the constellation size $M(\gamma)$ can only be selected from a finite discrete set, which will be detailed in Section 2.

Using (3.12), the service rate of continuous rate MQAM, denoted by R_{M}, can be expressed as

$$R_{\text{M}} = T_f B \log_2\Big(1 + K\mu\left(\theta,\gamma\right)\gamma\Big). \tag{3.13}$$

Comparing (3.13) with (3.3), we can find that the only difference between these two is a constant power loss of K. Thus, the problem of maximizing the effective capacity for continuous MQAM can be solved in a similar manner to that for deriving the maximum effective-capacity in Section C. Skipping the detailed derivations, we obtain the optimal power and rate adaptation policy for continuous rate adaptive MQAM, denoted by $\mu_{\text{opt}}^{\text{M}}(\theta, \gamma)$, as follows:

$$\mu_{\text{opt}}^{\text{M}}(\theta, \gamma) = \begin{cases} \dfrac{1}{K\gamma_K^{\frac{1}{\beta+1}}\gamma^{\frac{\beta}{\beta+1}}} - \dfrac{1}{K\gamma}, & \gamma \geq \gamma_K \\ 0, & \gamma < \gamma_K \end{cases} \tag{3.14}$$

where γ_K is the new cutoff threshold, which needs to meet the mean power constraint:

$$\int_{\gamma_K}^{\infty} \left(\frac{1}{K\gamma_K^{\frac{1}{\beta+1}}\gamma^{\frac{\beta}{\beta+1}}} - \frac{1}{K\gamma} \right) p_\Gamma(\gamma) d\gamma = 1. \tag{3.15}$$

Once γ_K is obtained, we can show the resulting expression of the effective capacity is the same as (3.9), except that γ_0 in (3.9) should be replaced by γ_K. It is clear that the power adaptation law of (3.14) also follows the same trends as (3.5), which adjusts the power assignment between the water-filling and the total channel inversion, depending on the specific value of θ. Similarly, for the other non-optimal power and rate adaptation policies, we can also derive their corresponding effective capacity expressions, which are omitted for lack of space, but are evaluated by the numerical solutions as shown in Fig. 14.

Fig. 14 illustrates the normalized effective capacity comparisons between the Shannon theory-based upper-bound and the continuous rate adaptive MQAM. We can observe that as the BER requirement becomes more stringent, the effective capacities of both optimal and non-optimal schemes decrease accordingly. However, our proposed power and rate adaptations are always the optimal schemes in each group with the same BER requirement. Agreeing with our observations, the effective

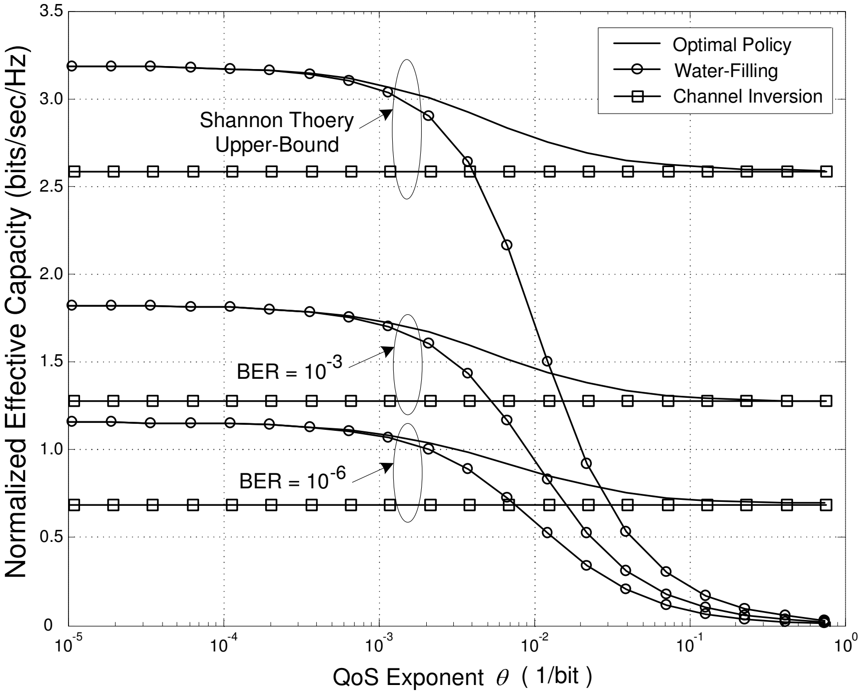


Fig. 14. The effective capacity with adaptive MQAM. The average SNR $\overline{\gamma} = 10$ dB and the fading parameter $m = 2$.

capacities of the optimal scheme converge to the water-filling for small θ and to the total channel inversion for large θ, respectively.

2. Discrete MQAM

The continuous rate assumption for the adaptive MQAM is not too practical. In this section, we relax this assumption by requesting that there be only N possible constellation sizes available. Specifically, we partition the entire SNR range by N non-overlapping consecutive intervals, resulting in $N+1$ boundary points denoted by $\{\Gamma_n\}_{n=0}^{N}$, where $\Gamma_0 < \Gamma_1 < \cdots < \Gamma_N$ with $\Gamma_0 = 0$ and $\Gamma_N = \infty$. Correspondingly, the adaptive modulation is selected to be in mode n if the SNR γ falls into the range $\Gamma_n \leq \gamma < \Gamma_{n+1}$. The constellation used for the zero-th mode is $M_0 = 0$ and for the nth mode is M_n-QAM, where $M_n = 2^n$ with $n = 1, 2, ..., N-1$. Thus, the spectral efficiency by using the nth mode is n bits/sec/Hz. The service rate of the nth mode, denoted by ν_n, is given by

$$\nu_n = T_f B n, \quad \text{for } n = 0, 1, ..., N-1. \tag{3.16}$$

To find the optimal power and rate adaptation policy for discrete rate adaptive MQAM, we first need to know how to choose the boundary points $\{\Gamma_n\}_{n=1}^{N-1}$ to maximize the effective capacity. Substituting (3.14) into (3.12), we get

$$M(\gamma) = \left(\frac{\gamma}{\gamma_K}\right)^{\frac{1}{\beta+1}} \Longrightarrow \gamma = [M(\gamma)]^{\beta+1} \gamma_K. \tag{3.17}$$

Although (3.17) is originally derived from continuous rate adaptive MQAM, it provides the guideline in choosing the boundaries for the discrete rate MQAM. Based on (3.17), we obtain the SNR boundaries $\{\Gamma_n\}_{n=1}^{N-1}$ for discrete rate MQAM as follows:

$$\Gamma_n = M_n^{\beta+1} \gamma_K^* \tag{3.18}$$

where γ_K^* denotes the new cutoff threshold for discrete rate MQAM. For each given γ_K^*, the boundaries $\{\Gamma_n\}_{n=1}^{N-1}$ are determined by (3.18). Then, the power-control policy is to retain a constant power for each mode $n > 0$ such that the BER requirements are satisfied. Thus, we obtain the optimal policy for the nth mode, denoted by $\mu_{\text{opt}}^n(\theta, \gamma)$, as follows:

$$\mu_{\text{opt}}^n(\theta, \gamma) = \begin{cases} \dfrac{(M_n - 1)}{K\gamma}, & 1 \leq n \leq N-1 \\ 0, & n = 0. \end{cases} \tag{3.19}$$

Let us further define $M_N \triangleq \infty$. Then, the cutoff threshold γ_K^* is determined by the mean power constraint

$$\sum_{n=0}^{N-1} \int_{M_n^{\beta+1}\gamma_K^*}^{M_{n+1}^{\beta+1}\gamma_K^*} \mu_n(\theta, \gamma) p_\Gamma(\gamma) d\gamma = 1 \tag{3.20}$$

which can be solved numerically.

We can observe from (3.18) that as $\theta \to 0$, the boundary selection policy becomes:

$$\lim_{\theta \to 0} \Gamma_n = M_n \gamma_K^* \tag{3.21}$$

which is same as the selection policy for the discrete rate water-filling algorithm [64, eq. (29)]. On the other hand, as $\theta \to \infty$, the threshold γ_K^* vanishes to zero, making mode 0 infinitely small. At the same time, one of the other $(N-1)$ modes dominates the entire SNR range. Again, the power-control policy converges to the total channel inversion in this case.

Using (3.19), we plot the instantaneous power assignments of the optimal power-adaptation policy in Fig. 15. We can observe from Fig. 15 that the power control curve has the zigzag shape due to the constellation constraint. However, the power control policy still varies between the discrete rate water-filling at small θ and the total channel inversion at large θ, which is consistent with that of the continuous rate

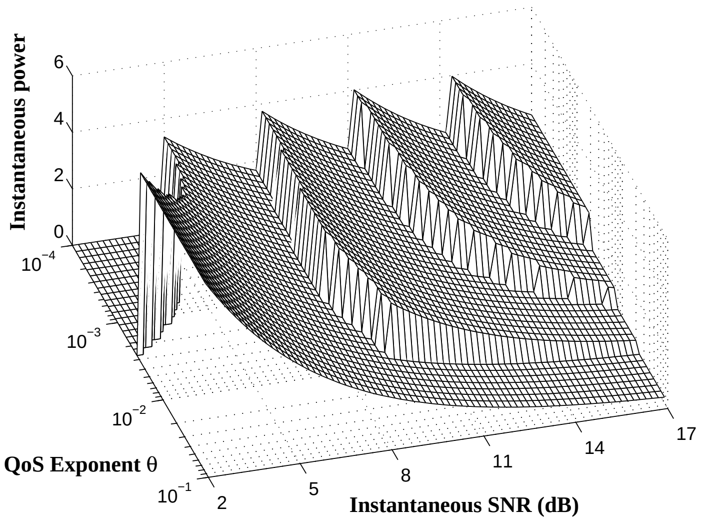


Fig. 15. The power-adaptation strategy for discrete rate adaptive MQAM. The average SNR $\overline{\gamma} = 10$ dB, the fading parameter $m = 2$, BER$= 10^{-3}$, and the number of modes $N = 5$.

MQAM.

Using (2.15), we derive the effective capacity under the optimal policy, denoted by $\hat{\mathrm{E}}_{\mathrm{C}}^{\mathrm{opt}}(\theta)$, as follows:

$$\hat{\mathrm{E}}_{\mathrm{C}}^{\mathrm{opt}}(\theta) = -\frac{1}{\theta}\log\left(\sum_{n=0}^{N-1}\pi_n e^{-\theta\nu_n}\right) \tag{3.22}$$

where ν_n is given by (3.16) and

$$\pi_n = \int_{\Gamma_n}^{\Gamma_{n+1}} p_\Gamma(\gamma)d\gamma = \frac{1}{\Gamma(m)}\left[\gamma\left(m, \frac{m}{\overline{\gamma}}\Gamma_{n+1}\right) - \gamma\left(m, \frac{m}{\overline{\gamma}}\Gamma_n\right)\right] \tag{3.23}$$

with $\{\Gamma_n\}_{n=0}^{N}$ given by (3.18).

Fig. 16 compares the normalized effective capacities between continuous rate MQAM and discrete rate MQAM using both the optimal and non-optimal policies. As shown by Fig. 16, the discrete rate MQAM under the optimal policy suffers from a certain loss in performance as compared to the continuous rate MQAM due to the

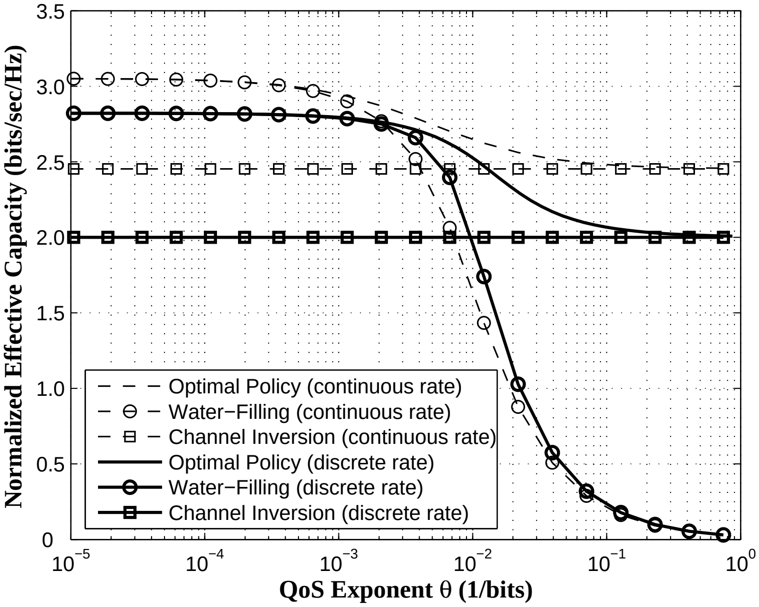


Fig. 16. The effective capacity with continuous rate and discrete rate using different power-control policies. The average SNR $\overline{\gamma} = 15$ dB, the fading parameter $m = 2$, BER$= 10^{-3}$, and the number of modes $N = 4$.

discrete constellation constraint. However, such a performance loss is not significant. An interesting phenomenon is that the effective capacity of discrete rate water-filling is even larger than that of the continuous rate water-filling for large θ, which is because the service rate of discrete rate scheme has smaller variance than that of the continuous rate scheme, where a service process with smaller variance can support more stringent delay QoS requirement.

E. Impact of Channel Correlation

1. Effective Capacity for FSMC-Modeled Channel

We derive the above analytical results by using a block fading channel model. However, this model is not always valid. In most scenarios, it is more practical to consider the correlated wireless channel models. There exist a number of models characterizing the correlated channel fading processes. For instance, the Jake's model [39] has been widely accepted as an accurate modeling approach. Based on the Jake's model, the autocorrelation of the channel gain, denoted by $A_g(\tau)$, can be expressed as $A_g(\tau) = J_0^2(2\pi f_d \tau)$ [39], where $J_0(\cdot)$ denotes the zero-th order Bessel function of the first kind and f_d is the maximum Doppler frequency. However, if using the Jake's model in our systems, it is hard to derive the effective capacity expression from (2.14). Then, it is even harder to find the power and rate adaptation policies. Therefore, we apply FSMC to model the correlated service process for simplicity.

Integrating the FSMC model with our discrete rate adaptive MQAM, the state of FSMC corresponds to the mode of adaptive modulation. Let $p_{i,j}$ denote the transition probability from state i to state j. We assume a slow-fading channel model such that transition only happens between adjacent states [65, 66]. Under this assumption, we have $p_{ij} = 0$, if $|i-j| > 1, \forall i, j \in \{0, 1, ..., N-1\}$. The adjacent transition probabilities can be approximated as follows [65]:

$$\begin{cases} p_{n,n+1} \approx \frac{N_\Gamma(\Gamma_{n+1})T_f}{\pi_n}, & \text{for } n = 0, 1, ..., N-2, \\ p_{n,n-1} \approx \frac{N_\Gamma(\Gamma_n)T_f}{\pi_n}, & \text{for } n = 1, 2, ..., N-1 \end{cases} \tag{3.24}$$

where $\{\Gamma_n\}_{n=0}^{N-1}$ and $\{\pi_n\}_{n=0}^{N-1}$ are given by (3.18) and (3.23), respectively, and $N_\Gamma(\gamma)$ is the level-crossing rate (LCR) calculated at SNR value of γ, which is given by [34]

$$N_\Gamma(\gamma) = \frac{\sqrt{2\pi} f_d}{\Gamma(m)} \left(\frac{m\gamma}{\overline{\gamma}}\right)^{m-\frac{1}{2}} \exp\left(-\frac{m\gamma}{\overline{\gamma}}\right). \tag{3.25}$$

Then, the remaining transition probabilities can be derived by using (3.24) as follows:

$$\begin{cases} p_{0,0} = 1 - p_{0,1} \\ p_{N-1,N-1} = 1 - p_{N-1,N-2} \\ p_{n,n} = 1 - p_{n,n-1} - p_{n,n+1}, \;\; n = 1, ..., N-2. \end{cases} \tag{3.26}$$

Thus, applying (3.24) and (3.26), we obtain the transition probability matrix of the FSMC, which is denoted by $\mathbf{P} = [p_{ij}]_{N\times N}$. Based on FSMC-modeled service process and discussion in Chapter II-D, the effective capacity of the service process can be expressed as

$$\mathrm{E_C}(\theta) = -\frac{1}{\theta}\log\Big(\rho\big\{\mathbf{P}\,\mathbf{\Phi}(\theta)\big\}\Big) \tag{3.27}$$

where $\mathbf{P}$ is the transition probability matrix of the FSMC mentioned above, $\mathbf{\Phi}(\theta) \triangleq \mathrm{diag}\left\{e^{-\nu_0\theta}, e^{-\nu_1\theta}, ..., e^{-\nu_{N-1}\theta}\right\}$, where $\{\nu_n\}_{n=0}^{N-1}$ is given by (3.16), and $\rho\{\cdot\}$ denotes the spectral radius of the matrix.

2. Resource Allocation for FSMC-Modeled Channel

Although we employ the FSMC-model-based service process, it is still difficult to *directly* derive the power and rate adaptation policy to maximize the effective capacity described by (3.27). Fortunately, the wireless channel offers a unique feature that allows us to obtain the simple but *near-optimal* solution. We observe that the FSMC-modeled service process satisfy the properties described by the following proposition.

Proposition 4. ***If*** *we denote the effective capacity functions of two FSMC-based service processes by* $\mathrm{E}_{\mathrm{C}_1}(\theta)$ *and* $\mathrm{E}_{\mathrm{C}_2}(\theta)$*, and they have the same marginal statistics, but differ in Doppler frequencies denoted by* f_{d_1} *and* f_{d_2}*, respectively,* ***then*** *the following equation holds:*

$$\mathrm{E}_{\mathrm{C}_1}(\theta) \approx \mathrm{E}_{\mathrm{C}_2}\left(\frac{f_{d_2}}{f_{d_1}}\theta\right) \tag{3.28}$$

Proof. The proof is provided in Appendix E. □

Remarks: Proposition 4 says that $\mathrm{E}_{\mathrm{C}_1}(\theta)$ is approximately a horizontal-shifted version of $\mathrm{E}_{\mathrm{C}_2}(\theta)$ along θ-axis (when θ-axis uses the logarithmic scale), where the difference between these two functionals is $10\log_{10}(f_{d_2}/f_{d_1})$ dB. Specifically, if $(f_{d_2}/f_{d_1}) > 1$, then $\mathrm{E}_{\mathrm{C}_1}(\theta)$ is a left-shifted version of $\mathrm{E}_{\mathrm{C}_2}(\theta)$; otherwise, $\mathrm{E}_{\mathrm{C}_1}(\theta)$ is a right-shifted version of $\mathrm{E}_{\mathrm{C}_2}(\theta)$.

It is well known that the Doppler frequency f_d characterizes the time correlation of channel fading processes. The larger the Doppler frequency f_d, the lower the correlation of the service process. When the Doppler frequency is large enough, the channel process can be approximately considered as uncorrelated, just like the block fading channel. For example, based on our system parameters and the standard Jake's channel model, the autocorrelation $A_g(T_f)$ passes through its first zero-point at the Doppler frequency of 191.25 Hz, which we denote by f_d^{Jake}. However, due to the inaccuracy of the FSMC-based channel model, such a Doppler frequency, denoted by f_d^{FSMC}, is about $f_d^{\mathrm{FSMC}} = 300$ Hz with the same system parameters. In the following discussions, when it is unnecessary to distinguish between these two, we denote both f_d^{Jake} and f_d^{FSMC} by f_d^*, which characterizes the Doppler frequency where the channel process can be approximately considered as uncorrelated, like the block fading channel.

Let $\mathrm{E}_{\mathrm{C}}^*(\theta)$ denote the effective capacity in a block fading channel model. Then, based on Proposition 4, for an FSMC-correlated channel with the same marginal statistics and a Doppler frequency f_d ($f_d \ll f_d^*$), its effective capacity, denoted by $\mathrm{E}_{\mathrm{C}}^{(f_d)}(\theta)$, can be approximated as $\mathrm{E}_{\mathrm{C}}^{(f_d)}(\theta) \approx \mathrm{E}_{\mathrm{C}}^*(\kappa\theta)$, where

$$\kappa = \frac{f_d^*}{f_d}. \tag{3.29}$$

Likewise, for each power-adaptation policy $\mu(\theta, \gamma)$ that is used for block fading channels, the new policy $\mu(\kappa\theta, \gamma)$ can be applied to the correlated channels with Doppler frequency f_d. This policy generates a new effective-capacity functional, which is approximately a left-shifted version of the original one for the block fading channel, with a difference of $10\log_{10}(\kappa)$ dB along θ-axis. Thus, given the optimal power-adaptation policy $\mu_{\text{opt}}(\theta, \gamma)$ for block fading channel, the optimal power-adaptation policy for correlated channel is approximately $\mu_{\text{opt}}(\kappa\theta, \gamma)$. Note that $\kappa > 1$ due to $f_d < f_d^*$ in (3.29), as θ increases, the policy $\mu_{\text{opt}}(\kappa\theta, \gamma)$ makes the power-control policy converge faster to the total channel inversion than the case under the block fading channel model. The higher the correlation is, the faster the power-control policy converges to the total channel inversion. Specifically, for our FSMC-based channel model with Doppler frequency f_d, the policy of choosing the boundary points $\{\Gamma_n\}_{n=1}^{N-1}$ becomes:

$$\Gamma_n = M_n^{\kappa\beta+1}\gamma_K^*(\kappa\theta) \tag{3.30}$$

where $\gamma_K^*(\kappa\theta)$ denotes the cutoff threshold obtained by (3.18) at the QoS exponent of $\kappa\theta$.

Proposition 4 plays an important role in deriving the power-control policy for

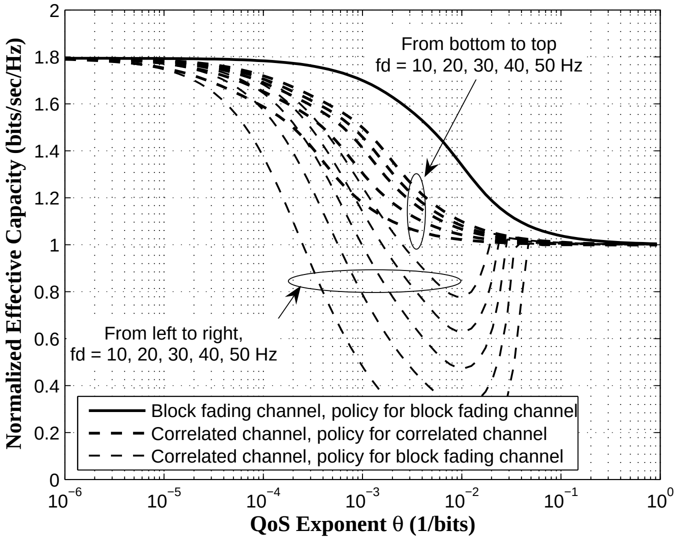


Fig. 17. The effective capacity with different power and rate control policies. The average SNR $\overline{\gamma} = 10$ dB, the fading parameter $m = 2$, BER$= 10^{-3}$, $f_d^{\text{FSMC}} = 300$ Hz, and the number of modes for adaptive MQAM $N = 5$.

the correlated channel. Applying Proposition 4, we can simply shift the existing optimal power-control policy for $10\log_{10}(\kappa)$ dB to obtain the new policies. However, since (3.28) given in Proposition 4 is only an approximation result, our obtained new power and rate adaptation policy is just a *near*-optimal solution.

Fig. 17 shows the normalized effective capacities of both block fading channel and FSMC-based correlated channel under different power-adaptation policies. The optimal policy $\mu_{\text{opt}}(\theta, \gamma)$ for the block fading channel is derived from Section D, which generates the highest effective-capacity curve as shown by the solid line in Fig. 17. Then, according to the analyses in this section, we apply the policy $\mu_{\text{opt}}(\kappa\theta, \gamma)$ to the correlated channel, which numerically generates a group of effective-capacity curves as shown by a set of dashed lines in Fig. 17. These dashed lines are virtually "parallel to" the solid line of the original block fading channel effective capacity. This is consistent with *Remarks* on Proposition 4. However, if we apply the block fading channel policy

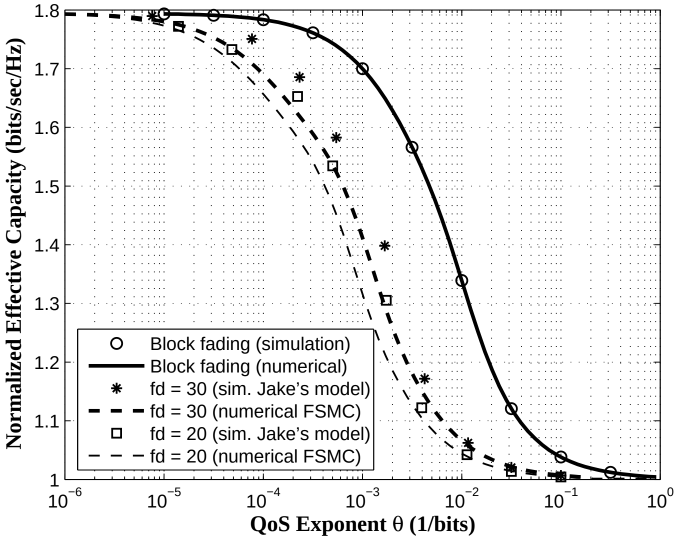


Fig. 18. The effective capacity comparisons with optimal power and rate adaptations between the FSMC-based process (numerical results) and the Jake's model (simulations). The average SNR $\overline{\gamma} = 10$ dB, the fading parameter $m = 2$, BER$= 10^{-3}$, the number of modes for adaptive MQAM is $N = 5$, the Doppler frequencies $f_d^{\text{FSMC}} = 300$ Hz, and $f_d^{\text{Jake}} = 191.25$ Hz.

$\mu_{\text{opt}}(\theta, \gamma)$ directly to the correlated channel, the resulting effective capacities decrease significantly, as shown by a group of dotted lines in Fig. 17.

F. Simulation Results

Using FSMC-based channel model, we obtain the analytical expression for the effective-capacity and the near-optimal power and rate adaptation policies. However, it is important to verify that the policy derived from the FSMC model can also be applied to the more general scenarios, e.g., the Jake's model, without losing the performance satisfactions. Thus, in this section we simulate the Jake's channel process and compare its outcomes with the analytical results obtained in previous sections.

Applying the optimal power and rate adaptation policy given by (3.30), Fig. 18

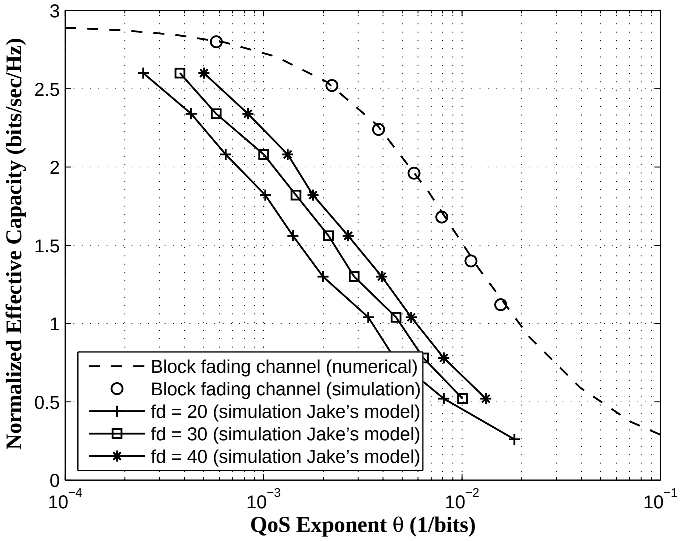


Fig. 19. The effective capacity of constant power approach. The average SNR $\overline{\gamma} = 10$ dB, the fading parameter $m = 1$ (Rayleigh channel).

shows the normalized effective capacity comparisons between the Jake's channel model and the FSMC-based channel model. We can observe from Fig. 18 that for the block fading channel, the simulation results perfectly match with the numerical results. On the other hand, when considering the channel correlation, the outcomes from simulations and numerical solutions for these two models share the same trends but differ very slightly. Such a difference can be explained as follows. As stated in Section E, due to the inaccuracy of the channel model, the Jake's model and the FSMC model have different f_d^*'s, where $f_d^{\text{FSMC}} \approx 300$ Hz and $f_d^{\text{Jake}} = 191.25$ Hz, respectively. Thus, for the same Doppler frequency f_d, the resulting κ in (3.29) is different. Theoretically, the difference of the effective-capacity curves between these two models is $10\log_{10}\left(f_d^{\text{FSMC}}/f_d^{\text{Jake}}\right) = 1.96$ dB, which is consistent with the effective capacity difference observed in Fig. 18.

The above analyses described in Fig. 18 verify that the power-adaptation policy derived by using the FSMC model can be well applied to the general Jake's channel model, where the system employs the *discrete rate MQAM*. In the following, we further show that the policy can also be applied to the general Jake's channel model where the system uses *continuous rate transmission*. Fig. 19 plots the normalized effective capacity of the constant-power approach, where we assume that the Shannon capacity can be achieved for each channel realization. For block fading channel, the simulated effective capacity agrees well with the analytical results. On the other hand, for the correlated fading channel, as the Doppler frequency f_d increases, the effective capacity also increases, with the resulting effective-capacity curves roughly "parallel to" each other. This observation implies that the effective capacity of general channel process also follows the similar trends as described in Proposition 4. Therefore, for continuous rate transmissions, the near-optimal power and rate adaptation law also has the form similar to $\mu_{\text{opt}}(\kappa\theta, \gamma)$ for a certain coefficient κ, where $\mu_{\text{opt}}(\theta, \gamma)$ is given by (3.5).

G. Summary

In this chapter, we proposed and analyzed the QoS-driven power and rate control policies by applying the concept of effective capacity. Our analyses in block fading channel identified the key fact that there exists a fundamental tradeoff between spectral efficiency and QoS provisioning. Depending on the specific QoS requirements, the optimal power-adaptation policy dynamically changes between water-filling and channel inversion. For the more practical adaptive MQAM modulation-based systems, we also developed the corresponding optimal power and rate adaptation scheme. When taking the channel correlation into consideration, we proposed the simple, but efficient, power-control scheme for Markov modeled fading channels. The simulation results verified that such an approach can also be applied to the more general channel models.

As a natural extension of single channel communications, in the next chapter, we will focus on resource allocation problem for multichannel communication systems.

CHAPTER IV

RESOURCE ALLOCATION: MULTIPLE CHANNELS

A. Introduction

The increasing demand for wireless network services such as wireless Internet accessing, mobile computing, and cellular telephoning motivates an unprecedented revolution in wireless broadband communications [71]. This also imposes great challenges in designing the wireless networks since the time-varying fading channel has a significant impact on supporting diverse *quality-of-service* (QoS) requirements for heterogeneous mobile users. In response to these challenges, a great deal of research has been devoted to the techniques that can enhance the spectral-efficiency of the wireless communications systems [72]. The framework used to evaluate these techniques is mainly based on information theory, using the concept of Shannon capacity [6, 7, 73]. While this framework is suitable for an analysis of maximizing the system throughput, it may overlook the mobile users' QoS requirements, since Shannon theory does not place any restrictions on the complexity and delay [64]. Consequently, to provide QoS guarantees for diverse mobile users, it is necessary to take the QoS metrics into account when applying the prevalent information theory to mobile wireless network designs.

In Chapter III, we proposed a QoS-driven power and rate adaptation scheme for *single-input-single-output* (SISO) systems over flat-fading channels. The proposed scheme aims at maximizing the system throughput subject to a given delay-QoS constraint. Specifically, by integrating information theory with the concept of *effective capacity* [20–23], we convert the original problem to the one with the target at maximizing the effective capacity, by which the delay-QoS constraint is characterized by the QoS exponent θ. Using the effective capacity, a smaller θ corresponds to a looser

QoS guarantee, while a larger θ implies a more stringent QoS requirement. In the limiting case, when $\theta \to 0$, the system can tolerate an arbitrarily long delay, which is the scenario studied in information theory. On the other hand, when $\theta \to \infty$, the system cannot tolerate any delay, which corresponds to an extremely stringent delay-bound. In Chapter III, we derived the optimal power-control policy which is adaptive to the QoS exponent θ. The results obtained in Chapter III show that when the QoS constraint becomes loose ($\theta \to 0$), the optimal power-control policy converges to the well-known *water-filling* scheme [7, 64], where the Shannon (or ergodic) capacity can be achieved. In contrast, when the QoS constraint gets stringent ($\theta \to \infty$), the optimal policy converges to the *total channel inversion* scheme [9, 64] under which the system operates at a constant rate. Our analyses also demonstrate that there exists a *fundamental tradeoff* between the throughput and the QoS provisioning. For instance, over a flat-fading Rayleigh channel, the SISO system cannot support stringent delay QoS ($\theta \to \infty$), no matter how much power and spectral bandwidth resources are assigned for the transmission.

As the sequel of Chapter III, in this chapter we focus on QoS provisioning for *multichannel* communications over wireless networks. The motivation of this chapter is mainly based on recent advances in physical layer developments, where a large number of promising schemes can be considered as utilizing multichannels to enhance the system performance. The multichannel communications architecture discussed in this chapter is in a broad sense, which models either multiple diversity branches for *diversity combining* or a number of parallel subchannels for *multiplexing* [75]. Examples of diversity-based systems include code-division-multiple-access (CDMA) RAKE receivers which take the advantage of frequency diversity [76, 77] and multiple input multiple output (MIMO) diversity systems which utilize spatial diversity [36]. On the other hand, examples of multiplexing-based systems include multicarrier systems employing orthogonal-frequency-division-multiplexing (OFDM) mechanism [78] and MIMO multiplexing systems [79].

In this chapter, we show that multichannel transmission can significantly improve the delay-QoS provisioning for wireless communications. In particular, when the QoS constraint is loose ($\theta \to 0$), the optimal power-control policy also converges to the water-filling scheme that achieves the Shannon (ergodic) capacity. By contrast, when the QoS constraint is stringent ($\theta \to \infty$), the optimal policy converges to a scheme which operates at a constant rate (the zero-outage capacity), where an important observation is that, by using only a limited number of subchannels, the above resulting constant rate (the zero-outage capacity) is close to the Shannon capacity. This implies that the optimal effective capacity function connects the ergodic capacity and the zero-outage capacity as the QoS constraint varies. Furthermore, unlike the single channel transmission scheme which has to tradeoff the throughput for QoS provisioning, the multichannel transmission scheme can achieve both high throughput and stringent QoS at the same time. For instance, our simulation results show that over the Rayleigh fading channel, a multicarrier system with 64 independent subchannels can achieve more than 99% of the Shannon capacity, while still guaranteeing a constant rate transmission, as if the transmission was over a wireline network. The above observation demonstrates from another perspective that the zero-outage capacity approaches the ergodic capacity as the number of parallel subchannels increases [11].

The rest of the chapter is organized as follows. Section B describes our general multichannel wireless system model. Sections C derives the optimal power adaptation for diversity-based systems. Section D formulates the optimization problem for the multiplexing-based systems, and Section E develops the corresponding optimal solutions. Section F further investigates the special cases of our proposed optimal

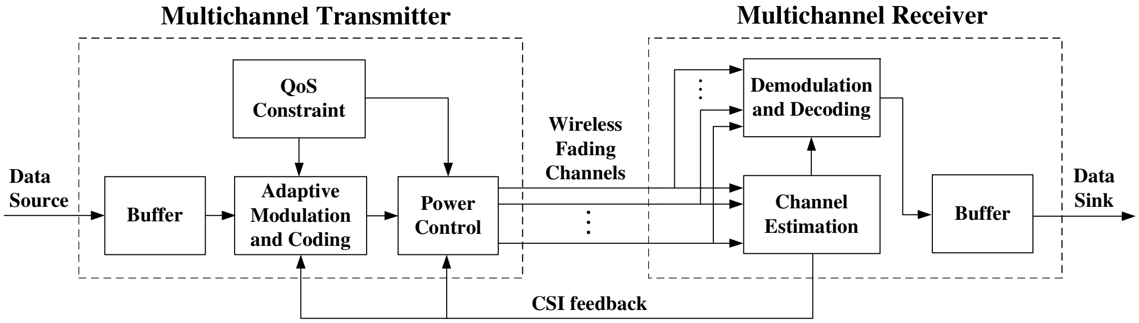


Fig. 20. The multichannel system model.

power-control scheme. Section G conducts simulations to evaluate the performance of our proposed scheme. The chapter concludes with Section H.

B. System Model

The general multichannel system model over a wireless link is shown in Fig. 20. We concentrate on a discrete-time system with a point-to-point link between the transmitter and the receiver in mobile wireless networks. Let us denote the system total spectral-bandwidth by B and the mean transmit power by $\overline{P}$, respectively. The power spectral density (PSD) of the complex additive white Gaussian noise (AWGN) is denoted by $N_0/2$ per dimension. We assume that AWGN is independent identically distributed (i.i.d.) on each subchannel. Unless otherwise stated, throughout this chapter we use "subchannels" to represent either diversity or multiplexing branches in our multichannel system model.

First, the upper-layer packets are divided into *frames* at the datalink layer, which forms the "data source" as shown in Fig. 20. The frame duration is denoted by T_f, which is assumed to be less than the fading coherence time, but sufficiently long so that the information-theoretic assumption of infinite code-block length is meaningful [80]. The frames are stored at the transmit buffer and split into bit streams at the physical layer. Then, based on the QoS constraint and channel-state information (CSI) fed back from the receiver, adaptive modulation and coding (AMC), as well as power control are applied, respectively, at the transmitter. Depending on the specific transmission mechanism, the bit streams are transmitted through N subchannels to the receiver. The reverse operations are executed at the receiver side. Finally, the frames are recovered at the "data sink" for further processing. We also make the following two assumptions.

A1: The discrete-time channel is assumed to be block fading. The path gains are invariant within a frame's time duration T_f, but vary independently from one frame to another. Making such an assumption is mainly based on the following reasons. First, the effective capacity expression in a block fading channel (2.15) only depends on marginal statistics of a service process, which is much simpler than the general expression given by (2.14), where higher order statistics of a service process are required. Second, more importantly, through the study of Chapter III, we observe that there exists a simple and efficient approach to convert the power adaptation policy obtained in block-fading channels to that over correlated-fading channels, making the investigation of power adaptation in block-fading channels more applicable.

A2: We further assume that given the transmit power, the specific multichannel transmission scheme, and an instantaneous channel gain, the AMC scheme can achieve the Shannon capacity. Based on the above two assumptions, for each given power-control policy, the resulting effective capacity reaches its maximum for all modulation/coding schemes and all channel realizations.

The wireless channel may be modeled as being frequency-selective (e.g., in the context of multicarrier OFDM system or CDMA RAKE receiver-based system), but each subchannel experiences the flat-fading. Denote the nth subchannel envelope process by $\{\alpha_n[i], n \in \mathcal{N}_0, i = 1, 2, ...\}$, where $\mathcal{N}_0 = \{1, 2, ..., N\}$ represents the index-set of the subchannels and $i \in \{1, 2, ...\}$ is time-index of the frame. Let $\lambda_n[i] \triangleq \alpha_n^2[i]$ denote the path-gain process. Then, the joint probability density function (pdf) of the path-gains $\boldsymbol{\lambda}[i] \triangleq (\lambda_1[i], \lambda_2[i], ..., \lambda_N[i])$ can be expressed as $p_{\boldsymbol{\Lambda}}(\boldsymbol{\lambda}) = p_{\Lambda_1,\Lambda_2,...,\Lambda_N}(\lambda_1, \lambda_2, ..., \lambda_N)$. Although the scheme discussed in this chapter can be applied to any channel distribution models, we just assume the Rayleigh channel model for convenience.

Throughout this chapter, we also assume that the CSI is perfectly estimated at the receiver and reliably fed back to the transmitter without delay. Moreover, the datalink-layer buffer size is assumed to be infinite. In the following discussions, since the block-fading channel process is stationary and ergodic, its instantaneous-time marginal statistics is independent of the time-index i, and thus we may omit the time-index i for simplicity.

C. Diversity Systems

We first focus on the QoS-driven power adaptation for *diversity-based systems*. The key idea of diversity-based systems is to transmit multiple copies of the same data through different subchannels. At the receiver side, the multiple copies are combined together such that the transmission reliability can be enhanced.

For diversity combining systems, the system performance is determined by the combined signal-to-noise ratio (SNR) at the receiver. If we assume that no power control is used, then the SNR at the receiver combiner can be denoted by $\gamma[i]$, which depends not only on the instantaneous channel condition, but also on the specific diversity scheme used. For instance, in a maximal-ratio combining (MRC) system with N diversity branches [34], the SNR at the output of the combiner can be expressed as $\gamma[i] = \sum_{n=1}^{N} \overline{P}\lambda_n[i]/(N_0B)$, with its mean $\overline{\gamma} = \overline{P}\sum_{n=1}^{N} \mathbb{E}\{\lambda_n[i]\}/(N_0B)$, where

$\mathbb{E}\{\cdot\}$ denotes the expectation. On the other hand, in a selection combining (SC) system [34], the SNR at the output of the combiner is given by $\gamma[i] = \overline{P}\lambda_{\max}[i]/(N_0B)$, with the mean $\overline{\gamma} = \overline{P}\mathbb{E}\{\lambda_{\max}[i]\}/(N_0B)$, where $\lambda_{\max}[i] = \max\{\lambda_n[i], n \in \mathcal{N}_0\}$. Let the pdf of $\gamma[i]$ be denoted by $p_\Gamma(\gamma)$. It is well known that there have been a great deal of research efforts in deriving the analytical expressions of the pdf $p_\Gamma(\gamma)$ under different channel conditions and different diversity combining techniques.

Using diversity combining, the original vector channel (i.e., multichannel) transmission problem is converted into a scalar channel (i.e., single channel) transmission problem. Therefore, the scheme discussed in Chapter III can be directly applied to obtain the optimal power-adaptation policy, where the only difference is at the pdf $p_\Gamma(\gamma)$ of the SNR at the combiner output. Specifically, the optimal policy, denoted by $\mu_{\text{opt}}(\theta, \gamma)$, can be expressed similar as (3.5):

$$\mu_{\text{opt}}(\theta, \gamma) = \begin{cases} \dfrac{1}{\gamma_0^{\frac{1}{\beta+1}} \gamma^{\frac{\beta}{\beta+1}}} - \dfrac{1}{\gamma}, & \gamma \geq \gamma_0 \\ 0, & \gamma < \gamma_0 \end{cases} \tag{4.1}$$

where we define

$$\beta \triangleq \frac{\theta T_f B}{\log 2} \tag{4.2}$$

as the normalized QoS exponent and γ_0 as the cutoff SNR threshold, which can be numerically obtained by meeting the following mean power constraint:

$$\int_{\gamma_0}^{\infty} \left(\frac{1}{\gamma_0^{\frac{1}{\beta+1}} \gamma^{\frac{\beta}{\beta+1}}} - \frac{1}{\gamma} \right) p_\Gamma(\gamma) d\gamma = 1. \tag{4.3}$$

Note that the threshold $\gamma_0 = \gamma_0(\theta, p_\Gamma(\gamma))$ depends not only on the fading distribution $p_\Gamma(\gamma)$, but also on the QoS exponent θ. Similar to the conclusion obtained in Chapter III, we can observe that when the QoS constraint is loose ($\theta \to 0$), the optimal power-control law converges to the water-filling scheme, where the Shannon capacity can be achieved. On the other hand, when the QoS constraint is stringent ($\theta \to \infty$), the optimal power-control law converges to the total channel inversion scheme[1] such that the system operates at a constant service rate.

Once obtaining γ_0, we can derive the optimal effective capacity, denoted by $\mathrm{E}_{\mathrm{C}}^{\mathrm{opt}}(\theta)$ as:

$$\mathrm{E}_{\mathrm{C}}^{\mathrm{opt}}(\theta) = -\frac{1}{\theta}\log\left(\int_{0}^{\gamma_0} p_{\Gamma}(\gamma)d\gamma + \int_{\gamma_0}^{\infty}\left(\frac{\gamma}{\gamma_0}\right)^{-\frac{\beta}{(\beta+1)}} p_{\Gamma}(\gamma)d\gamma\right). \tag{4.4}$$

Given the specific diversity scheme and channel statistics, the optimal effective capacity given in (4.4) can be calculated either by the closed-form expression or by numerical solution.

D. Multiplexing Systems: Problem Formulation

In the following, we consider QoS-driven power adaptation for *multiplexing-based systems*. The core idea of multiplexing systems is to transmit different data streams through different subchannels. At the receiver side, the parallel data steams are recovered separately. This transmission strategy can either combat the frequency selective fading channel (e.g., in multicarrier systems) or increase the throughput (e.g., in MIMO multiplexing systems), which are elaborated on, respectively, as follows.

1. Multicarrier Systems

Consider a multicarrier system with N subchannels corresponding to N subcarriers. If we assume a constant equal-power distribution among all subcarriers, then the

[1] In this scheme, the transmission power is proportional to the reciprocal of the channel power gain.

instantaneous transmit power for the nth subchannel at the ith frame, denoted by $P_n[i]$, is equal to $P_n[i] = \overline{P}/N$ for all n and i. The corresponding instantaneous received SNR, denoted by $\gamma_n[i]$, can be expressed as

$$\gamma_n[i] = \frac{\lambda_n[i](\overline{P}/N)}{N_0(B/N)} = \frac{\overline{P}\lambda_n[i]}{N_0 B}, \text{ for } n \in \mathcal{N}_0. \tag{4.5}$$

Denote the joint pdf of the SNR vector $\boldsymbol{\gamma}[i] = (\gamma_1[i], \gamma_2[i], ..., \gamma_N[i])$ for all subchannels by $p_{\boldsymbol{\Gamma}}(\boldsymbol{\gamma}) = p_{\Gamma_1,\Gamma_2,...,\Gamma_N}(\gamma_1, \gamma_2, ..., \gamma_N)$, and the corresponding power-adaptation policy for the nth subchannel by $\mu_n(\theta, \boldsymbol{\gamma}[i])$, respectively. Then, the instantaneous transmit power for the nth subchannel becomes $P_n[i] = \mu_n(\theta, \boldsymbol{\gamma}[i])\,\overline{P}/N$. Note that we limit the mean transmit power by $\overline{P}$. Therefore, the power-control policy needs to satisfy the mean power constraint as follows:

$$\sum_{n=1}^{N} \underbrace{\int_0^\infty \cdots \int_0^\infty}_{N-\text{fold}} \mu_n(\theta, \boldsymbol{\gamma}) p_{\boldsymbol{\Gamma}}(\boldsymbol{\gamma}) d\gamma_1 \cdots d\gamma_N = N \tag{4.6}$$

where

$$\mu_n(\theta, \boldsymbol{\gamma}) \geq 0, \text{ for all } n \in \mathcal{N}_0. \tag{4.7}$$

Recall that we assume that the AMC scheme can achieve the Shannon capacity. Thus, the instantaneous service rate of the frame i, denoted by $R[i]$, can be expressed as

$$R[i] = \sum_{n=1}^{N} \left(\frac{T_f B}{N}\right) \log_2\bigg(1 + \mu_n(\theta, \boldsymbol{\gamma}[i])\,\gamma_n[i]\bigg). \tag{4.8}$$

Thus, from (2.15), the effective capacity, denoted by $\mathrm{E_C}(\theta)$, can be expressed as follows:

$$\begin{aligned}\mathrm{E_C}(\theta) &= -\frac{1}{\theta}\log\left(\mathbb{E}\left\{e^{-\theta R[i]}\right\}\right)\\ &= -\frac{1}{\theta}\log\Bigg(\underbrace{\int_0^\infty\!\!\cdots\int_0^\infty}_{N-\mathrm{fold}}\prod_{n=1}^{N}\left[1+\mu_n(\theta,\boldsymbol{\gamma})\gamma_n\right]^{-\frac{\beta}{N}}p_{\boldsymbol{\Gamma}}(\boldsymbol{\gamma})d\gamma_1\cdots d\gamma_N\Bigg) \qquad (4.9)\end{aligned}$$

where β is also given by (4.2). To maximize the effective capacity, we can formulate an optimization problem as follows:

$$\mathrm{E_C^{opt}}(\theta) = \max_{\mu_n(\theta,\boldsymbol{\gamma}),n\in\mathcal{N}_0}\left\{-\frac{1}{\theta}\log\Bigg(\underbrace{\int_0^\infty\!\!\cdots\int_0^\infty}_{N-\mathrm{fold}}\prod_{n=1}^{N}\left[1+\mu_n(\theta,\boldsymbol{\gamma})\gamma_n\right]^{-\frac{\beta}{N}}p_{\boldsymbol{\Gamma}}(\boldsymbol{\gamma})d\gamma_1\cdots d\gamma_N\Bigg)\right\} \qquad (4.10)$$

subject to constraints given by (4.6) and (4.7).

2. MIMO Systems

Let N_t and N_r denote the number of transmit and receive antennas, respectively, and let $\mathbb{C}$ denote the space of complex numbers. Then, the MIMO multiplexing-based transmission can be expressed as $\mathbf{y}[i] = \mathbf{H}[i]\mathbf{x}[i] + \mathbf{n}[i]$, where $\mathbf{y}[i] \in \mathbb{C}^{N_r}$ denotes the received signal, $\mathbf{H}[i] \in \mathbb{C}^{N_r\times N_t}$ represents the complex channel matrix, $\mathbf{x}[i] \in \mathbb{C}^{N_t}$ stands for the input signal, and $\mathbf{n}[i] \in \mathbb{C}^{N_r}$ is the complex AWGN, where, without loss of generality, we assume $\mathbb{E}\{\mathbf{n}[i]\mathbf{n}[i]^\dagger\} = \mathbf{I}_{N_r}$, with † denoting the conjugate transpose. It is well known that for MIMO multiplexing systems, the data streams are equivalent to transmitting through N parallel singular-value channels [8], where $N = \min\{N_t, N_r\}$. Mathematically, the transmitted signals can be modeled as [8]

$$\widetilde{y}_\ell[i] = \sqrt{\lambda_\ell[i]}\,\widetilde{x}_\ell[i] + \widetilde{n}_\ell[i], \text{ for all } \ell \in \mathcal{N}_0 \qquad (4.11)$$

where $\{\sqrt{\lambda_\ell[i]}\}_{\ell=1}^N$ are nonzero singular values of the channel matrix $\mathbf{H}[i]$. Corresponding to our system description in Section B, $\{\sqrt{\lambda_\ell[i]}\}_{\ell=1}^N$ and $\{\lambda_\ell[i]\}_{\ell=1}^N$ can be considered as the *virtual* envelope process and path-gain process for MIMO multiplexing system, respectively. There have been abundant literatures investigating the joint pdf $p_{\mathbf{\Lambda}}(\boldsymbol{\lambda})$ for $\boldsymbol{\lambda}[i] = (\lambda_1[i], \lambda_2[i], ..., \lambda_N[i])$. For instance, when the channels between all transmit and receive antenna pairs are i.i.d. Rayleigh distributed with unit energy, the pdf $p_{\mathbf{\Lambda}}(\boldsymbol{\lambda})$ follows the well-known Wishart distribution as [8]

$$p_{\mathbf{\Lambda}}(\boldsymbol{\lambda}) = \left[N!\left(\prod_{i=1}^{N}(N-i)!(M-i)!\right)\right]^{-1} \exp\left(-\sum_{i=1}^{N}\lambda_i\right)\prod_{i=1}^{N}\lambda_i^{M-N}\prod_{1\le i<j\le N}(\lambda_i-\lambda_j)^2 \tag{4.12}$$

where $M = \max\{N_t, N_r\}$. Using (2.15), we also derive the effective capacity $\mathrm{E_C}(\theta)$ for MIMO multiplexing system as follows:

$$\begin{aligned}\mathrm{E_C}(\theta) &= -\frac{1}{\theta}\log\left(\mathbb{E}\left\{e^{-\theta R[i]}\right\}\right)\\ &= -\frac{1}{\theta}\log\left(\underbrace{\int_0^\infty\cdots\int_0^\infty}_{N-\text{fold}}\prod_{\ell=1}^{N}\left[1+\mu_\ell(\theta,\boldsymbol{\lambda})\lambda_\ell\right]^{-\beta}p_{\mathbf{\Lambda}}(\boldsymbol{\lambda})d\lambda_1\cdots d\lambda_N\right)\end{aligned} \tag{4.13}$$

where $\mu_\ell(\theta, \boldsymbol{\lambda})$ denotes the power-adaptation policy and β is also given by (4.2). To maximize the effective capacity, we formulate the optimization problem as follows:

$$\mathrm{E_C^{opt}}(\theta) = \max_{\mu_\ell(\theta,\boldsymbol{\lambda}),\ell\in\mathcal{N}_0}\left\{-\frac{1}{\theta}\log\left(\underbrace{\int_0^\infty\cdots\int_0^\infty}_{N-\text{fold}}\prod_{\ell=1}^{N}\left[1+\mu_\ell(\theta,\boldsymbol{\lambda})\lambda_\ell\right]^{-\beta}p_{\mathbf{\Lambda}}(\boldsymbol{\lambda})d\lambda_1\cdots d\lambda_N\right)\right\} \tag{4.14}$$

subject to the mean power constraint:

$$\sum_{\ell=1}^{N}\underbrace{\int_0^\infty\cdots\int_0^\infty}_{N-\text{fold}}\mu_\ell(\theta,\boldsymbol{\lambda})p_{\mathbf{\Lambda}}(\boldsymbol{\lambda})d\lambda_1\cdots d\lambda_N = \overline{P} \tag{4.15}$$

and also the constraint:

$$\mu_\ell(\theta, \boldsymbol{\lambda}) \geq 0, \text{ for all } \ell \in \mathcal{N}_0. \tag{4.16}$$

Comparing (4.10) with (4.14), we can observe that the two optimization problems have the same structure except for certain constant-scalar differences. Therefore, we can develop a unified approach to derive the optimal power-adaptation policy. To simplify the presentation, in the next section, we will mainly focus on multicarrier systems. The detailed derivations for MIMO multiplexing systems are similar to those of the multicarrier systems, but omitted in this chapter for lack of space.

3. Independent Optimization

Before getting into details of maximizing the effective capacity expressed in (4.10) and (4.14), we first consider an alternative strategy, namely, the independent optimization approach for the following reasons. Since we already obtain the optimal power-adaptation policy for the single channel transmission in Chapter III, can we directly apply this strategy to multiplexing systems? For instance, in a multiplexing system with N i.i.d. subchannels, one possible solution is to maximize the effective capacity at each subchannel *independently* using the optimal single channel power-adaptation policy. Is this resulting scheme optimal?

Surprisingly, the answer to the above questions is *no*. In fact, this independent optimization approach turns out to be optimal in maximizing the Shannon capacity (e.g., water-filling power control for multichannel transmissions). Note that when $\theta \to 0$, the maximum effective capacity approaches the Shannon capacity. Therefore, this strategy can maximize the effective capacity as $\theta \to 0$. However, as will be shown in the following sections, the independent optimization approach is *not* the optimal policy to maximize the effective capacity for an arbitrary θ.

To characterize the performance of independent power adaptation over i.i.d. subchannels, we have the following proposition.

Proposition 5. *Under the same power and spectral-bandwidth constraints,* ***if*** *we apply an arbitrary power-adaptation policy to a single channel transmission system, and apply the same power-adaptation policy to each of N i.i.d. subchannels of a multichannel transmission system* independently*,* ***then*** *the resulting effective capacities, denoted by ${\rm E_C}^{(1)}(\theta)$ for the single channel system and ${\rm E_C}^{(N)}(\theta)$ for the multichannel system, respectively, satisfy ${\rm E_C}^{(N)}(\theta) = {\rm E_C}^{(1)}\left(\theta/N\right)$.*

Proof. Denote the service rate of the nth subchannel of multichannel transmission by $R_n[i]$ and the service rate of single channel transmission by $R[i]$, respectively. When the channel condition is the same, we have $R_n[i] = R[i]/N$, $\forall n$ and $\forall i$, which is because the nth subchannel only occupies $1/N$ of the total spectral-bandwidth. Then, the following equations hold:

$$\begin{aligned} {\rm E_C}^{(N)}(\theta) &= -\frac{1}{\theta}\log\left(\mathbb{E}\left\{e^{-\theta\sum_{n=1}^{N}R_n[i]}\right\}\right) \\ &= -\frac{1}{\theta}\log\left(\mathbb{E}\left\{e^{-\theta R_n[i]}\right\}\right)^N \\ &= -\frac{1}{\left(\frac{\theta}{N}\right)}\log\left(\mathbb{E}\left\{e^{-\left(\frac{\theta}{N}\right)R[i]}\right\}\right) \\ &= {\rm E_C}^{(1)}\left(\frac{\theta}{N}\right). \end{aligned} \tag{4.17}$$

Thus, the proof follows. □

Remark 3. Proposition 5 says that as compared to the single channel transmission, the effective capacity gain of the multichannel transmission using the independent power control is $10\log_{10}N$ dB. In other words, ${\rm E}_{\rm C}^{(N)}(\theta)$ is a right-shifted version of ${\rm E}_{\rm C}^{(1)}(\theta)$ along θ-axis using the logarithmic scale, where the difference between these two is $10\log_{10}N$ dB. Over the single channel Rayleigh fading environment, we prove in

Chapter III that the effective capacity always approaches zero as $\theta \to \infty$. Therefore, according to Proposition 5, by using the independent power-adaptation policies, as long as the number of subchannels N is finite, the effective capacity $\mathrm{E_C}^{(N)}(\theta)$ also approaches zero as $\theta \to \infty$. In the following sections, we propose a joint optimization approach, which performs much better than the independent optimization.

E. Multiplexing Systems: Optimal Allocation Policy

Since $\log(\cdot)$ is a monotonically increasing function, for each given $\theta > 0$, the original maximization problem of (4.10) is equivalent to the following minimization problem:

$$\min_{\mu_n(\theta,\boldsymbol{\gamma}),n\in\mathcal{N}_0}\left\{\underbrace{\int_0^\infty\cdots\int_0^\infty}_{N-\text{fold}}\prod_{n=1}^{N}\left[1+\mu_n(\theta,\boldsymbol{\gamma})\gamma_n\right]^{-\frac{\beta}{N}}p_{\mathbf{\Gamma}}(\boldsymbol{\gamma})d\gamma_1\cdots d\gamma_N\right\} \tag{4.18}$$

which is subject to the same set of constraints given by (4.6) and (4.7). As derived in Appendix F, we prove that the objective function in (4.18) is strictly convex on the space spanned by $\big(\mu_1(\theta,\boldsymbol{\gamma}),...,\mu_N(\theta,\boldsymbol{\gamma})\big)$. In addition, it is clear that the constraints given by (4.6) and (4.7) are linear with respect to $\big(\mu_1(\theta,\boldsymbol{\gamma}),...,\mu_N(\theta,\boldsymbol{\gamma})\big)$. Therefore, the problem can be considered as a convex optimization problem which has the unique optimal solution. Then, using standard optimization technique, we can construct the Lagrangian function as follows:

$$\begin{aligned}\mathcal{J} = &\underbrace{\int_0^\infty\cdots\int_0^\infty}_{N-\text{fold}}\prod_{n=1}^{N}\left[1+\mu_n(\theta,\boldsymbol{\gamma})\gamma_n\right]^{-\frac{\beta}{N}}p_{\mathbf{\Gamma}}(\boldsymbol{\gamma})d\gamma_1\cdots d\gamma_N\\ &+\kappa_0\left\{\sum_{n=1}^{N}\underbrace{\int_0^\infty\cdots\int_0^\infty}_{N-\text{fold}}\mu_n(\theta,\boldsymbol{\gamma})p_{\mathbf{\Gamma}}(\boldsymbol{\gamma})d\gamma_1\cdots d\gamma_N-N\right\}-\sum_{n=1}^{N}\kappa_n\mu_n(\theta,\boldsymbol{\gamma})\end{aligned} \tag{4.19}$$

where all the Lagrangian multipliers $\{\kappa_n\}_{n=0}^{N}$ satisfy $\kappa_n \geq 0$. Differentiating the Lagrangian function and setting the derivative equal to zero [81, Sec. 4.2.4], we obtain a set of N equations:

$$\begin{aligned}\frac{\partial \mathcal{J}}{\partial \mu_n(\theta,\boldsymbol{\gamma})} &= -\frac{\beta\gamma_n}{N}\left[1+\mu_n(\theta,\boldsymbol{\gamma})\gamma_n\right]^{-\frac{\beta}{N}-1}\prod_{i\in\mathcal{N}_0,\, i\neq n}\left[1+\mu_i(\theta,\boldsymbol{\gamma})\gamma_i\right]^{-\frac{\beta}{N}} p_{\mathbf{\Gamma}}(\boldsymbol{\gamma}) \\ &\quad +\kappa_0 p_{\mathbf{\Gamma}}(\boldsymbol{\gamma}) - \kappa_n = 0, \text{ for all } n\in\mathcal{N}_0. \end{aligned} \tag{4.20}$$

According to the concept of *complementary slackness* [82, Sec. 5.5.2], if the *strict* inequality $\mu_j(\theta,\boldsymbol{\gamma}) > 0$ holds for a certain $j\in\mathcal{N}_0$, then the Lagrangian multiplier κ_j corresponding to $\mu_j(\theta,\boldsymbol{\gamma})$ must be equal to zero. Based on this fact, we consider two different scenarios, respectively, as follows.

1. Scenario-1: $\mu_n(\theta,\boldsymbol{\gamma}) > 0$ Holds for All $n\in\mathcal{N}_0$.

Under the conditions of the above Scenario-1, all subchannels are assigned with power for data transmission. Then, according to the complementary slackness, except for κ_0, all the other Lagrangian multipliers $\{\kappa_n\}_{n=1}^{N}$ must be equal to zero. Thus, (4.20) reduces to:

$$\left[1+\mu_n(\theta,\boldsymbol{\gamma})\gamma_n\right]^{-\frac{\beta}{N}-1}\prod_{i\in\mathcal{N}_0,\, i\neq n}\left[1+\mu_i(\theta,\boldsymbol{\gamma})\gamma_i\right]^{-\frac{\beta}{N}} = \frac{N\gamma_0}{\gamma_n}, \text{ for all } n\in\mathcal{N}_0 \tag{4.21}$$

where we define $\gamma_0 \triangleq \kappa_0/\beta$, which is a cutoff threshold to be optimized later. Solving (4.21), we can obtain the optimal power-adaptation policy as follows:

$$\mu_n(\theta,\boldsymbol{\gamma}) = \frac{1}{\gamma_0^{\frac{1}{\beta+1}}\prod_{i\in\mathcal{N}_0}\gamma_i^{\frac{\beta}{N(\beta+1)}}} - \frac{1}{\gamma_n}, \; n\in\mathcal{N}_0. \tag{4.22}$$

Note that the policy given by (4.22) is optimal only if $\mu_n(\theta,\boldsymbol{\gamma}) > 0$ holds for all $n\in\mathcal{N}_0$. Specifically, define $\mathcal{N}_1$ as the index-set of SNRs which satisfy this strict inequality as follows:

$$\mathcal{N}_1 \triangleq \left\{ n\in\mathcal{N}_0 \,\middle|\, \frac{1}{\gamma_0^{\frac{1}{\beta+1}}\prod_{i\in\mathcal{N}_0}\gamma_i^{\frac{\beta}{N(\beta+1)}}} - \frac{1}{\gamma_n} > 0 \right\}. \tag{4.23}$$

Then, (4.22) is the optimal solution only if $\mathcal{N}_1 = \mathcal{N}_0$. Otherwise, if $\mathcal{N}_1 \subset \mathcal{N}_0$, we need to consider the following scenario.

2. Scenario-2: There Exists $\mu_n(\theta, \boldsymbol{\gamma})$ Such That $\mu_n(\theta, \boldsymbol{\gamma}) = 0$.

If $\mathcal{N}_1 \subset \mathcal{N}_0$, there must exist certain $\mu_n(\theta, \boldsymbol{\gamma})$ such that $\mu_n(\theta, \boldsymbol{\gamma}) = 0$. In other words, some subchannels are not assigned with any power. In order to identify the set of subchannels to which the system do not assign power, we introduce the following lemma.

Lemma 1. ***If*** $n \notin \mathcal{N}_1$, ***then*** $\mu_n(\theta, \boldsymbol{\gamma}) = 0$.

Proof. The proof is provided in Appendix G. □

Lemma 1 states that *all the power is assigned to the subchannels which belong to* $\mathcal{N}_1$. Thus, the original minimization problem of (4.18) reduces to

$$\min_{\mu_n(\theta,\boldsymbol{\gamma}),\, n\in\mathcal{N}_1} \left\{ \underbrace{\int_0^\infty \cdots \int_0^\infty}_{N-\text{fold}} \prod_{n\in\mathcal{N}_1} \left[1 + \mu_n(\theta, \boldsymbol{\gamma})\gamma_n\right]^{-\frac{\beta}{N}} p_{\mathbf{\Gamma}}(\boldsymbol{\gamma}) d\gamma_1 \cdots d\gamma_N \right\}. \qquad (4.24)$$

Comparing (4.24) with (4.18), we can observe that the two minimization problems have the same structure except that the optimization space shrinks from $\mathcal{N}_0$ to $\mathcal{N}_1$. The above observation suggests us to solve this minimization problem in a *recursive* manner.

Following the same procedure as that used in Section 1, if the strict inequality $\mu_n(\theta, \boldsymbol{\gamma}) > 0$ holds for all $n \in \mathcal{N}_1$, we can obtain the optimal power-control policy as follows:

$$\mu_n(\theta, \boldsymbol{\gamma}) = \begin{cases} \dfrac{1}{\gamma_0^{\frac{N}{N_1\beta+N}} \prod_{i\in\mathcal{N}_1} \gamma_i^{\frac{\beta}{N_1\beta+N}}} - \dfrac{1}{\gamma_n}, & n \in \mathcal{N}_1 \\ 0, & \text{otherwise} \end{cases}$$

```
$\mathbf{Algorithm:\ QoS - driven\ power\ adaptation}$.

(1) Initialization.
    a) Obtain $\mathcal{N}_1$, and $N_1$ by (4.23) and (4.25), respectively.
    b) $k = 1$.
(2) While ($\mathcal{N}_k \neq \mathcal{N}_{k-1}$) do
    a) $\mathcal{N}_{k+1} = \left\{ n \in \mathcal{N}_k \,\middle|\, \frac{1}{\gamma_0^{\frac{N}{N_k\beta+N}} \prod_{i\in\mathcal{N}_k} \gamma_i^{\frac{\beta}{N_k\beta+N}}} - \frac{1}{\gamma_n} > 0 \right\}$.
    b) $N_{k+1} = |\mathcal{N}_{k+1}|$.
    c) $k = k+1$.
(3) Obtain the optimal adaptation policy.
    a) Denote $\mathcal{N}^* = \mathcal{N}_k$ and $N^* = N_k$, respectively.
    b) $\mu_n(\theta,\boldsymbol{\gamma}) = \begin{cases} \frac{1}{\gamma_0^{\frac{N}{N^*\beta+N}} \prod_{i\in\mathcal{N}^*} \gamma_i^{\frac{\beta}{N^*\beta+N}}} - \frac{1}{\gamma_n}, & n \in \mathcal{N}^* \\ 0, & \text{otherwise.} \end{cases}$
```

Fig. 21. Algorithm of optimal power allocation for multicarrier system.

where N_1 denotes the number of subchannels belonging to $\mathcal{N}_1$, or, the cardinality of $\mathcal{N}_1$, i.e.,

$$N_1 \triangleq |\mathcal{N}_1|. \tag{4.25}$$

Otherwise, if not all of the subchannels $n \in \mathcal{N}_1$ satisfy the strict inequality $\mu_n(\theta,\boldsymbol{\gamma}) > 0$, we need to further divide $\mathcal{N}_1$ and repeat this procedure itself again. In summary, the QoS-driven optimal power-adaptation algorithm is described as the algorithm shown in Fig. 21.

The principle of the optimal power-adaptation algorithm is to search for the maximum set of SNRs which can simultaneously satisfy the strict inequality $\mu_n(\theta,\boldsymbol{\gamma}) > 0$, i.e., the maximum set of subchannels which can be assigned power simultaneously. Once we successfully identify such a set ($\mathcal{N}_k = \mathcal{N}_{k-1} = \mathcal{N}^*$), the optimal power-adaptation policy is obtained. Otherwise, we exclude those undesired SNRs from current optimization space and repeat this searching procedure itself again. If the "while-loop" ends up with $(\mathcal{N}_{k-1} = \mathcal{N}_k = \emptyset)$, then no subchannel can satisfy the strict inequality condition $\mu_n(\theta, \boldsymbol{\gamma}) > 0$. In this case, we set $\mu_n(\theta, \boldsymbol{\gamma}) = 0$ for all $n \in \mathcal{N}_0$. Thus, the system falls into an outage state and cannot send any data. Finally, we obtain the optimal resource allocation policy for multicarrier systems as follows:

$$\mu_n(\theta, \boldsymbol{\gamma}) = \begin{cases} \dfrac{1}{\gamma_0^{\frac{N}{N^*\beta+N}} \prod_{i\in\mathcal{N}^*} \gamma_i^{\frac{\beta}{N^*\beta+N}}} - \dfrac{1}{\gamma_n}, & n \in \mathcal{N}^* \\ 0, & \text{otherwise.} \end{cases} \tag{4.26}$$

Similarly, for MIMO multiplexing system, we can show that the optimal power-adaptation policy can be expressed as

$$\mu_\ell(\theta, \boldsymbol{\lambda}) = \begin{cases} \dfrac{1}{\lambda_0^{\frac{1}{N^*\beta+1}} \prod_{i\in\mathcal{N}^*} \lambda_i^{\frac{\beta}{N^*\beta+1}}} - \dfrac{1}{\lambda_\ell}, & \ell \in \mathcal{N}^* \\ 0, & \text{otherwise} \end{cases} \tag{4.27}$$

where $\mathcal{N}^*$ and N^* can be obtained by a similar algorithm as shown in Fig. 21.

Given the optimal power-adaptation algorithm, the cutoff threshold γ_0 is determined by meeting the mean power constraint (4.6). Note that γ_0 is jointly determined by the QoS exponent θ and channel model distribution $p_{\boldsymbol{\Gamma}}(\boldsymbol{\gamma})$. After obtaining the cutoff threshold, the optimal effective capacity can be calculated by (4.10).

F. Special Cases

1. Two-Subchannel Case ($N = 2$)

To demonstrate the execution procedure of our proposed algorithm, let us consider a particular case when the number of subcarriers $N = 2$. Using the algorithm described in Fig. 21, we can see that the joint optimal power-adaptation policy partitions the

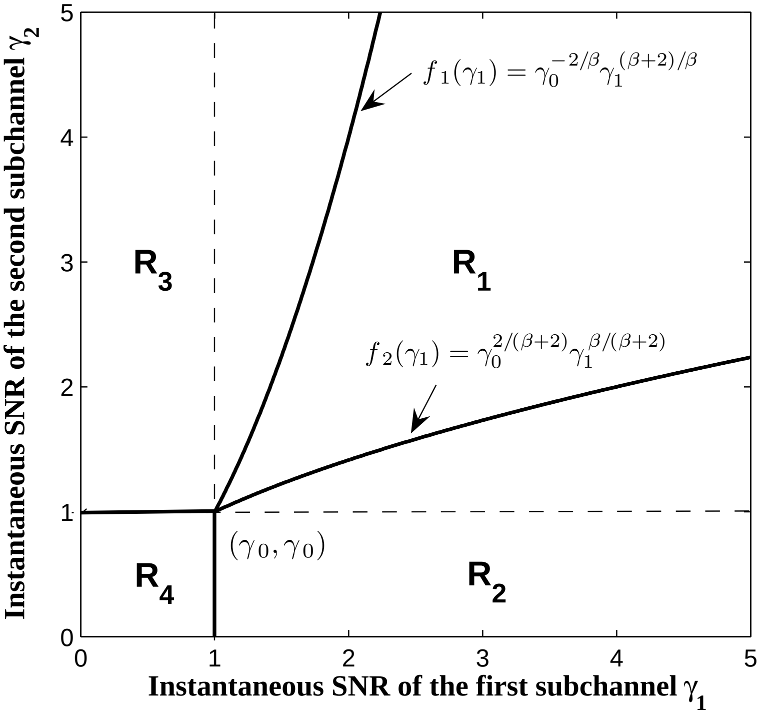


Fig. 22. The policy regions of different power-adaptation strategies when the number of subchannels $N = 2$. The solid lines depicts the result of joint optimization and the dashed lines depicts the result of independent optimization. In this example, we set $\gamma_0 = 1$ and $\beta = 2$. The four policy regions are: (R_1) both subchannels are allocated power; (R_2) only the first subchannel is allocated power; (R_3) only the second subchannel is allocated power; and (R_4) the system is in outage state.

SNR-plane (γ_1, γ_2) into four exclusive regions by the solid lines as shown in Fig. 22. If (γ_1, γ_2) falls into region R_1, both subchannels will be assigned with power for data transmission, where the boundaries of region R_1 is determined by $f_1(\gamma_1) = \gamma_0^{-2/\beta}\gamma_1^{(\beta+2)/\beta}$ and $f_2(\gamma_1) = \gamma_0^{2/(\beta+2)}\gamma_1^{\beta/(\beta+2)}$.[2] On the other hand, if (γ_1, γ_2) falls into either region R_2 or R_3, then only one of the subchannels will be assigned with power. Otherwise, if (γ_1, γ_2) belongs to region R_4, the system will be in an outage state. As shown by Fig. 22, the four regions are functions of γ_0 and β, which change as the values of γ_0 and β vary. Thus, based on (4.6), the cutoff threshold γ_0 is determined by satisfying the following power constraint:

$$\int_{\mathrm{R}_1} \left[\mu_1^{(1)}(\theta, \boldsymbol{\gamma}) + \mu_2^{(1)}(\theta, \boldsymbol{\gamma})\right] p_{\mathbf{\Gamma}}(\boldsymbol{\gamma}) d\gamma_1 d\gamma_2 + \int_{\mathrm{R}_2} \mu_1^{(2)}(\theta, \boldsymbol{\gamma}) p_{\mathbf{\Gamma}}(\boldsymbol{\gamma}) d\gamma_1 d\gamma_2 \\ + \int_{\mathrm{R}_3} \mu_2^{(2)}(\theta, \boldsymbol{\gamma}) p_{\mathbf{\Gamma}}(\boldsymbol{\gamma}) d\gamma_1 d\gamma_2 = 2 \tag{4.28}$$

where

$$\mu_n^{(1)}(\theta, \boldsymbol{\gamma}) = \frac{1}{\gamma_0^{\frac{1}{\beta+1}} (\gamma_1\gamma_2)^{\frac{\beta}{2(\beta+1)}}} - \frac{1}{\gamma_n} \tag{4.29}$$

and

$$\mu_n^{(2)}(\theta, \boldsymbol{\gamma}) = \frac{1}{\gamma_0^{\frac{2}{\beta+2}} \gamma_n^{\frac{\beta}{\beta+2}}} - \frac{1}{\gamma_n} \tag{4.30}$$

[2] The functions $f_1(\gamma_1)$ and $f_2(\gamma_1)$ are obtained by solving the boundary condition $\mathcal{N}_1 = \mathcal{N}_0$, where $\mathcal{N}_1$ is given by (4.23).

for $n = 1$ and $n = 2$, respectively. After obtaining γ_0 and using (4.10), the optimal effective capacity can be derived as follows:

$$\begin{aligned}\mathrm{E}_{\mathrm{C}}^{\mathrm{opt}}(\theta) \;=\; -\frac{1}{\theta}\log\Bigg(&\int_{\mathrm{R}_4} p_{\mathbf{\Gamma}}(\boldsymbol{\gamma})d\gamma_1 d\gamma_2 + \int_{\mathrm{R}_1}\prod_{n=1}^{2}\Big[1+\mu_n^{(1)}(\theta,\boldsymbol{\gamma})\gamma_n\Big]^{-\frac{\beta}{2}} p_{\mathbf{\Gamma}}(\boldsymbol{\gamma})d\gamma_1 d\gamma_2 \\ &+\int_{\mathrm{R}_2}\Big[1+\mu_1^{(2)}(\theta,\boldsymbol{\gamma})\gamma_n\Big]^{-\frac{\beta}{2}} p_{\mathbf{\Gamma}}(\boldsymbol{\gamma})d\gamma_1 d\gamma_2 + \int_{\mathrm{R}_3}\Big[1+\mu_2^{(2)}(\theta,\boldsymbol{\gamma})\gamma_n\Big]^{-\frac{\beta}{2}} p_{\mathbf{\Gamma}}(\boldsymbol{\gamma})d\gamma_1 d\gamma_2\Bigg).\end{aligned} \tag{4.31}$$

From the above example, we can find that even for a simple case of $N = 2$, the cutoff threshold γ_0 and the optimal effective capacity $\mathrm{E}_{\mathrm{C}}^{\mathrm{opt}}(\theta)$ generally do not have simple closed-form solutions. For the case with $N > 2$, the situation becomes even more complicated. However, by executing the proposed algorithm, γ_0 and $\mathrm{E}_{\mathrm{C}}^{\mathrm{opt}}(\theta)$ can be easily found through simulations for any given joint channel distribution $p_{\mathbf{\Gamma}}(\boldsymbol{\gamma})$. Thus, in this chapter, except for the trivial case of $N = 1$, we use simulation to find γ_0 and $\mathrm{E}_{\mathrm{C}}^{\mathrm{opt}}(\theta)$ for multiplexing-based systems. It is also worth noting that by using independent optimization approach, the power-adaptation policy partitions the SNR-plane (γ_1, γ_2) into four exclusive regions by the dashed lines as shown in Fig. 22.

2. Limiting Cases

One of the most significant differences between our proposed QoS-driven power adaptation and most other existing power-control approaches, such as the conventional water-filling algorithm, constant power scheme, and the independent optimization approach mentioned above, is that our proposed algorithm is executed in a *joint* fashion. Specifically, *the power assigned to one subchannel depends not only on its own channel quality, but also on the other subchannels' qualities*, by which the statistics of the aggregate service rate from all subchannels can be controlled to meet a certain delay-QoS requirement. In the following, we further study some limiting cases of our proposed optimal power-adaptation algorithms.

Case 1: When $N = 1$, or equivalently, all the subchannels are fully correlated, i.e., $\gamma_1 = \gamma_2 = \cdots = \gamma_N = \gamma$, the multichannel transmission reduces to single channel transmission. In this case, the joint pdf $p_{\mathbf{\Gamma}}(\boldsymbol{\gamma})$ reduces to $p_{\Gamma}(\gamma)$. Then, the optimal power-adaptation policy and power constraint turn out to be the ones reducing to our previous results in Chapter III, which is expected since single channel transmission is a special case of our multichannel communications.

Case 2: When the QoS exponent $\theta \to 0$, indicating that the system can tolerate an arbitrarily long delay, the optimal power-adaptation policy reduces to:

$$\lim_{\theta \to 0} \mu_n(\theta, \boldsymbol{\gamma}) = \begin{cases} \dfrac{1}{\gamma_0} - \dfrac{1}{\gamma_n}, & \gamma_n \geq \gamma_0, \\ 0, & \text{otherwise} \end{cases} \tag{4.32}$$

for all $n \in \mathcal{N}_0$, which is the water-filling formula for multichannel communications, where, as expected, the joint optimization reduces to the independent optimization. This observation verifies that the independent optimization approach is optimal to maximize the effective capacity as $\theta \to 0$. Thus, our QoS-driven power-adaptation scheme converges to water-filling algorithm when the system can tolerate an arbitrarily long delay. It also follows that the optimal effective capacity converges to the Shannon capacity as $\theta \to 0$.

Case 3: When the QoS exponent $\theta \to \infty$, then the system cannot tolerate any delay. In this case, the cutoff threshold $\gamma_0 \to 0$ (note that $\gamma_0 = \kappa_0/\beta$), which implies that the system does not enter the outage state almost surely. Letting $\theta \to \infty$ in (4.26) [i.e., Step (3)-b) in Fig. 21], we obtain the corresponding optimal strategy as follows:

$$\lim_{\theta\to\infty}\mu_n(\theta,\boldsymbol{\gamma}) = \begin{cases} \dfrac{1}{\phi^{\frac{N}{N^*}}\prod_{i\in\mathcal{N}^*}\gamma_i^{\frac{1}{N^*}}} - \dfrac{1}{\gamma_n}, & n\in\mathcal{N}^* \\ 0, & \text{otherwise} \end{cases} \tag{4.33}$$

where $\phi \triangleq \lim_{\theta\to\infty}\gamma_0^{\frac{1}{\beta+1}}$ and $\mathcal{N}^* \neq \varnothing$ almost surely. The power-control law given by (4.33) is just the policy to achieve the zero-outage capacity of the system [10, 11]. Thus, when the QoS exponent $\theta\to\infty$, the optimal throughput approaches the zero-outage capacity of the system. In summary, *as the QoS exponent θ increases from zero to infinity, the optimal effective capacity decreases accordingly from the ergodic capacity to zero-outage capacity.*

Plugging the power-control strategy given by (4.33) into (4.8), we can derive the resulting instantaneous service rate $R = R[i]$ when $\theta\to\infty$ as follows:

$$\begin{aligned} R &= \sum_{n=1}^{N^*}\left(\frac{T_fB}{N}\right)\log_2\Big(1+\mu_n(\theta,\boldsymbol{\gamma})\gamma_n\Big) \\ &= \left(\frac{T_fB}{N}\right)\log_2\left(\prod_{n\in\mathcal{N}^*}\left[\frac{\gamma_n}{\phi^{\frac{N}{N^*}}\prod_{i\in\mathcal{N}^*}\gamma_i^{\frac{1}{N^*}}}\right]\right) \\ &= \left(\frac{T_fB}{N}\right)\log_2\left(\frac{\prod_{n\in\mathcal{N}^*}\gamma_n}{\phi^N\prod_{i\in\mathcal{N}^*}\gamma_i}\right) \\ &= T_fB\log_2\left(\frac{1}{\phi}\right). \end{aligned} \tag{4.34}$$

That is, no matter what the channel realization is, the system maintains a constant service rate $T_fB\log_2(1/\phi)$. This result is also consistent with our previous work on single channel transmissions in Chapter III, where as the delay-QoS constraint becomes stringent, the optimal power control operates at a constant service rate. Since the service rate is constant, the effective capacity is also equal to this constant, i.e.,

$$\lim_{\theta \to \infty} \mathrm{E}_{\mathrm{C}}^{\mathrm{opt}}(\theta) = T_f B \log_2 \left(\frac{1}{\phi} \right). \tag{4.35}$$

From (4.35), we can observe that the smaller the value ϕ is, the larger the effective capacity $\mathrm{E}_{\mathrm{C}}^{\mathrm{opt}}(\theta)$ becomes. Our numerical results show that ϕ is a monotonic decreasing function of N. Consequently, when $\theta \to \infty$, the optimal effective capacity $\mathrm{E}_{\mathrm{C}}^{\mathrm{opt}}(\theta)$ increases as the number of subchannels N increases. In contrast, as mentioned in Remark 3 for Proposition 5, by using the independent power-control policies, as long as the number N of subchannels is finite, the effective capacity $\mathrm{E}_{\mathrm{C}}(\theta)$ always approaches zero as $\theta \to \infty$. Thus, our proposed joint optimization-based power control shows significant advantages over all the other independent power-control strategies as the delay-QoS constraint becomes stringent. For the MIMO multiplexing system, by using a similar procedure, we can show that

$$\lim_{\theta \to \infty} \mathrm{E}_{\mathrm{C}}^{\mathrm{opt}}(\theta) = N T_f B \log_2 \left(\frac{1}{\varphi} \right) \tag{4.36}$$

where $\varphi \triangleq \lim_{\theta \to \infty} \lambda_0^{\frac{1}{N\beta+1}}$. From (4.36), we can observe that when the QoS exponent $\theta \to \infty$, the effective capacity of the MIMO multiplexing system is almost a linearly increasing function of the number of subchannels $N = \min\{N_t, N_r\}$, which implies the significant superiority of employing the MIMO infrastructure for the QoS provisioning in mobile wireless networks.

G. Simulation Evaluations

We evaluate the performance of proposed QoS-driven power-adaptation algorithms by simulations. In this section, we mainly focus on three different diversity-based and multiplexing-based multichannel systems. We first simulate the multicarrier system

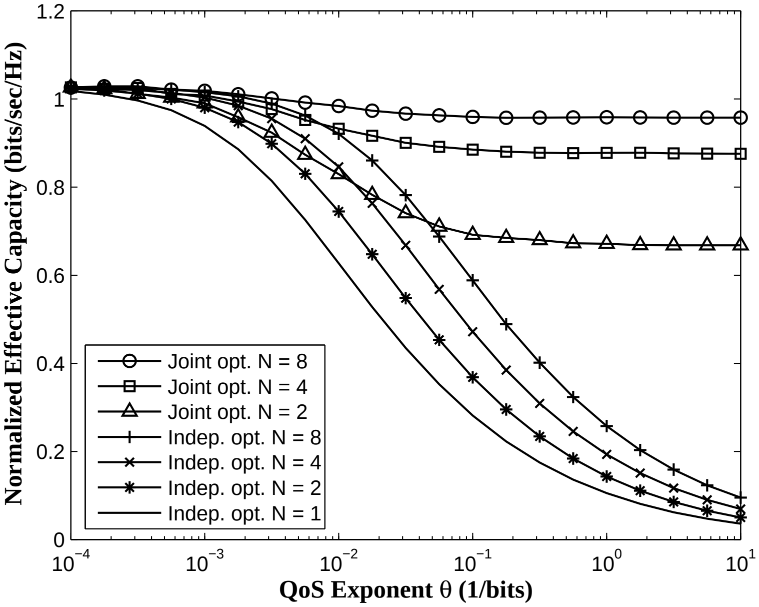


Fig. 23. The effective capacity comparisons between the joint optimization-based and independent optimization-based power-adaptation policies for N i.i.d. subchannels in a multicarrier system.

which utilizes frequency domain multiplexing. The fading statistics of different subcarriers are assumed to be i.i.d. Rayleigh distributed with average SNR $\overline{\gamma} = 0$ dB. We then simulate two MIMO systems which apply either diversity combining or multiplexing. For simplicity, we also assume that the fading statistics between all transmit and receive antenna pairs are i.i.d. Rayleigh distributed with average SNR $\overline{\gamma} = 0$ dB per receive antenna. The diversity combining MIMO scheme is Tx-beamforming/Rx-MRC (briefly termed as "beamforming" in the following for convenience) since this scheme provides the maximum spectral-efficiency among all MIMO diversity schemes. Furthermore, the system total spectral-bandwidth B is fixed to $B = 100$ KHz and the frame duration T_f is set to $T_f = 2$ ms for all simulations.

Fig. 23 plots the optimal effective capacity of multicarrier system against the QoS exponent θ with different number of subcarriers, where for comparison purpose,

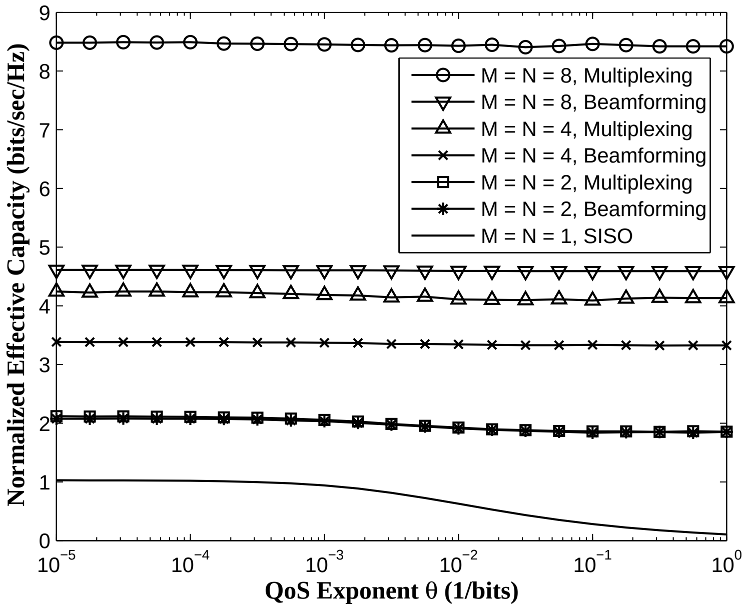
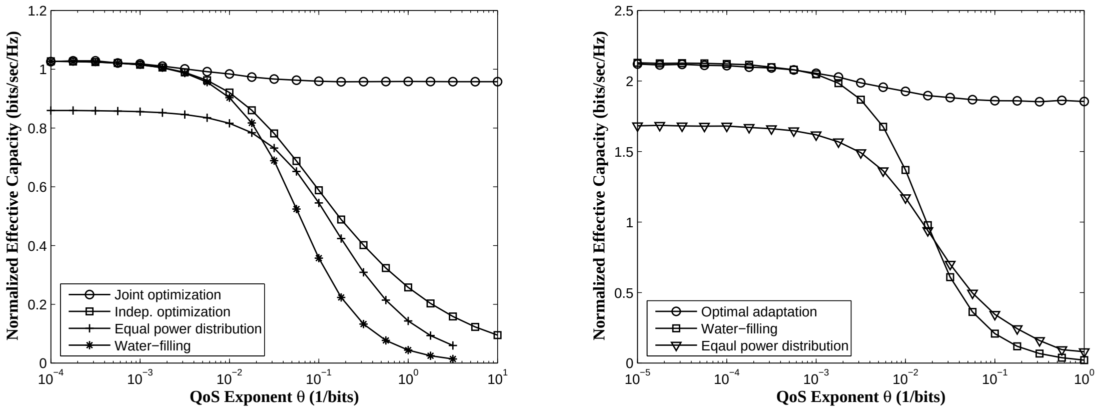


Fig. 24. The optimal effective capacity comparisons for MIMO systems using the different numbers of antennas.

we also plot the effective capacity using independent optimization approach. As mentioned in Section D, independent optimization of N subcarriers can right-shift the effective capacity curves for $10 \log_{10} N$ dB, compared to single-carrier system. Consequently, all the effective capacity curves approach zero as the QoS exponent θ increases. In contrast, based on our proposed joint optimization, the effective capacities are significantly larger than those of independent optimizations. As the QoS exponent θ increases, the effective capacity approaches a nonzero constant, where the larger the number of subcarriers, the higher the effective capacity. For example, by using only $N = 8$ i.i.d. subcarriers, the proposed scheme can achieve more than 90% of the Shannon capacity while still guaranteeing a constant rate transmission (as $\theta \to \infty$).

Fig. 24 plots the optimal effective capacities of MIMO diversity and multiplexing systems with different numbers of transmit and receive antennas. We can observe from

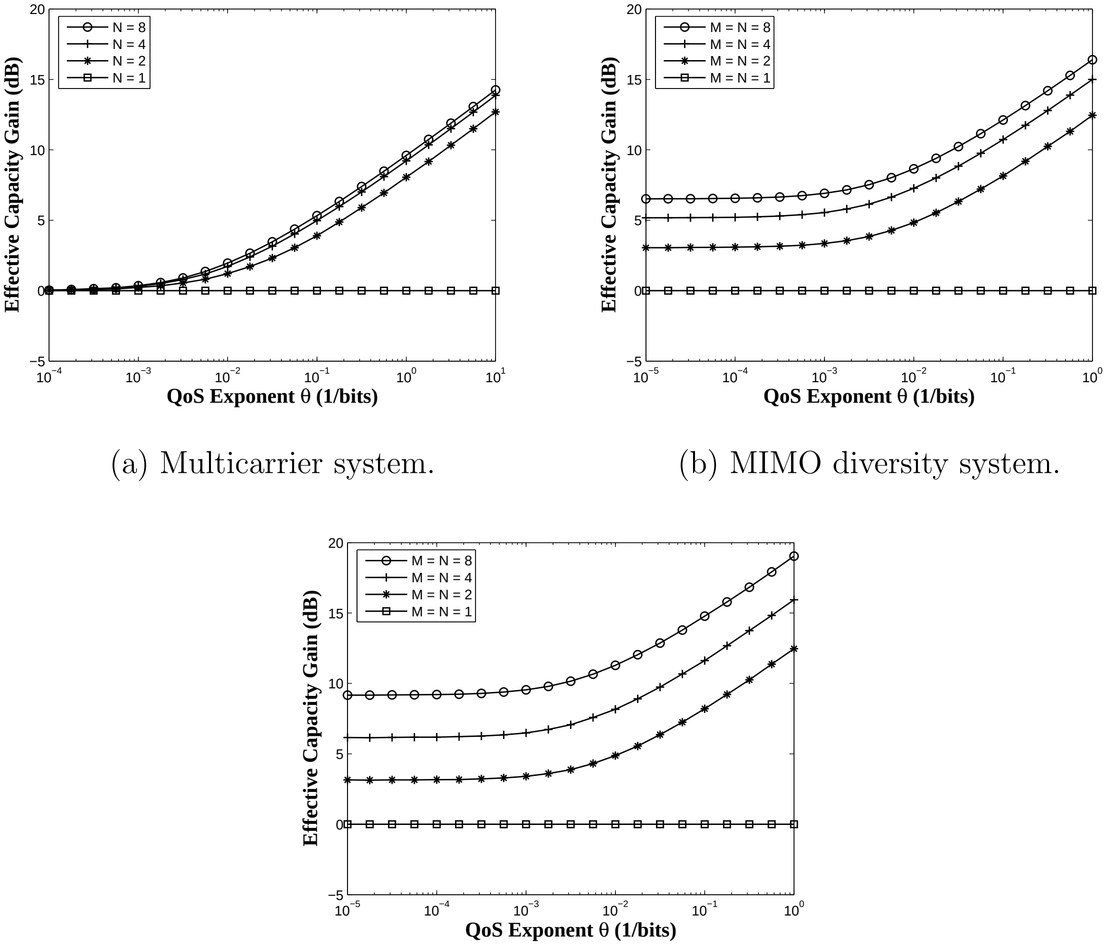


(a) Multicarrier system ($N = 8$). (b) MIMO multiplexing system (2×2).

Fig. 25. The effective capacity comparisons among different power allocation strategies for multiplexing-based systems.

Fig. 24 that the effective capacity increases as the number of antennas increases. When $M = N = 2$, where as defined in the above, $M = \max\{N_t, N_r\}$ and $N = \min\{N_t, N_r\}$, the performance loss of beamforming system compared to multiplexing system is virtually indistinguishable. However, as the number of antennas increases, the diversity gain is limited, but the multiplexing gain almost linearly increases with N. On the other hand, we can observe that just using a small number of transmit and receive antennas, the effective capacity of MIMO transmission is close to the Shannon capacity as $\theta \to \infty$, since all effective capacities are virtually constants, which implies that the MIMO system can guarantee stringent QoS with the service rate near Shannon-capacity.

To compare the impact of different power adaptations on QoS provisioning, Fig. 25 plots the effective capacities of multicarrier system and MIMO multiplexing system under different power-control policies. The power-adaptation schemes shown in Fig. 25 include our proposed optimal optimization, independent optimization for i.i.d multicarrier system, water-filling scheme, and equal power distribution



(a) Multicarrier system.

(b) MIMO diversity system.

(c) MIMO multiplexing system.

Fig. 26. The effective capacity gains compared to single channel (SISO) transmissions.

scheme. As expected, our proposed optimal power adaptation achieves the maximum effective capacity among all power-control policies. The optimal scheme converges to the water-filling for a small θ and converges to a constant for a large θ, where the effective capacity of all other schemes converges to zero for a large θ, which implies the significant advantage of our proposed scheme on supporting stringent QoS over other existing schemes.

Fig. 26 compares the effective-capacity gain of multichannel ($N > 1$) transmission with the single channel ($N = 1$) transmission. We can observe from Fig. 26 that by using the optimal power adaptation, our multichannel transmission-based scheme has the significant advantage over single channel transmission-based scheme, where the larger the QoS exponent θ, the higher the effective capacity gain. This means that multichannel transmission can support much more stringent QoS than single channel transmission. In particular, since the effective-capacity gain at $\theta \to 0$ is actually the spectral-efficiency gain, we can observe that for MIMO diversity and multiplexing system, the superiority of employing MIMO infrastructure in terms of enhancing QoS-guarantees is even more significant than that in terms of improving the spectral-efficiency.

Finally, Fig. 27 shows how much percentage of the Shannon capacity that the constant service rate can achieve by using our proposed optimal power adaptation (as $\theta \to \infty$). As expected, when the number of subchannels increases, the service rate gets closer and closer to the Shannon capacity. The percentage of Shannon capacity achieved is approximately proportional to the diversity order of the system, where for multicarrier systems, the diversity order is N, but for MIMO systems, the diversity order is $M \times N$. We can observe from Fig. 27 that when the system diversity order is 64, all multichannel systems can achieve more than 99% of the Shannon capacity, while still guaranteeing a constant rate transmission. In this case, a simple and efficient approach is to just use the fixed power-adaptation policy of our proposed scheme with $\theta \to \infty$, no matter what the delay-QoS constraint is, since this fixed power-adaptation policy can support both loose and stringent QoS requirements with only a slight throughput loss compared to the optimal Shannon capacity.

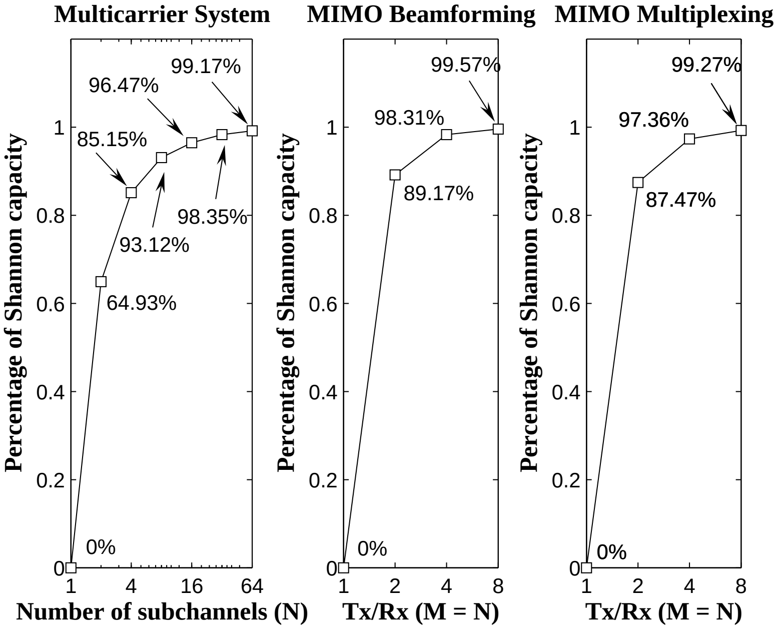


Fig. 27. The optimal effective capacity improvements as the function of the system diversity order compared to the Shannon capacity when the QoS exponent $\theta \rightarrow \infty$.

H. Summary

We have proposed and analyzed the QoS-driven power and rate adaptation schemes for diversity and multiplexing systems by integrating information theory with the effective capacity. The proposed resource allocation policies are general and applicable to different fading channel distributions. Our results showed that as the QoS exponent increases from zero to infinity, the optimal effective capacity decreases accordingly from the ergodic capacity to zero-outage capacity. Moreover, the multichannel transmission provides a significant advantage over single channel transmission for the stringent delay-QoS guarantees. Compared to the single channel transmission which has to deal with the tradeoff between throughputs and delay, the multichannel transmissions can achieve high throughput and stringent QoS at the same time.

Until now, we studied the resource allocation for single channel and multichannel systems, when assuming that the CSI feedback is perfect. However, in practice, CSI can never be perfect. Motivated by this practical concern, in the next chapter, we will consider the impact of channel estimation error on resource allocation and QoS provisioning.

CHAPTER V

RESOURCE ALLOCATION WITH CHANNEL ESTIMATION ERRORS

A. Introduction

The explosive demand for wireless services motivates a rapid evolution of wireless wideband communications. In order to efficiently support a large number of distinct wireless applications, such as wireless Internet, mobile computing, and cellular telephoning, diverse quality-of-service (QoS) guarantees play the increasing important role to the future wireless networks. Over the wireless environment, the most scarce radio resources are power and spectral bandwidth. In response, a great deal of research has been devoted to the techniques that can enhance the spectral efficiency of the wireless transmissions. The framework used to evaluate these techniques is mainly based on information theory [6, 7], using the concept of either *ergodic capacity* [8, 9] or *outage capacity* [10, 11]. The ergodic capacity maximizes the average spectral efficiency with an infinite long delay. The outage capacity, on the other hand, maintains a constant rate transmission with a certain outage probability. From the point-of-view of delay QoS, such an information-theoretic framework maximizes the system throughput either without any delay constraint (i.e., ergodic capacity), or with a stringent delay constraint (i.e., outage capacity). These two extremes may not refine enough for the user's satisfactions, where a wide range of delay constraints may be requested for different applications. Consequently, to provide diverse QoS guarantees, it is necessary to take the QoS metrics into account when applying the prevalent information theory.

In Chapters III and IV, we proposed QoS-driven power allocation schemes for single-input-single-output (SISO) and also multiple-input-multiple-output (MIMO) systems, respectively, when assuming *perfect* channel state information (CSI) available at both the transmitter and receiver. The proposed scheme aims at maximizing the system throughput subject to a given delay constraint. Specifically, by integrating information theory with the concept of *effective capacity* [20–23], we convert the original problem to the one with the target at maximizing the effective capacity, in which the delay QoS constraint is characterized by the QoS exponent θ. Applying the effective capacity, a smaller θ corresponds to a looser QoS guarantee, while a larger θ implies a more stringent QoS requirement. In the limiting case, when $\theta \to 0$, the system can tolerate an arbitrarily long delay, which is the scenario to derive the ergodic capacity. In contrast, when $\theta \to \infty$, the system cannot tolerate any delay, which corresponds to the case to obtain the zero-outage capacity. Thus, as θ dynamically varies, the optimal power allocation builds up a bridge between the ergodic capacity and the zero-outage capacity.

As the sequel of Chapters III and IV, this chapter focuses on QoS provisioning over parallel channels in the presence of channel estimation errors. Our study is based on the *block-fading* (also known as quasi-static) channel model. The physical validity of this model is discussed in [83]. Due to its analytical convenience, the block-fading channel model is commonly used in literatures [10, 11, 21, 22, 25, 84, 85], which also greatly simplifies our analyses. We concentrate on communications over *parallel channels*, since this is a fundamental communication mechanism, where a large number of promising techniques fall into this category. For instance, multicarrier systems employing orthogonal-frequency-division-multiplexing (OFDM) can be considered as parallel communications at the *frequency* domain [85, 86]. In contrast, the MIMO system is an typical example which utilizes *spatial* domain parallel channels [8, 11, 25]. The emerging MIMO-OFDM architecture combines parallel channels in a joint spatial-frequency domain. On the other hand, the simple SISO system is also a special case of parallel communications, where the number of parallel channels is one.

The research of this chapter is mainly motivated by a practical concern, where a perfect CSI is hard to obtain in real wireless networks [25, 84, 85, 87]. Therefore, it becomes critically important to investigate how to deal with such an imperfectness, and what is its impact on QoS provisioning. Compared to the case with perfect CSI, imperfect CSI imposes new challenges to our throughput maximization problem. In particular, the problem is *not* convex in nature. To overcome this mathematical difficulty, we divide the original non-convex problem into two orthogonal sub-problems, each of which turns out to be convex and can be solved efficiently. The main contributions of this chapter can be summarized as follows.

1. We derive the power allocation algorithm under the *total* power constraint (Theorem 2), which shows that the optimal policy is actually classic water-filling, regardless of delay requirement.

2. We propose the power allocation scheme under the *average* power constraint (Theorem 3), which shows that as the QoS exponent θ increases from zero to infinity, the optimal effective capacity decreases from the ergodic capacity to the zero-outage capacity.

3. Under stringent delay requirement, we provide necessary and sufficient conditions for the convergence of the average power (Theorem 4), which shows that in the presence of channel estimation errors, the average power always diverges. Furthermore, a positive zero-outage capacity is proved to be unattainable. Alternatively, we explicitly obtain the power allocation scheme to minimize the outage probability (Theorem 5).

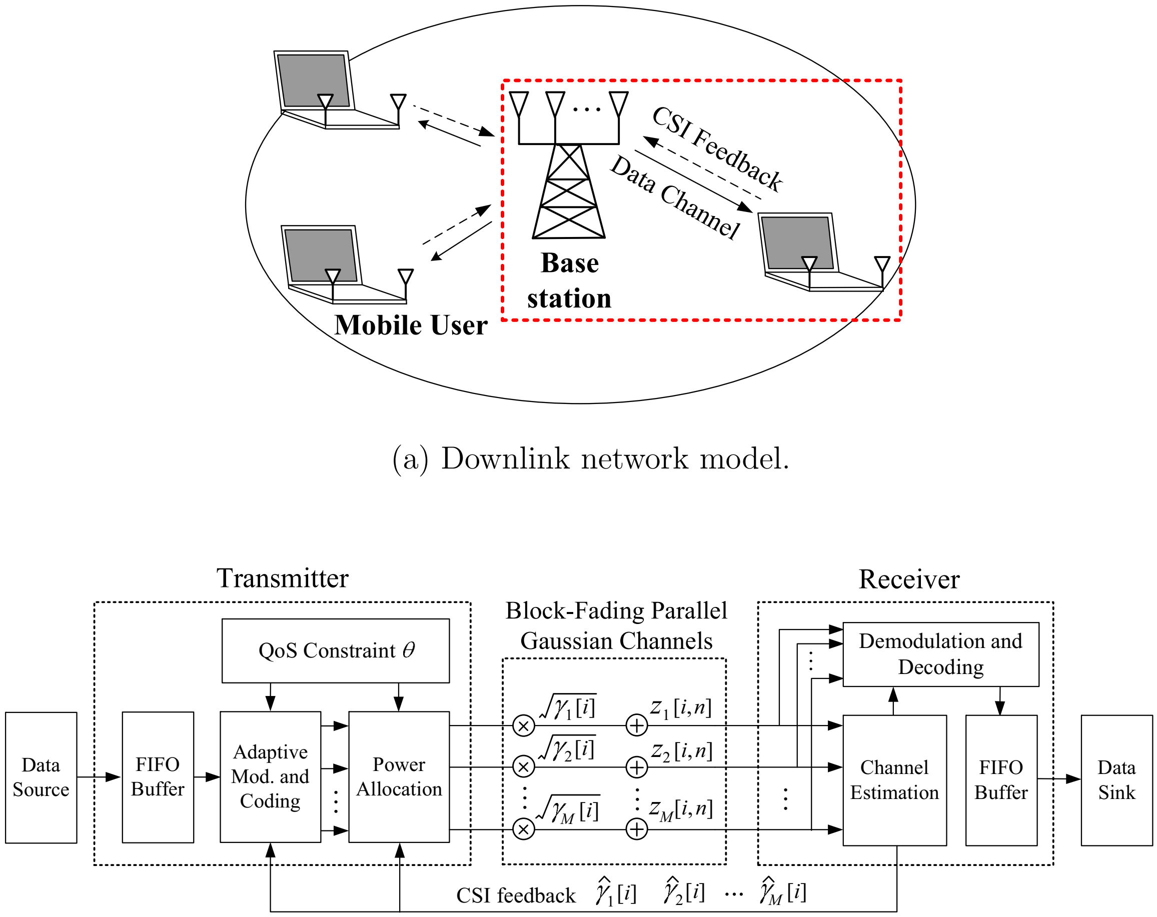
Our results also suggest that a larger number of parallel channels can provide higher throughput and more stringent QoS, while offering better robustness against the channel estimation errors.

The rest of the chapter is organized as follows. Section B describes our parallel system model. Sections C derives the optimal power allocation policy with different power constraints. Section D discusses the power allocation strategy for stringent QoS provisioning. Section E conducts simulations to evaluate the performance of our proposed scheme. The chapter concludes with Section F.

Notations. We use upper- and lower-case boldface letters to denote matrices and vectors, respectively. $\mathbb{R}$ and $\mathbb{C}$ indicate the space of real and complex numbers, respectively, with possible superscript denoting the dimension of the matrices or vectors. $\mathbb{R}_+$ and $\mathbb{R}_{++}$ represent the nonnegative and positive real numbers, respectively. $(x)^+ \triangleq \max\{0, x\}$. $\mathbb{E}[\cdot]$ stands for the expectation, $\mathbb{E}_{\boldsymbol{x}}[\cdot]$ represents that the expectation is with respect to $\boldsymbol{x}$. $\boldsymbol{I}_K$ denotes a $K \times K$ identity matrix. $\boldsymbol{x} \sim \mathcal{CN}(\boldsymbol{u}, \boldsymbol{\Sigma})$ means that the complex random vector $\boldsymbol{x}$ follows a jointly Gaussian distribution with mean $\boldsymbol{u}$ and covariance matrix $\boldsymbol{\Sigma}$.

B. System Model

The system model is illustrated in Fig. 28. We concentrate on a discrete-time point-to-point link between the base station (transmitter) and one of the mobile users (receivers) in downlink wireless networks, as shown in Fig 28(a). In particular, the transmitter and the receiver are communicating through M parallel fading channels over spectral bandwidth B. As shown in Fig. 28(b), a first-in-first-out (FIFO) buffer is equipped at the transmitter, which buffers the data *frames* to be transmitted to the receiver. Each frame consists of $M \times N$ symbols. The frame duration is denoted by T_f, which is assumed to be less than the fading coherence time, but sufficiently



(a) Downlink network model.

(b) Point-to-point link.

Fig. 28. The downlink network model and the point-to-point model between the base station and the mobile user.

long so that the information-theoretic assumption of infinite code-block length (i.e., $N \to \infty$) is meaningful [10, 11]. The frame is then divided into M substreams, each with N symbols transmitted through one of the parallel channels. Based on a given QoS constraint θ requested by the mobile session and CSI fed back from the mobile receiver, the transmitter needs to find an optimal codeword (implemented by the adaptive modulation and coding) and a corresponding power allocation strategy, which can maximize the throughput subject to the QoS constraint θ.

The discrete-time channel process is assumed to be block-fading. Specifically, the path gains are constant within a frame's duration T_f, but vary *independently* from one frame to another, following a certain continuous distribution. Note that the most commonly used channel distributions, such as Rayleigh, Rice, Nakagami, Log-normal, and Wishart, are all continuous and thus belong to this category. The transmission for the nth symbol of the ith frame can be modeled as

$$\boldsymbol{y}[i,n] = \sqrt{\boldsymbol{\Gamma}[i]}\boldsymbol{x}[i,n] + \boldsymbol{z}[i,n]$$

where $i = 1, 2, ...$ denotes the frame index, $n = 1, 2, ..., N$ denotes the symbol index, $\boldsymbol{x}[i,n] \in \mathbb{C}^M$ and $\boldsymbol{y}[i,n] \in \mathbb{C}^M$ are complex channel input and output symbols, respectively, $\sqrt{\boldsymbol{\Gamma}[i]} \triangleq \text{diag}\{\sqrt{\gamma_1[i]}, \sqrt{\gamma_2[i]}, ..., \sqrt{\gamma_M[i]}\} \in \mathbb{R}_+^{M\times M}$ denotes the diagonal channel gain matrix, and $\boldsymbol{z}[i,n] \sim \mathcal{CN}(\boldsymbol{0}, \boldsymbol{I}_M)$ is i.i.d. complex additive white Gaussian noise (AWGN), which, by a properly transmit power scaling, can be normalized to have the unit variance.

Let $\boldsymbol{\gamma}[i] \triangleq (\gamma_1[i], \gamma_2[i], ..., \gamma_M[i])$ denote the instantaneous CSI. When the receiver knows *perfectly* about $\boldsymbol{\gamma}[i]$, for a given power allocation $\boldsymbol{\mu}[i] \triangleq (P_1[i], P_2[i], ..., P_M[i]) \in \mathbb{R}_+^M$, the maximum instantaneous mutual information between channel inputs and outputs, denoted by $\mathcal{I}(\boldsymbol{\mu}[i], \boldsymbol{\gamma}[i])$, can be expressed as

$$\mathcal{I}(\boldsymbol{\mu}[i], \boldsymbol{\gamma}[i]) \triangleq \frac{T_f B}{K} \sum_{m=1}^{M} \log_2 \Big(1 + \gamma_m[i] P_m[i] \Big) \tag{5.1}$$

which can be achieved by the independent complex Gaussian inputs expressed by $\boldsymbol{x}[i,n] \sim \mathcal{CN}(\mathbf{0}, \mathrm{diag}\{\boldsymbol{\mu}[i]\})$. In Eq. (5.1), the parameter K with $1 \leq K \leq M$ is a scaling constant dependent on the specific parallel transmission scheme. For instance, when $\boldsymbol{\gamma}[i]$ corresponds to M singular-values of the spatial MIMO channel, we have $K = 1$. On the other hand, when $\boldsymbol{\gamma}[i]$ corresponds to M subchannel gains of a multicarrier system, K is equal to M.

In this chapter, we are interested in the scenario where $\boldsymbol{\gamma}[i]$ is *imperfectly* known to the receiver. Let $\widehat{\boldsymbol{\gamma}}[i] \in \mathbb{R}_+^M$ denote the estimation of the actual CSI $\boldsymbol{\gamma}[i]$. Given $\boldsymbol{\mu}[i]$ and $\widehat{\boldsymbol{\gamma}}[i]$, the closed-form expression for the maximum instantaneous mutual information between the channel inputs and outputs turns out to be intractable, even in the simple case of $M = 1$ [84]. However, under sufficient conditions, a tight lower-bound, denoted by $\widehat{\mathcal{I}}(\boldsymbol{\mu}[i], \widehat{\boldsymbol{\gamma}}[i])$, can be obtained as [25, 84, 85]

$$\widehat{\mathcal{I}}(\boldsymbol{\mu}[i], \widehat{\boldsymbol{\gamma}}[i]) \triangleq \frac{T_f B}{K} \sum_{m=1}^{M} \log_2 \left(1 + \frac{\widehat{\gamma}_m[i] P_m[i]}{1 + \sigma_e^2 \sum_{m=1}^{M} P_m[i]} \right) \tag{5.2}$$

where σ_e^2 denotes the variance of the channel estimation errors, which depends on the channel dynamics and channel estimation schemes employed [25], and is assumed to be known *a priori* at the both ends of the link. The mutual information lower-bound in Eq. (5.2) can be achieved by the independent complex Gaussian inputs and nearest neighbor decoding rule, see, e.g., [25, 87] for a detailed discussion. It is also clear that when $\sigma_e^2 \to 0$, we have $\widehat{\gamma}_m[i] \to \gamma_m[i]$, and Eq. (5.2) reduces to Eq. (5.1).

In this chapter, we also make the following assumptions.

A1: We assume that the estimated CSI $\widehat{\boldsymbol{\gamma}}[i]$ is reliably fed back to the transmitter without delay. The issues of feedback delay and unreliable feedback channels can be modeled as a *channel mean feedback* problem [88], which is not the focus of this chapter. In addition, the preliminary work about the impact of feedback delay on the QoS provisioning can be found in [60] (also see in Chapter VI for details).

A2: We further assume that given a power allocation $\boldsymbol{\mu}[i]$ and the estimated CSI $\widehat{\boldsymbol{\gamma}}[i]$, the adaptive modulation and coding can choose an ideal channel code for each frame, such that the transmission rate, denoted by $R(\boldsymbol{\mu}[i], \widehat{\boldsymbol{\gamma}}[i])$, achieves the mutual information lower-bound $\widehat{\mathcal{I}}(\boldsymbol{\mu}[i], \widehat{\boldsymbol{\gamma}}[i])$ given in Eq. (5.2). Based on this assumption, the derived effective capacity using Eq. (5.2) also serves as a lower-bound for the optimal effective capacity.

A3: In practice, the channel estimation itself may cause a certain power loss. In this chapter, since our focus is to study the impact of imperfect CSI on QoS provisioning, we ignore such a performance degradation factor. Based on our framework, the results can be easily extended to the case considering the cost of channel estimations.

In the following discussions, since the block-fading channel process is i.i.d., its instantaneous marginal statistics is independent of the frame index i, and thus we may omit the frame index i for simplicity.

C. Power Allocation for QoS Provisioning

1. Problem Formulation

Let us define $\boldsymbol{\nu} \triangleq (\theta, \widehat{\boldsymbol{\gamma}})$ as *network state information* (NSI). Then, based on Eq. (5.2) and assumption **A2**, the transmission rate, denoted by $R\big(\boldsymbol{\mu}(\boldsymbol{\nu}), \widehat{\boldsymbol{\gamma}}\big)$, can be expressed as

$$R\big(\boldsymbol{\mu}(\boldsymbol{\nu}), \widehat{\boldsymbol{\gamma}}\big) = \frac{T_f B}{K} \sum_{m=1}^{M} \log_2 \left(1 + \frac{\widehat{\gamma}_m P_m(\boldsymbol{\nu})}{1 + \sigma_e^2 \sum_{m=1}^{M} P_m(\boldsymbol{\nu})}\right) \tag{5.3}$$

where the power allocation policy $\boldsymbol{\mu}(\theta, \widehat{\boldsymbol{\gamma}}) = \boldsymbol{\mu}(\boldsymbol{\nu}) = (P_1(\boldsymbol{\nu}), P_2(\boldsymbol{\nu}), ..., P_M(\boldsymbol{\nu})) \in \mathbb{R}_+^M$ is not only the function of the estimated CSI $\widehat{\boldsymbol{\gamma}}$, but also the function of the QoS exponent θ. For a given QoS constraint specified by θ, in order to find the optimal power allocation policy, denoted by $\boldsymbol{\mu}^*(\boldsymbol{\nu})$, that maximizes the effective capacity of Eq. (2.15), we can formulate a maximization problem as follows:

$$\boldsymbol{\mu}^*(\boldsymbol{\nu}) = \arg\max_{\boldsymbol{\mu}(\boldsymbol{\nu})} \left\{ -\frac{1}{\theta} \log\left(\mathbb{E}_{\widehat{\boldsymbol{\gamma}}} \Big[\mathcal{F}\big(\boldsymbol{\mu}(\boldsymbol{\nu}), \widehat{\boldsymbol{\gamma}}\big) \Big] \right) \right\}$$

where

$$\begin{aligned} \mathcal{F}\big(\boldsymbol{\mu}(\boldsymbol{\nu}), \widehat{\boldsymbol{\gamma}}\big) &\triangleq e^{-\theta R\left(\boldsymbol{\mu}(\boldsymbol{\nu}), \widehat{\boldsymbol{\gamma}}\right)} \\ &= \prod_{m=1}^{M} \left(1 + \frac{\widehat{\gamma}_m P_m(\boldsymbol{\nu})}{1 + \sigma_e^2 \sum_{m=1}^{M} P_m(\boldsymbol{\nu})} \right)^{-\beta} \end{aligned} \tag{5.4}$$

with $\beta \triangleq \theta T_f B/(K \log 2)$ defined as normalized QoS exponent. Since $\log(\cdot)$ is a monotonically increasing function, for each given QoS constraint $\theta \in \mathbb{R}_{++}$, the maximization problem above is equivalent to the following minimization problem:

$$\boldsymbol{\mu}^*(\boldsymbol{\nu}) = \arg\min_{\boldsymbol{\mu}(\boldsymbol{\nu})} \left\{ \mathbb{E}_{\widehat{\boldsymbol{\gamma}}} \Big[\mathcal{F}\big(\boldsymbol{\mu}(\boldsymbol{\nu}), \widehat{\boldsymbol{\gamma}}\big) \Big] \right\}. \tag{5.5}$$

In this chapter, we mainly consider two different power constraints. A simple and practical constraint is known as *total power constraint*, also called short-term power constraint [10]. Specifically, the transmit power for each frame cannot exceed a certain threshold P_{total}, i.e.,

$$\sum_{m=1}^{M} P_m(\boldsymbol{\nu}) \le P_{\text{total}} \tag{5.6}$$

for all realizations of $\widehat{\boldsymbol{\gamma}} \in \mathbb{R}_+^M$.

On the other hand, the *average power constraint*, also known as long-term power constraint [10], is often investigated from an information-theoretic point-of-view. Under the average power constraint, the mean of the transmit power cannot exceed a certain threshold P_{avg}, but no restriction is imposed on the instantaneous transmit power, i.e.,

$$\mathbb{E}_{\widehat{\gamma}}\left[\sum_{m=1}^{M} P_m(\boldsymbol{\nu})\right] \leq P_{\text{avg}}. \tag{5.7}$$

A system may be subject to the total power or/and average power constraints, which are elaborated on in the followings, respectively.

2. Power Allocation With Total Power Constraint

We first consider the problem of minimizing Eq. (5.5) subject to the total power constraint given by Eq. (5.6). It is clear that Eq. (5.5) achieves its minimum when the constraint in Eq. (5.6) is satisfied with equality. Accordingly, let us define a convex set, denoted by $\mathcal{S}$, for the power allocation policy as follows:

$$\mathcal{S} \triangleq \left\{ \boldsymbol{\mu}(\boldsymbol{\nu}) : \boldsymbol{\mu}(\boldsymbol{\nu}) \in \mathbb{R}_+^M, \sum_{m=1}^{M} P_m(\boldsymbol{\nu}) = P_{\text{total}} \right\}.$$

Then, we have the following lemma.

Lemma 2. *The objective function $\mathbb{E}_{\widehat{\gamma}}\left[\mathcal{F}\left(\boldsymbol{\mu}(\boldsymbol{\nu}), \widehat{\gamma}\right)\right]$ given in Eq. (5.5) is strictly convex on $\mathcal{S}$.*

Proof. It is easy to verify that $R(\boldsymbol{\mu}(\boldsymbol{\nu}), \widehat{\gamma})$ given in Eq. (5.3) is strictly concave on $\mathcal{S}$. On the other hand, $f(x) = e^{-\theta x}$ is a strictly convex and non-increasing function for any fixed $\theta \in \mathbb{R}_{++}$. Using the property given by [82, eq. (3.10)], we know that $\mathcal{F}\left(\boldsymbol{\mu}(\boldsymbol{\nu}), \widehat{\gamma}\right) = \exp(-\theta R(\boldsymbol{\mu}(\boldsymbol{\nu}), \widehat{\gamma}))$ is strictly convex on $\mathcal{S}$. Finally, since the expectation is a linear operation, it preserves the strictly convexity. The proof follows. □

Since the objective function given in Eq. (5.5) is strictly convex on $\mathcal{S}$, we can use the standard Lagrangian method to find the unique optimal power allocation policy, denoted by $\boldsymbol{\mu}^*_{\text{total}}(\boldsymbol{\nu}) \in \mathcal{S}$. Construct the Lagrange as follows:[1]

$$\mathcal{J}_1 = \mathbb{E}_{\widehat{\boldsymbol{\gamma}}}\left[\prod_{m=1}^{M}\left(1 + \frac{\widehat{\gamma}_m P_m(\boldsymbol{\nu})}{1+\sigma_e^2 P_{\text{total}}}\right)^{-\beta}\right] + \lambda_1 \sum_{m=1}^{M} P_m(\boldsymbol{\nu}) \tag{5.8}$$

where λ_1 denotes the Lagrangian multiplier. By solving the Karush-Kuhn-Tucker (KKT) condition [82] of Eq. (5.8), we obtain the optimal power allocation policy $\boldsymbol{\mu}^*_{\text{total}}(\boldsymbol{\nu})$, which can be described by the following theorem.

Theorem 2. *For each estimated fading state $\widehat{\boldsymbol{\gamma}}$, let $\pi(\cdot)$ be defined as a permutation of $\widehat{\boldsymbol{\gamma}}$ such that $\widehat{\gamma}_{\pi(1)} \geq \widehat{\gamma}_{\pi(2)} \geq \cdots \geq \widehat{\gamma}_{\pi(M)}$. Also define*

$$\widetilde{\gamma}_{\pi(m)} \triangleq \frac{\widehat{\gamma}_{\pi(m)}}{1+\sigma_e^2 P_{\text{total}}} \tag{5.9}$$

for all $m = 1, 2, ..., M$. Then, the $\pi(m)$-th component of $\boldsymbol{\mu}^*_{\text{total}}(\boldsymbol{\nu})$, denoted by $P^*_{\pi(m)}(\boldsymbol{\nu})$, follows the classic water-filling formula and is determined by

$$P^*_{\pi(m)}(\boldsymbol{\nu}) = \left(\omega(\boldsymbol{\nu}, k) - \frac{1}{\widetilde{\gamma}_{\pi(m)}}\right)^+ \tag{5.10}$$

where $\omega(\boldsymbol{\nu}, k)$ denotes the time-varying water-level, which is chosen such that the total power constraint is satisfied, and is given by

$$\omega(\boldsymbol{\nu}, k) = \frac{1}{k}\left(P_{\text{total}} + \sum_{i=1}^{k} \frac{1}{\widetilde{\gamma}_{\pi(i)}}\right). \tag{5.11}$$

The parameter k in Eqs. (5.10) and (5.11) denotes the number of active *channels allocated with nonzero power, which is the unique integer in $\{1, 2, ..., M\}$ such that $\omega(\boldsymbol{\nu}, k) > 1/\widetilde{\gamma}_{\pi(m)}$ for $m \leq k$ and $\omega(\boldsymbol{\nu}, k) \leq 1/\widetilde{\gamma}_{\pi(m)}$ for $m > k$.*

Proof. The sketch of the proof is provided in Appendix H. □

Remark 4. The water-level $\omega(\boldsymbol{\nu}, k)$ and the number of active channels k are jointly

[1] In this chapter, the explicit Lagrangian multipliers corresponding to the constraint $\boldsymbol{\mu}(\boldsymbol{\nu}) \in \mathbb{R}^M_+$ are omitted.

determined by the channel state $\widehat{\boldsymbol{\gamma}}$. As a result, different fading states $\widehat{\boldsymbol{\gamma}}$ correspond to different $\omega(\boldsymbol{\nu}, k)$ and k.

Remark 5. Although our objective is to maximize the throughput subject to the QoS constraint θ, Theorem 2 states that the optimal power allocation under the total power constraint is actually independent of θ. This implies that under the total power constraint, the water-filling formula is *always* the optimal power allocation policy, regardless of θ. On the other hand, since this policy does not distinguish the services with different QoS constraints, the power is not allocated in favor of the QoS provisioning.

Substituting Eqs. (5.10) and (5.11) into Eq. (5.5) with some algebraic manipulations, we obtain the minimum objective function under the total power constraint as follows:

$$\mathbb{E}_{\widehat{\boldsymbol{\gamma}}}\Big[\mathcal{F}\left(\boldsymbol{\mu}^*_{\text{total}}(\boldsymbol{\nu}), \widehat{\boldsymbol{\gamma}}\right)\Big] = \mathbb{E}_{\widehat{\boldsymbol{\gamma}}}\left[\left\{\frac{k\Sigma_k\Pi_k\left(1+\sigma_e^2 P_{\text{total}}\right)}{1+(\sigma_e^2+\Sigma_k)P_{\text{total}}}\right\}^{k\beta}\right] \tag{5.12}$$

where, for notational convenience, we define $\Sigma_k \triangleq 1/\big(\sum_{i=1}^{k}\widehat{\gamma}_{\pi(i)}^{-1}\big)$ and also $\Pi_k \triangleq \prod_{i=1}^{k}\widehat{\gamma}_{\pi(i)}^{-1/k}$.

3. Power Allocation With Average Power Constraint

In this section, we focus on minimizing Eq. (5.5) subject to the average power constraint given by Eq. (5.7). This problem is more difficult than that under the total power constraint, since when $\sigma_e^2 > 0$, the objective function in Eq. (5.5) is not convex on the entire space spanned by $\boldsymbol{\mu}(\boldsymbol{\nu}) \in \mathbb{R}_+^M$. Alternatively, we obtain the optimal solution by a two-step approach.

Noticing that for each given total power P_{total}, by Theorem 2, we already know the optimal power allocation policy $\boldsymbol{\mu}^*_{\text{total}}(\boldsymbol{\nu})$. However, under the average power constraint, the instantaneous total power P_{total} changes with each fading state. In response, we rewrite P_{total} by $P_{\text{total}}(\boldsymbol{\nu})$ to emphasize such a temporal variation. To obtain the optimal power allocation under the average power constraint, we can solve the problem into two steps. The first step is to find the optimal *temporal* power allocation policy, denoted by $P^*_{\text{total}}(\boldsymbol{\nu}) \in \mathbb{R}_+$, which minimizes the objective function Eq. (5.5) while meeting the average power constraint:

$$\mathbb{E}_{\widehat{\boldsymbol{\gamma}}}\left[P^*_{\text{total}}(\boldsymbol{\nu})\right] = P_{\text{avg}}. \tag{5.13}$$

Once the optimal policy $P^*_{\text{total}}(\boldsymbol{\nu})$ is obtained, the second step is to assign power along the M parallel channels according to the water-filling algorithm described in Theorem 2, satisfying $\sum_{m=1}^{M} P^*_m(\boldsymbol{\nu}) = P^*_{\text{total}}(\boldsymbol{\nu})$.

Noting that when deriving the optimal policy $P^*_{\text{total}}(\boldsymbol{\nu})$ for the first step, an underlying assumption is that at the second step, the policy $\boldsymbol{\mu}^*_{\text{total}}(\boldsymbol{\nu})$ is applied for each $P^*_{\text{total}}(\boldsymbol{\nu})$. Therefore, the objective function for the first step can be expressed as Eq. (5.12), instead of the original one in Eq. (5.5). Based on Eq. (5.12), we formulate the new optimization problem as follows:

$$P^*_{\text{total}}(\boldsymbol{\nu}) = \arg\min_{P_{\text{total}}(\boldsymbol{\nu})} \left\{ \mathbb{E}_{\widehat{\boldsymbol{\gamma}}}\left[\left\{ \frac{k\Sigma_k \Pi_k \left[1 + \sigma_e^2 P_{\text{total}}(\boldsymbol{\nu})\right]}{1 + (\sigma_e^2 + \Sigma_k) P_{\text{total}}(\boldsymbol{\nu})} \right\}^{k\beta} \right] \right\} \tag{5.14}$$

subject to the average power constraint given in Eq. (5.13).

Let us define a convex set, denoted by $\mathcal{S}'$, for the temporal power allocation policy as follows:

$$\mathcal{S}' \triangleq \left\{ P_{\text{total}}(\boldsymbol{\nu}) : P_{\text{total}}(\boldsymbol{\nu}) \in \mathbb{R}_+, \mathbb{E}_{\widehat{\boldsymbol{\gamma}}}\left[P_{\text{total}}(\boldsymbol{\nu})\right] = P_{\text{avg}} \right\}. \tag{5.15}$$

Then, we have the following lemma.

Lemma 3. *The objective function given in Eq. (5.14) is strictly convex on set $\mathcal{S}'$.*

Proof. The proof is provided in Appendix I. □

Due to the convexity of Eq. (5.14) on set $\mathcal{S}'$, we divide the original non-convex problem into two sub-problems, each of which is convex. An illustration of the 2-dimensional convex optimization is shown in Fig. 29. Although the objective function is not convex, the two optimized dimensions are always convex, respectively. From Lemma 3, once again, we can use the Lagrangian technique to derive the unique optimal temporal power allocation policy $P_{\text{total}}^*(\boldsymbol{\nu}) \in \mathcal{S}'$. Construct the Lagrange as follows:

$$\mathcal{J}_2 = \mathbb{E}_{\widehat{\gamma}}\left[\left\{\frac{k\Sigma_k\Pi_k\left[1+\sigma_e^2 P_{\text{total}}(\boldsymbol{\nu})\right]}{1+(\sigma_e^2+\Sigma_k)P_{\text{total}}(\boldsymbol{\nu})}\right\}^{k\beta}\right] + \lambda_2\mathbb{E}_{\widehat{\gamma}}\left[P_{\text{total}}(\boldsymbol{\nu})\right] \tag{5.16}$$

where λ_2 denotes the Lagrangian multiplier. Solving the above Lagrangian problem, we obtain the optimal temporal power allocation policy $P_{\text{total}}^*(\boldsymbol{\nu})$ under the average power constraint, which can be described by the following theorem.

Theorem 3. *The optimal temporal power allocation policy $P_{\text{total}}^*(\boldsymbol{\nu}) \in \mathbb{R}_{++}$, if exists, is the unique positive solution of the following equation:*

$$\frac{\left[1+(\sigma_e^2+\Sigma_k)P_{\text{total}}(\boldsymbol{\nu})\right]^{\frac{k\beta+1}{M\beta+1}}}{(k\Sigma_k)^{\frac{k\beta+1}{M\beta+1}}\,\Pi_k^{\frac{k\beta}{M\beta+1}}\left[1+\sigma_e^2 P_{\text{total}}(\boldsymbol{\nu})\right]^{\frac{k\beta-1}{M\beta+1}}} = \omega^*. \tag{5.17}$$

Otherwise, if such a solution $P_{\text{total}}^*(\boldsymbol{\nu}) \in \mathbb{R}_{++}$ does not exist, then $P_{\text{total}}^*(\boldsymbol{\nu}) = 0$. In Eq. (5.17), $\omega^* \in \mathbb{R}_{+}$ is a constant, which is chosen such that the average power constraint is satisfied.

Proof. The proof is provided in Appendix J. □

Remark 6. The constant ω^* can be called *water-level coefficient*, which is proportional to the average power constraint. The higher the average power constraint P_{avg}, the larger the water-level coefficient ω^*. Once ω^* is determined, it remains as a constant

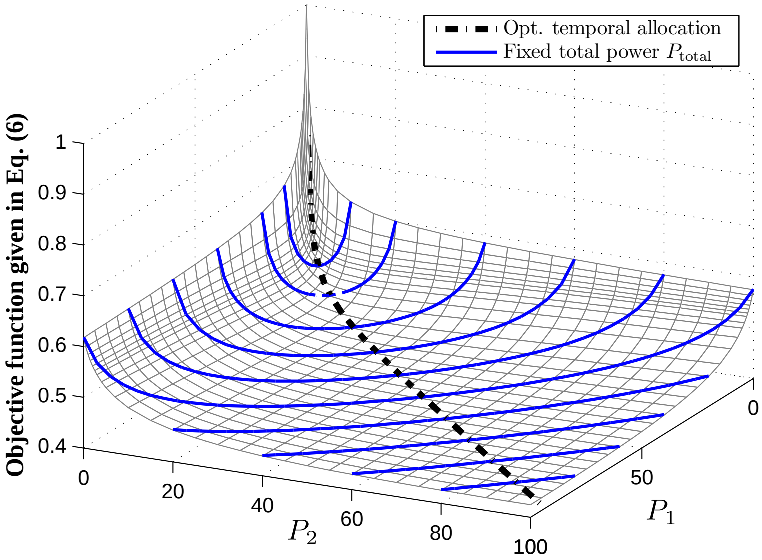


Fig. 29. The 2-dimensional convex optimization for the case of $M = 2$, where $K = 1$, $\widehat{\gamma}_1 = 5$, $\widehat{\gamma}_2 = 3$, $\beta = 0.1$, and $\sigma_e^2 = 0.1$.

regardless of the instantaneous channel realizations.

Unfortunately, the general closed-form solution for Eq. (5.17) turns out to be intractable. However, since the left-hand side of Eq. (5.17) is a monotonically increasing function of $P_{\text{total}}(\boldsymbol{\nu}) \in \mathbb{R}_+$, the solution, if exists, can be easily obtained numerically. Moreover, under a number of special cases, Eq. (5.17) can be solved in closed-form expressions.

a. $\beta \to 0$

When the normalized QoS exponent $\beta \to 0$, Eq. (5.17) becomes a quadratic polynomial of $P_{\text{total}}(\boldsymbol{\nu})$ and can be easily solved in closed-form. In this case, we get the following optimal temporal power allocation policy:

$$P^*_{\text{total}}(\boldsymbol{\nu})\Big|_{\beta\to 0} = \left(\frac{-(2\sigma_e^2+\Sigma_k)+\sqrt{\Sigma_k^2+4\omega^* k\Sigma_k\sigma_e^2(\sigma_e^2+\Sigma_k)}}{2\sigma_e^2(\sigma_e^2+\Sigma_k)}\right)^+$$

which is the optimal temporal power allocation policy given by [25, eq. (17)] (and also [84, eq. (3)] for the case with $M = 1$) to achieve the *ergodic capacity* of the parallel channels with channel estimation errors. This is expected since when $\beta \rightarrow 0$, implying that the system can tolerate an arbitrarily long delay, the optimal effective capacity approaches the ergodic capacity.

b. $\beta \rightarrow \infty$

When the normalized QoS exponent $\beta \rightarrow \infty$, implying stringent delay constraint, Eq. (5.17) becomes a linear function of $P_{\text{total}}(\boldsymbol{\nu})$. The optimal temporal power allocation policy can be easily derived as follows:

$$P^*_{\text{total}}(\boldsymbol{\nu})\Big|_{\beta\rightarrow\infty} = \left(\frac{\eta_k - 1}{\Sigma_k - \sigma_e^2\left(\eta_k - 1\right)}\right)^+ \tag{5.18}$$

where $\eta_k \triangleq (\omega^*)^{M/k}\, k\Sigma_k\Pi_k$. Substituting Eq. (5.18) into Eq. (5.10), the power assigned to each parallel channels can be expressed as

$$P^*_{\pi(m)}(\boldsymbol{\nu})\Big|_{\beta\rightarrow\infty} = \left(\left(\frac{1}{\Sigma_k - \sigma_e^2\left(\eta_k - 1\right)}\right)\left(\frac{\eta_k}{k} - \frac{\Sigma_k}{\widehat{\gamma}_{\pi(m)}}\right)\right)^+. \tag{5.19}$$

The optimal effective capacity approaches the *zero-outage capacity*[2] as $\beta \rightarrow \infty$. Therefore, Eq. (5.19) provides the optimal power allocation policy to achieve zero-outage capacity lower-bound with channel estimation errors. The details about zero-outage capacity and outage minimization will be presented in the next section.

[2] The zero-outage capacity is also termed delay-limited capacity [10, 11].

c. $\sigma_e^2 \to 0$

When the channel estimation is perfect, Eq. (5.17) also becomes a linear function of $P_{\text{total}}(\boldsymbol{\nu})$. Likewise, the power assigned to each parallel channels can be obtained as

$$P^*_{\pi(m)}(\boldsymbol{\nu})\Big|_{\sigma_e^2 \to 0} = \left((\omega^*)^{\frac{M\beta+1}{k\beta+1}} \Pi_k^{\frac{k\beta}{k\beta+1}} - \frac{1}{\widehat{\gamma}_{\pi(m)}} \right)^+ \tag{5.20}$$

which becomes the optimal power allocation for parallel channels under perfect CSI discussed in Chapter V. Moreover, in the limiting cases for loose QoS constraint (i.e., $\beta \to 0$), Eq. (5.20) reduces to the classic water-filling to achieve the ergodic capacity, which is expected, as discussed before. On the other hand, for stringent QoS constraint (i.e., $\beta \to \infty$), Eq. (5.20) reduces to the power allocation policy given in [10, eq. (28)] and [11, eq. (21)] to achieve the zero-outage capacity of the parallel fading channels. This is also expected since when $\beta \to \infty$, implying that the system cannot tolerate any delay, the power allocation needs to be designed to guarantee a zero-outage.

4. Power Allocation With Both Constraints

In this section, we consider the scenario where the system is subject to *both* total power constraint P_{total} and average power constraint P_{avg}. The motivation of this study is the following. First, in practice, the system requires both total power and average power constraints due to hardware limitations. More importantly, as will be seen in the next section, under *only* an average power constraint, the average power does not always converge. When the average power cannot be bounded away from infinity, it is necessary to impose a total power constraint to avoid this divergence. In the following, we assume $P_{\text{total}} \geq P_{\text{avg}}$. Otherwise, P_{avg} is unattainable.

To address the total power constraint, let us define another convex set, denoted by $\mathcal{S}''$, for the temporal power allocation policy as follows:

$$\mathcal{S}'' \triangleq \left\{ P_{\text{total}}(\boldsymbol{\nu}) : P_{\text{total}}(\boldsymbol{\nu}) \in \mathcal{S}', P_{\text{total}}(\boldsymbol{\nu}) \leq P_{\text{total}} \right\}$$

where $\mathcal{S}'$ is defined in Eq. (5.15). It is clear that $\mathcal{S}'' \subseteq \mathcal{S}'$. Now the problem becomes maximizing the objective given in Eq. (5.14) on set $\mathcal{S}''$, instead of on set $\mathcal{S}'$ investigated in Section 3. By the similar procedure used in Section 3, we derive the optimal temporal power allocation under both constraints. The optimal temporal power assigned to each fading state, denoted by $P^*_{\text{both}}(\boldsymbol{\nu})$, is simply a *truncated* version of the power derived from Theorem 3, i.e.,

$$P^*_{\text{both}}(\boldsymbol{\nu}) = \min\left\{P^*_{\text{total}}(\boldsymbol{\nu}), P_{\text{total}}\right\} \tag{5.21}$$

where $P^*_{\text{total}}(\boldsymbol{\nu})$ is obtained by Theorem 3. Accordingly, the water-level coefficient ω^* needs to be recalculated to meet the average power constraint.

D. Performance under Stringent Delay Constraint

When designing the QoS-driven power allocation algorithm, we are more interested in the region where the QoS constraint is stringent. Therefore, in this section, we take a close look at the power allocation performance under stringent delay constraint.

1. Convergence Analyses for the Average Power

As the delay constraint becomes stringent ($\beta \to \infty$), Eq. (5.18) provides the optimal temporal power allocation. However, using Eq. (5.18), the average power may *diverge*. In other words, for a given P_{avg}, we probably cannot find ω^* such that $\mathbb{E}_{\widehat{\boldsymbol{\gamma}}}\left[P^*_{\text{total}}(\boldsymbol{\nu})\right] = P_{\text{avg}}$. In order to guarantee that the average power converges, we need to upper-bound the expectation of Eq. (5.18) away from infinity, which is equivalent to

$$\mathbb{E}_{\widehat{\boldsymbol{\gamma}}}\left[\frac{\eta_k - 1}{\Sigma_k - \sigma_e^2\left(\eta_k - 1\right)}\right]^+ < \infty \tag{5.22}$$

where $\mathbb{E}[x]^+ \triangleq \mathbb{E}\left[x|x \geq 0\right]$. [3] Explicitly characterizing the left-hand side of Eq. (5.22) is hard since k is time-varying depending on $\widehat{\boldsymbol{\gamma}}$. Alternatively, it is more convenient to find the necessary and sufficient conditions for the convergence. The result can be summarized in the following theorem.

Theorem 4. *If $\omega^* > 1$, a necessary condition to guarantee that the average power converges to a finite number is given by*

$$\mathbb{E}_{\widehat{\boldsymbol{\gamma}}}\left[\frac{1}{\widehat{\gamma}_{\pi(1)} - \sigma_e^2(\omega^* - 1)}\right]^+ < \infty, \tag{5.23}$$

while a sufficient condition is given by

$$\mathbb{E}_{\widehat{\boldsymbol{\gamma}}}\left[\frac{\Pi_M}{1 - \sigma_e^2(\omega^*)^M M \Pi_M}\right]^+ < \infty. \tag{5.24}$$

Otherwise, if $\omega^* \leq 1$, then $P_{\text{total}}(\boldsymbol{\nu}) = 0$ always holds, and thus $\mathbb{E}_{\widehat{\boldsymbol{\gamma}}}\left[P_{\text{total}}(\boldsymbol{\nu})\right] = 0$.

Proof. The proof is provided in Appendix K. □

Remark 7. When the channel estimation is perfect ($\sigma_e^2 = 0$), the sufficient condition given by Eq. (5.24) reduces to

$$\mathbb{E}_{\widehat{\boldsymbol{\gamma}}}\left[\Pi_M\right]^+ = \mathbb{E}_{\widehat{\boldsymbol{\gamma}}}\left[\Pi_M\right] < \infty \tag{5.25}$$

which is the condition termed *regular fading* in [11, def. 4] to achieve a positive zero-outage capacity with perfect CSI. Furthermore, when $\sigma_e^2 = 0$, it is easy to show that

[3] If the condition of Eq. (5.22) fulfills, then the average power converges since $P_{\text{avg}} = \mathbb{E}_{\widehat{\boldsymbol{\gamma}}}\left[\frac{\eta_k - 1}{\Sigma_k - \sigma_e^2(\eta_k - 1)}\right]^+ \Pr\left\{\frac{\eta_k - 1}{\Sigma_k - \sigma_e^2(\eta_k - 1)} \geq 0\right\} < \infty$.

Eq. (5.25) implies Eq. (5.23), which is also expected since Eq. (5.23) is a necessary condition.

Remark 8. For most commonly used channel distributions (e.g., Rayleigh, Nakagami, Rice, and Wishart), if $\sigma_e^2 = 0$, the *sufficient* condition given by Eq. (5.25) always fulfilled (additional condition of $M > 1$ may be required). Therefore, the average power always converges. However, if $\sigma_e^2 > 0$, the *necessary* condition given by Eq. (5.23) cannot be fulfilled. Thus, the average power always diverges.

2. Outage Minimization

When $\beta \to \infty$, substituting Eqs. (5.18) and (5.19) into Eq. (5.3) with some algebraic manipulations, the instantaneous spectral efficiency $R/(T_f B)$, denoted by $\mathcal{C}$ (bits/s/Hz), can be obtained as

$$\mathcal{C} = \begin{cases} \frac{M}{K}\log_2(\omega^*), & \text{if } P^*_{\text{total}}(\boldsymbol{\nu}) > 0 \\ 0, & \text{if } P^*_{\text{total}}(\boldsymbol{\nu}) = 0 \end{cases} \tag{5.26}$$

which implies that the transmission is either with a *constant rate* or in an *outage.* If the outage probability is nonzero, we know from definition that the zero-outage capacity of the system is zero. The following lemma describes the impact of channel estimation error on system outage probability.

Proposition 6. *If $\sigma_e^2 > 0$, then the outage probability is nonzero.*

Proof. The proof is provided in Appendix L. □

Remark 9. As long as $\sigma_e^2 > 0$, from Proposition 6 we know $\Pr\{P^*_{\text{total}}(\boldsymbol{\nu}) = 0\} > 0$. Any outage probability smaller than $\Pr\{P^*_{\text{total}}(\boldsymbol{\nu}) = 0\}$ is unattainable. In other words, the probability $\Pr\{P^*_{\text{total}}(\boldsymbol{\nu}) = 0\}$ indicates an *outage floor.* Based on the proof of

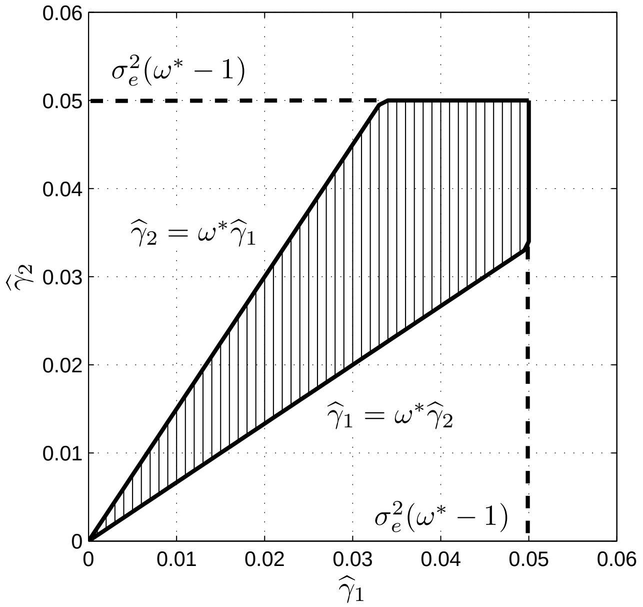


Fig. 30. The constructed outage region (shadowed area) for $M = 2$. The variance $\sigma_e^2 = 0.1$ and water-level coefficient $\omega^* = 1.5$.

Proposition 6 (see Appendix L for details), the constructed outage region is shown by Fig. 30 for the case of $M = 2$.

Corollary 1. *If $\sigma_e^2 > 0$, then the system zero-outage capacity is always zero, regardless of the channel fading distributions.*

Proof. The proof follows from Proposition 6. □

Remark 10. As compared to the case with perfect CSI, where the zero-outage capacity is always *positive* when the channel is regular fading [11], the zero-outage capacity of the system with imperfect CSI is always zero, due to the presence of nonzero σ_e^2. In this case, it makes more sense to study the *outage capacity*, instead of zero-outage capacity.

To transmit at a constant code rate $\mathcal{R}$ (bits/s/Hz), the following theorem provides the optimal power allocation policy that minimizes the outage probability under an average power constraint P_{avg}.

Theorem 5. *The optimal temporal power allocation policy, denoted by $P^*_{\text{out}}(\boldsymbol{\nu})$, that minimizes the outage probability while transmitting at a constant code rate $\mathcal{R}$, can be expressed as*

$$P^*_{\text{out}}(\boldsymbol{\nu}) = \begin{cases} P^*_{\text{total}}(\boldsymbol{\nu}), & \text{if } P^*_{\text{total}}(\boldsymbol{\nu}) \leq s^* \\ 0, & \text{otherwise} \end{cases} \tag{5.27}$$

where $P^*_{\text{total}}(\boldsymbol{\nu})$ is the solution of Eq. (5.18) with $\omega^* = 2^{\mathcal{R}K/M}$, and $s^* \in \mathbb{R}_+$ is a constant chosen such that the average power constraint is satisfied. Based on this policy, the resulting minimum outage probability, denoted by p_{out}, is determined by

$$p_{\text{out}} = \Pr\{P^*_{\text{total}}(\boldsymbol{\nu}) = 0\} + \Pr\{P^*_{\text{total}}(\boldsymbol{\nu}) > s^*\}.$$

Proof. It can be easily observed from Eq. (5.26) that ω^* should be chosen as $\omega^* = 2^{\mathcal{R}K/M}$. The rest of the proof is based on the result of [10], which is omitted for lack of space. □

Remark 11. The parameter s^* has the same role as the total power constraint P_{total} in previous sections. However, the power allocation policies are different. In an effective-capacity maximization problem, when the instantaneous power exceeds P_{total}, the system still use the maximum available power to transmit data [see Eq. (5.21)], avoiding the outage. In contrast, in an outage minimization problem, when the instantaneous power exceeds s^*, the system stops transmitting data to save the transmit power [see Eq. (5.27)], making the system fall into an outage.

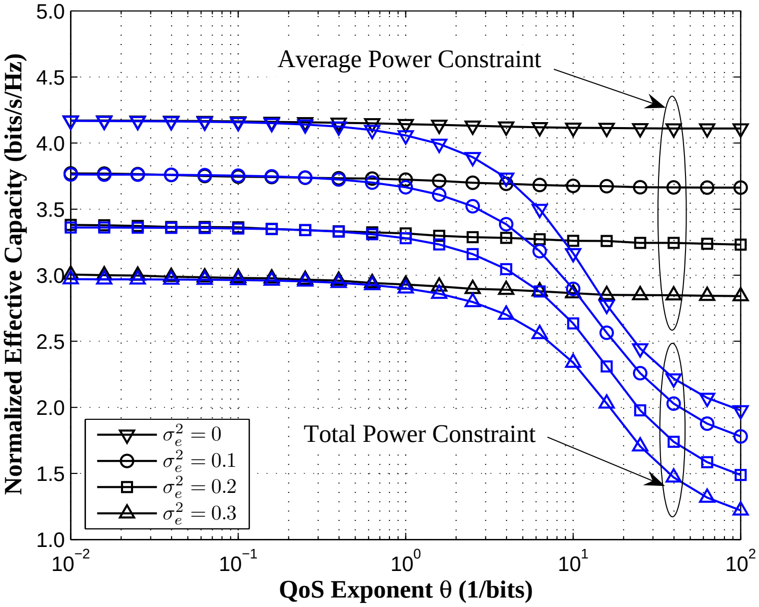


Fig. 31. The optimal effective capacity for a 4 × 4 MIMO system with different power constraints. The average SNR is set equal to 0 dB for both cases.

E. Performance Evaluations

In this section, we evaluate the performance of our proposed QoS-driven power allocation by simulations. As a typical application over parallel Gaussian channels, we simulate the MIMO system with N_t transmit antennas and N_r receive antennas. The channels between all transmit and receive antenna pairs are assumed to be i.i.d. complex Gaussian with $\mathcal{CN}(0, 1)$. In this case, the parameters $K = 1$ and $M = \min\{N_t, N_r\}$. By using the minimum mean squared error (MMSE) estimator at the receiver, the range of the error variance is $0 \leq \sigma_e^2 \leq 1$, and the estimated channels are i.i.d. with $\mathcal{CN}(0, 1 - \sigma_e^2)$. Furthermore, we set the product $T_f B = \log 2$ such that $\theta = \beta$ for convenience. The other system parameters are detailed, respectively, in each of the figures.

Fig. 31 plots the optimal effective capacity of a 4 × 4 MIMO system with dif-

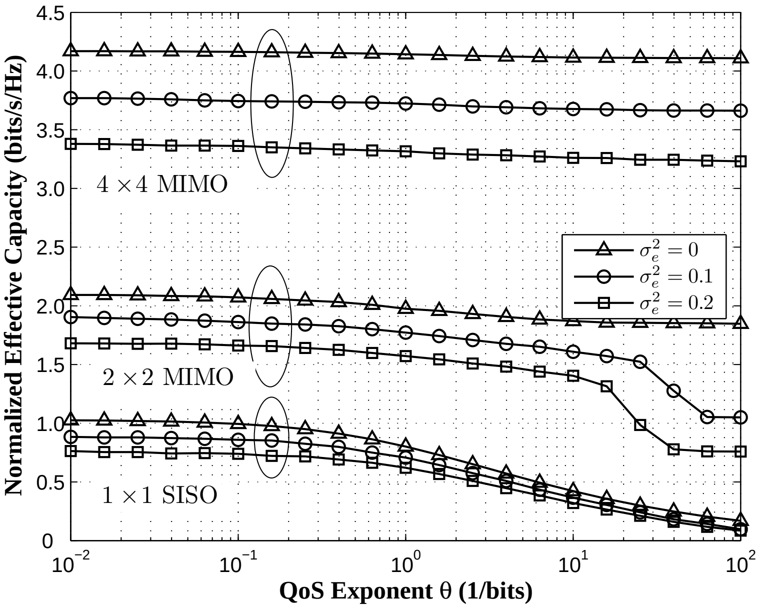


Fig. 32. The optimal effective capacity with different numbers of antennas under both total and average power constraints. The average power constraint $P_{\text{avg}} = 0$ dB and the total power constraint $P_{\text{total}} = 20$ dB.

ferent power constraints. When the QoS constraint is loose, we can observe from Fig. 31 that total power constraint and average power constraint have neglectable performance difference. However, as the QoS constraint becomes more stringent, the average power constraint shows significant performance advantages over the total power constraint. In particular, the effective capacity under the average power constraint virtually does not decrease as θ increases, while the effective capacity under the total power constraint drops quickly as θ increases, which verifies the importance of temporal power allocation on QoS provisioning. On the other hand, the impact of channel estimation error on effective capacity is also significant.

Fig. 32 plots the optimal effective capacity under both average power constraint and total power constraint, when σ_e^2 ranges from 0 to 0.2. As shown by Fig. 32, for the 4×4 MIMO system, the effective capacities are all virtually independent of θ. On the

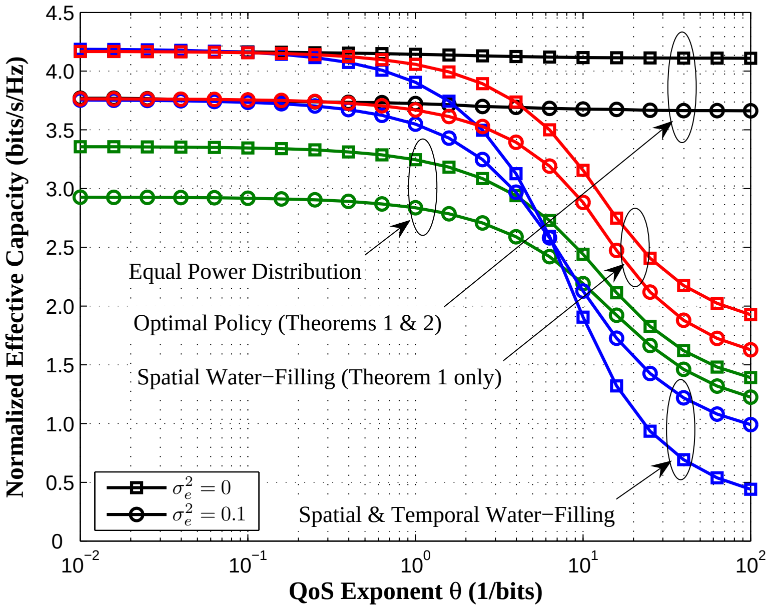


Fig. 33. The optimal effective capacity for a 4 × 4 MIMO system with different power allocation strategies. The average SNR is 0 dB for all the cases.

other hand, for the 2 × 2 MIMO system, the QoS constraint θ does not significantly affect the effective capacity when the channel estimation is perfect ($\sigma_e^2 = 0$). However, when $\sigma_e^2 = 0.1$ or $\sigma_e^2 = 0.2$, the effective capacities significantly decrease as the QoS constraint becomes stringent. Finally, for the 1 × 1 SISO system, all effective capacities converges to zero as θ increases, even when the channel estimation is perfect. Thus, Fig. 32 verifies that a larger number of antennas not only provides the higher throughput, but also offers better robustness against the channel estimation error, in terms of supporting stringent QoS requirements.

Fig. 33 plots the effective capacity under different power allocation strategies. Besides our proposed power allocation with total power constraint (referred as "spatial water-filling") and with average power constraint (referred as "optimal policy"), we also simulate equal power distribution strategy, and joint spatial-temporal water-filling strategy. Note that equal power distribution is the optimal power allocation

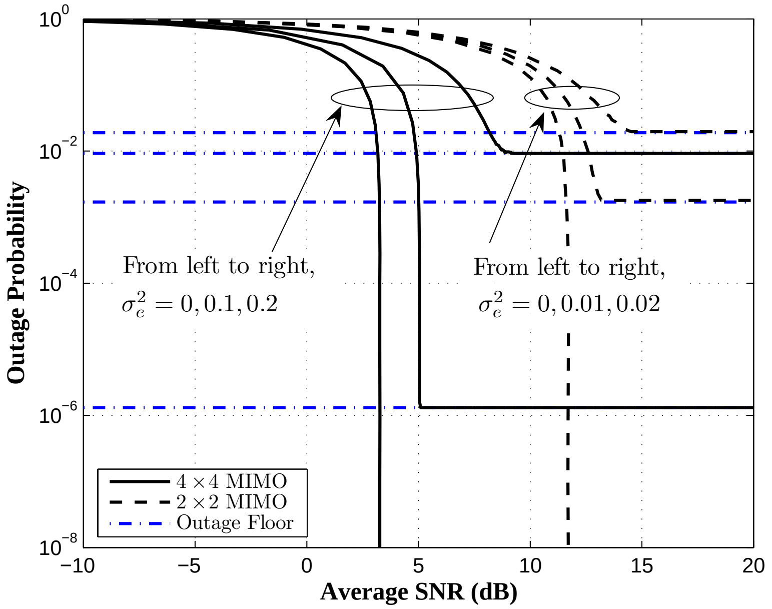


Fig. 34. The outage probability for MIMO system with different channel estimation errors. The code rate $\mathcal{R} = 6$ bits/s/Hz.

without CSI at the transmitter, and joint spacial-temporal water-filling is the optimal power allocation to achieve the ergodic capacity of the MIMO system. We can observe from Fig. 33 that for a given σ_e^2, our proposed optimal policy always achieves the highest effective capacity among all power allocation strategies. The advantage is more significant when the QoS constraint is stringent.

Finally, Fig. 34 plots the outage probability for MIMO systems with different channel estimation errors. Specifically, when the channel estimation is perfect ($\sigma_e^2 = 0$), the outage probability approaches zero when the average SNR is sufficiently high, which means a positive zero-outage capacity is achievable. However, when the channel estimation is imperfect ($\sigma_e^2 > 0$), the outage floor prevents the outage probability from further decreasing, no matter how much power is assigned. Fig. 34 also demonstrates that for a given code rate, the system with a larger number of antennas may tolerate severer channel estimation errors, while still maintaining better outage performance.

F. Summary

We proposed and analyzed QoS-driven power allocation over parallel fading channels by taking the imperfect channel estimations into consideration. Solving the original non-convex problem by a 2-dimensional convex optimization approach, we developed power allocation algorithms for different QoS and power constraints in a general system setting. As the QoS exponent θ increases from zero to infinity, the optimal effective capacity function connects the ergodic capacity with the zero-outage capacity, which is consistent with our previous work in the case of perfect CSI. Our analyses indicate that the imperfect channel estimations have a significant impact on QoS provisioning, especially when the delay constraint is stringent. In particular, a positive zero-outage capacity is unattainable in the presence of channel estimation errors. On the other hand, our simulation results for the MIMO systems also suggest that a larger number of parallel channels can provide higher throughput and more stringent QoS, while offering better robustness against the channel estimation errors.

In Chapters II–V, we studied QoS provisioning problem over a point-to-point communication link. In the next chapter, we turn our attention to multiuser scenario. Specifically, we will apply the derived framework into the downlink cellular wireless networks and propose the corresponding resource allocation scheme.

CHAPTER VI

RESOURCE ALLOCATION FOR CELLULAR NETWORKS

A. Introduction

The diverse quality-of-service (QoS) guarantees for the real-time multimedia transmissions play a critically important role in the next-generation mobile wireless networks. Unlike its wired counterpart networks, supporting the QoS requirement in wireless environment is much more challenging since the time-varying fading channel has the significant impact on the network performance. For wireless QoS guarantees, link adaptation (LA) techniques have been widely considered as the key solution to overcome the impact of the wireless channel. At the physical layer, the most scarce resources are power and spectral-bandwidth. As a result, the LA techniques such as adaptive modulation and power control are developed to enhance the spectral efficiency while maintaining a certain target error performance [64]. However, for real-time wireless multimedia services, the main QoS metric is bounded-delay, instead of high spectral efficiency [89]. Therefore, to support the real-time wireless multimedia QoS, we need to consider the LA techniques not only at the physical-layer, but also at the upper-protocol-layers such as data-link layer when designing the wireless networks. To achieve this goal, in this chapter we develop the cross-layer-model based adaptive resource-allocation scheme to support the real-time multimedia QoS in the downlink heterogeneous mobile wireless networks.

QoS provisioning in wireless networks has been widely studied from different perspectives, such as packet scheduling, admission control, traffic specifications, resource reservations, etc. [21, 22, 27, 69, 89–93]. In [89] and [90], the authors investigated the real-time and non-real-time QoS provisioning for code-division-multiple-access

(CDMA)-based wireless networks. In [91–93], several architectures/algorithms were discussed for either implicit or explicit QoS provisioning. In [27, 69], the authors integrated the finite-state Markov chain (FSMC) with adaptive modulation and coding (AMC), and then jointly considered the physical-layer channel and data-link-layer queuing characteristics. The idea of resource allocation in [27, 69] is to calculate the reserved bandwidth for each user by appropriate admission control and scheduling. This scheme is developed across the physical-layer and data-link-layer and is thus capable of characterizing the impact of physical-layer variation on the data-link-layer QoS performance. However, the main QoS requirement addressed in [27, 69] is the *average delay* of the wireless transmission, which does not effectively support the real-time multimedia services, where the key QoS metric is the *bounded delay*, as addressed in this chapter.

In [21, 22], the authors proposed a powerful concept termed “effective capacity”. This concept turns out to be the *dual problem* of the so-called “effective bandwidth”, which has been extensively studied in the early 90’s in the contexts of wired asynchronous transfer mode (ATM) networks. The effective capacity and effective bandwidth enable us to analyze the *statistical* delay-bound violation and buffer-overflow probabilities, which are critically important for multimedia wireless networks. Based on [21], the authors in [22, 23] proposed a set of resource-allocation schemes for statistical QoS guarantees in wireless networks. The key techniques used in [22, 23] are the integration of effective capacity with multiuser diversity [94], such that the scheme not only provides the statistical QoS for different mobile users, but also increases the total wireless-network’s throughput. However, the effective capacity approach has not been explored in cross-layer modeling and design for adaptive resource allocation and QoS guarantees in mobile wireless networks.

To overcome the aforementioned problems, in this chapter we propose a cross-layer-model based adaptive resource-allocation scheme for downlink heterogeneous mobile wireless networks. Based on our application of the effective capacity method in Chapter II, the system resources are allocated according to the heterogeneous fading channel statistics, the diverse QoS requirements, and different traffic characteristics. Specifically, our scheme adaptively assigns power-level and time-slots for real-time mobile users in a dynamic time-division multiple access (TDMA) mode to guarantee the bounded delays. We analytically derive the admission-control and power/time-slot allocation conditions to guarantee the *statistical* delay-bound for real-time mobile users. In this chapter, we do not employ multiuser diversity because of the following reasons. In a *centralized heterogenous* multiuser network, the multiuser diversity will cause the serious fairness problem — the users with good channels may occupy most of the resources, while the users with poor channels may hardly have opportunity for information transmission, which will result in large queueing delay and thus the user's delay-bound QoS cannot be guaranteed. On the other hand, the advantages of multiuser diversity only contribute to a small portion of mobile users whose channel quality is good, which may not lead to a significant QoS performance improvement from the entire network perspectives. Note that in [95], the authors proposed to use multiuser diversity under the "proportional fairness" constraint. However, this scheme can only support a *loose* delay-bound QoS requirements, which is also not suitable for real-time multimedia services where the delay-bound QoS requirement is *stringent*.

When designing the adaptive resource-allocation algorithm, we address the problems of the physical-layer impact on the statistical QoS provisioning performance. Specifically, we study how adaptive power-control and channel-state information (CSI) feedback delay influence our proposed scheme. Based on the results in Chapter III, we apply our proposed *QoS-driven power adaptation* for heterogeneous mobile users and compare its performance with conventional water-filling and constant power schemes. Our numerical and simulation results show that our proposed QoS-driven power control has significant advantages over the conventional power controls in terms of QoS-guarantees. On the other hand, our effective-capacity-based adaptive resource-allocation algorithm can efficiently support the QoS requirements for diverse real-time mobile users. In an in-door mobile environment, e.g., the widely used wireless local-area networks (WLAN), the proposed algorithm also provides sufficient robustness to the CSI feedback delay.

The rest of the chapter is organized as follows. Section B describes our system model. Section C develops the admission control and time-slot allocation algorithm with fixed average transmission-power. Section D proposes the joint power-level and time-slot allocation scheme. Section E analyzes the impact of feedback delay on the proposed scheme. The chapter concludes with Section F.

B. System Model

The system model is shown in Fig. 35. In this chapter, we concentrate on single-input-single-output (SISO) antenna system with the downlink transmission from the basestation to the mobile users. We denote the total number of mobile users by K, the total spectral-bandwidth of the system by B, and the average transmission-power of the basestation by $\overline{P}$, respectively. We first assume that the average transmission power $\overline{P}$ is fixed. In Section D, we will remove this constraint and let $\overline{P}$ vary within a discrete set. The K users are assumed to be heterogenous, i.e., they may experience different fading conditions and demand different QoS requirements.

As shown by Fig. 35, the upper-protocol-layer packets are first divided into a number of frames at data-link layer. The frames are stored at the transmitter infinite-

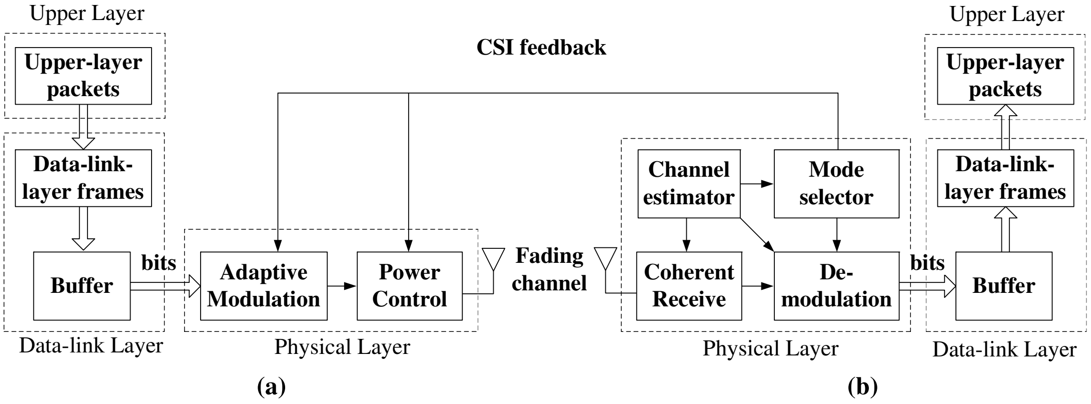


Fig. 35. The system model of downlink cellular wireless networks. (a) Basestation transmitter. (b) The kth mobile receiver.

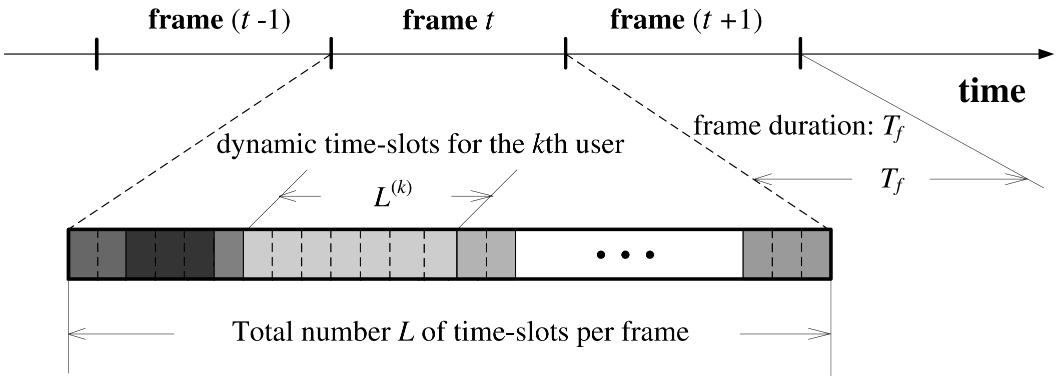


Fig. 36. The frame structure of the proposed system.

buffer and then split into bit-streams at physical layer, where the adaptive-modulation and power-control are employed, respectively, to enhance the system performance. The reverse operations are executed at the receiver side. Also, the CSI is estimated at the receiver and fed back to the transmitter for adaptive modulation and adaptive power-control, respectively.

1. Data-Link Layer Frame Structure

The frame structure of our proposed system is shown by Fig. 36. In our system, each frame at data-link layer consists of L number of time-slots. The time-duration of each frame is denoted by T_f. Due to the employment of adaptive modulation, the number of bits per frame varies depending on each user's modulation modes selected. As shown in Fig. 36, within the frame duration T_f, the system runs in a dynamic TDMA mode. The kth mobile user is assigned with a number $L^{(k)}$ of time-slots. The number $L^{(k)}$ is determined by the kth mobile user's QoS requirement, which will be detailed in Section C. Clearly, we have $\sum_{k=1}^{K} L^{(k)} \leq L$.

2. Channel Model

We assume that the wireless fading channel is flat-fading with Nakagami-m distribution. The fading statistics of different mobile users are independent of each other. In this section, we omit the user index k for simplicity. The probability density function (pdf) of the signal-to-noise ratio (SNR), denoted by $p_\Gamma(\gamma)$, can be expressed as [34]

$$p_\Gamma(\gamma) = \frac{\gamma^{m-1}}{\Gamma(m)} \left(\frac{m}{\overline{\gamma}}\right)^m \exp\left(-\frac{m}{\overline{\gamma}}\gamma\right), \ \gamma \geq 0 \tag{6.1}$$

where $\Gamma(\cdot)$ represents the complete Gamma function, m denotes the fading parameter of Nakagami-m distribution, and $\overline{\gamma}$ denotes the average SNR of the combined signal, which can be expressed as $\overline{\gamma} = \overline{P}E\{\alpha^2\}/(N_0 B)$, where $E\{\alpha^2\}$ is the average path-gain of the Nakagami fading channel and N_0 is the single-sided power spectral density (PSD) of the complex additive white Gaussian noise (AWGN). Note that when the average power-level $\overline{P}$ varies, the corresponding average SNR $\overline{\gamma}$ will change accordingly.

The channel is assumed to be invariant within a frame's time-duration T_f, but varies from one frame to another. Furthermore, we assume that the CSI is perfectly estimated at the receiver and reliably fed back to the transmitter with a time-delay denoted by τ. First, we assume $\tau = 0$, implying the perfect CSI feedback. We will address the scenario with delayed CSI feedback in Section E.

3. Adaptive Modulation

Adaptive modulation is an efficient LA technique to improve the spectral-efficiency at physical layer. In this chapter, we employ the adaptive QAM modulation proposed in [64]. The specific modulation modes for the adaptive-modulation scheme are constructed as follows. We partition the entire SNR range by N non-overlapping consecutive intervals, resulting in $N+1$ boundary points denoted by $\{\Gamma_n\}_{n=0}^{N}$, where $\Gamma_0 < \Gamma_1 < \cdots < \Gamma_N$ with $\Gamma_0 = 0$ and $\Gamma_N = \infty$. Correspondingly, the adaptive modulation is selected to be in mode n if the SNR, denoted by γ, falls into the range of $\Gamma_n \leq \gamma < \Gamma_{n+1}$. The zero-th mode corresponds to the "outage" mode of the system, i.e., the transmitter stops transmitting data in Mode 0. The constellation used for the nth mode is M_n-QAM, where $M_n = 2^n$ with $n \in \{1, 2, ..., N-1\}$. Let us further define $M_0 = 0$ and $M_N = \infty$. Thus, the spectral-efficiency of the adaptive modulation ranges from 0 to $N-1$ bits/sec/Hz. As the SNR increases, the system selects the mode with higher spectral-efficiency to transmit data. On the other hand, as the SNR gets worse, the system decreases the transmission rate to adapt to the degraded channel conditions. In the worst case, the transmitter stops transmitting data as in the "outage" mode.

The bit-error rate (BER) when using the nth mode for $n \in \{1, 2, ..., N-1\}$, denoted by BER_n, can be approximated as follows [64]:

$$\mathrm{BER}_n \approx 0.2 \exp\left(-g_n \gamma\right) \tag{6.2}$$

where $g_n = 3/[2(M_n - 1)]$. Based on the pdf given in Eq. (6.1), the probability π_n, that the SNR falls into mode n is determined by

$$\pi_n = \int_{\Gamma_n}^{\Gamma_{n+1}} p_\Gamma(\gamma) d\gamma = \frac{\Gamma\left(m, \frac{m}{\overline{\gamma}}\Gamma_n\right)}{\Gamma(m)} - \frac{\Gamma\left(m, \frac{m}{\overline{\gamma}}\Gamma_{n+1}\right)}{\Gamma(m)} \tag{6.3}$$

where $\Gamma(\cdot, \cdot)$ represents the incomplete Gamma function and $n \in \{0, 1, ..., N-1\}$.

In general, the forward-error control (FEC) and automatic retransmission request (ARQ) are also employed at the physical/data-link layer. However, in this chapter we only focus on uncoded system due to the following reasons. First, there exist the simple *analytical* power-control policies [53, 64] for uncoded transmissions, while for coded transmission, it is difficult to find such a policy. Thus, we assume uncoded transmission for analytical convenience. Second, based on our study in Chapter II, we observe that the performance trends of FEC/ARQ-based transmission is similar to that of uncoded systems, as long as the link BER is not too high. Therefore, the investigation of the uncoded system also provides a guideline on designing the coded system.

4. Power Control

We mainly investigate three different power-control strategies, namely, our proposed QoS-driven power control in Chapter III, the water-filling power control, and the constant-power approach. For different power-control strategies, the power-control law as well as the boundary points $\{\Gamma_n\}_{n=1}^{N-1}$ are different. We study how to adjust the power and decide the boundary points for the above three power-control strategies, respectively, as follows.

Strategy I: QoS-Driven Optimal Power Control. In Chapter III, we develop the QoS-driven optimal power-control strategy for the adaptive QAM modulation. Let the BER QoS requirement of the system be denoted by $\mathrm{P}_{\mathrm{tgt}}$. In order to achieve the target BER, i.e., $\mathrm{P}_{\mathrm{tgt}}$, the power-control law, denoted by $\mu_n(\gamma)$, for the nth mode can be derived as

$$\mu_n(\gamma) = \begin{cases} (M_n - 1)\dfrac{1}{\nu_n \gamma}, & M_n \le \frac{\gamma}{\gamma_0} < M_{n+1},\ (n \neq 0) \\ 0, & \frac{\gamma}{\gamma_0} < M_1,\ (n = 0) \end{cases} \tag{6.4}$$

where $\nu_n = -1.5/\log(5\mathrm{P}_{\mathrm{tgt}})$ and γ_0 is the cut-off threshold, which can be numerically obtained by meeting the following mean power constraint:

$$\sum_{n=1}^{N-1} \int_{\Gamma_n}^{\Gamma_{n+1}} \mu_n(\gamma) p_{\Gamma}(\gamma) d\gamma = 1 \tag{6.5}$$

where we have

$$\Gamma_n = \gamma_0 M_n^{\frac{\kappa T_f B \theta}{\log 2}+1} \tag{6.6}$$

where θ is QoS-exponent [21], and $\kappa \ge 1$ is a parameter to deal with the impact of channel correlation. Specifically, when the channel process is uncorrelated (i.e., block fading channel), then we have $\kappa = 1$. Otherwise, when the channel process is correlated, κ should be adjusted according to the channel Doppler frequency f_d. Once the cut-off threshold γ_0 is determined, the boundary points $\{\Gamma_n\}_{n=1}^{N-1}$ can be obtained by using Eq. (6.6). The QoS-driven power control makes the BER of each mode equal to $\mathrm{P}_{\mathrm{tgt}}$. Then, the resulting system BER is also equal to $\mathrm{P}_{\mathrm{tgt}}$.

Strategy II: Water-Filling Power Control. In [64], the authors proposed the optimal power-control strategy for adaptive MQAM that can maximize the spectral-efficiency, which is actually based on the time-domain water-filling algorithm. However, based on our study in Chapter III, we find that the water-filling power control can be considered as a special case of our proposed QoS-driven power control by letting the QoS exponent $\theta \to 0$. Thus, the power-control law and mean power constraint of the water-filling scheme are the same as those given by Eqs. (6.4) and (6.5), respectively. The boundary points are determined by

$$\Gamma_n = \lim_{\theta \to 0} \gamma_0 M_n^{\frac{\kappa T_f B\theta}{\log 2}+1} = \gamma_0 M_n. \tag{6.7}$$

Strategy III: Constant-Power Approach. Constant power-control approach is to keep the transmission power at the basestation as a constant. Using Eqs. (6.1) and (6.2), the average BER of the mode n, denoted by $\overline{\mathrm{BER}}_n$, can be derived as

$$\begin{aligned}\overline{\mathrm{BER}}_n &= \frac{1}{\pi_n}\int_{\Gamma_n}^{\Gamma_{n+1}} 0.2\exp(-g_n\gamma)p_\Gamma(\gamma)d\gamma \\ &= \frac{0.2\left(\frac{m}{b_n}\right)^m}{\pi_n\Gamma(m)}\left[\Gamma\left(m, \frac{b_n\Gamma_n}{\overline{\gamma}}\right) - \Gamma\left(m, \frac{b_n\Gamma_{n+1}}{\overline{\gamma}}\right)\right] \end{aligned} \tag{6.8}$$

where $b_n = g_n\overline{\gamma} + m$ for $n \in \{1, 2, ..., N-1\}$ and the boundary points are determined by

$$\Gamma_n = \frac{\eta}{g_n} \tag{6.9}$$

where the parameter η ($\eta > 0$) in Eq. (6.9) is numerically obtained by meeting the following constraint on the average BER requirement $\mathrm{P}_{\mathrm{tgt}}$:

$$\mathrm{P}_{\mathrm{tgt}} = \frac{\sum_{n=1}^{N-1} n\pi_n\overline{\mathrm{BER}}_n}{\sum_{n=1}^{N-1} n\pi_n}. \tag{6.10}$$

where $\overline{\mathrm{BER}}_n$ is the function of η through Eqs. (6.8) and (6.9). Once the parameter η is determined, the boundary points $\{\Gamma_n\}_{n=1}^{N-1}$ can be obtained by using Eq. (6.9).

5. Service Process Modeling by Using FSMC

In this chapter, we employ the FSMC model to characterize the variation of the wireless service process. Each state of FSMC corresponds to a mode of the adaptive-modulation scheme. Let $p_{i,j}$ denote the transition probability from state i to state

j. We assume a slow-fading channel model such that the transition only happens between adjacent states [65]. Under such an assumption, we have $p_{ij} = 0$ for all $|i-j| > 1$. The adjacent transition probability can be approximated as [65]

$$\begin{cases} p_{n,n+1} \approx \frac{N_\Gamma(\Gamma_{n+1})T_f}{\pi_n}, \text{ where } n = 0, 1, ..., N-2, \\ p_{n,n-1} \approx \frac{N_\Gamma(\Gamma_n)T_f}{\pi_n}, \quad \text{where } n = 1, 2, ..., N-1 \end{cases} \tag{6.11}$$

where $N_\Gamma(\gamma)$ is the level-crossing rate (LCR) determined by SNR of γ, which is given by [34]

$$N_\Gamma(\gamma) = \frac{\sqrt{2\pi} f_d}{\Gamma m} \left(\frac{m\gamma}{\overline{\gamma}}\right)^{m-\frac{1}{2}} \exp\left(-\frac{m\gamma}{\overline{\gamma}}\right) \tag{6.12}$$

where f_d is the maximum Doppler frequency of the mobile user. Then, the remaining transition probabilities can be derived by using Eq. (6.11) as follows:

$$\begin{cases} p_{0,0} = 1 - p_{0,1} \\ p_{N-1,N-1} = 1 - p_{N-1,N-2} \\ p_{n,n} = 1 - p_{n,n-1} - p_{n,n+1}, \; n = 1, ..., N-2. \end{cases} \tag{6.13}$$

Applying Eqs. (6.11) and (6.13), we obtain the probability transition matrix of the FSMC, denoted by $\mathbf{P} = [p_{ij}]_{N\times N}$. Correspondingly, we obtain the stationary distribution of the FSMC, denoted by $\boldsymbol{\pi}$, as follows:

$$\boldsymbol{\pi} = [\, \pi_0, \pi_1, ...\pi_{N-1} \,] \tag{6.14}$$

where π_n is given by Eq. (6.3) for $n \in \{0, 1, ..., N-1\}$.

C. Adaptive Resource Allocation With Fixed Average Power

The cross-layer modeling introduced in Chapter II establishes the analytical framework to investigate the impact of physical-layer infrastructure variations on the statistical QoS provisioning performance at the data-link-layer through the effective capacity function. In this section, we develop the adaptive resource-allocation algorithms based on our developed cross-layer model to guarantee the desired QoS requirements. Since our focus is mainly on resource allocation in this chapter, we only adopt the simple round-robin (RR) scheduling for the real-time mobile users.

1. The Effective Capacity of the Service Process

As described in Section 1, our proposed system operates in a dynamic TDMA mode. As shown in Fig. 36, the kth user is assigned with $L^{(k)}$ of time-slots per frame for information transmission. In order to determine the number $L^{(k)}$ of time-slots allocated to the kth user to support its statistical QoS, we first need to derive the effective capacity of the service-process. Consider only allocating $L^{(k)} = 1$ time-slot as a basic-unit to the kth user, the effective capacity of the kth user, denoted by $\mathrm{E}_{\mathrm{C}}^{(k,1)}(\theta)$, can be expressed as

$$\mathrm{E}_{\mathrm{C}}^{(k,1)}(\theta) = -\frac{1}{\theta}\log\Big(\rho\big\{\mathbf{P}^{(k)}\,\mathbf{\Phi}^{(1)}(\theta)\big\}\Big),\ \theta > 0 \tag{6.15}$$

where $\mathbf{P}^{(k)}$ is the transition probability matrix of the kth user, which is determined by the kth user's channel statistics and is independent of $L^{(k)}$, and $\mathbf{\Phi}^{(1)}(\theta)$ is given by $\mathbf{\Phi}^{(1)}(\theta) = \mathrm{diag}\left\{e^{-\lambda_0^{(1)}\theta}, e^{-\lambda_1^{(1)}\theta}, ..., e^{-\lambda_{N-1}^{(1)}\theta}\right\}$, where $\lambda_n^{(1)} = nT_f B/L,\ n \in \{0, 1, ..., N-1\}$, which is independent of the channel statistics.

When allocating $L^{(k)} = l$ time-slots for the user, applying the results developed in [22], the effective capacity, denoted by $\mathrm{E}_{\mathrm{C}}^{(k,l)}(\theta)$, can be expressed as

$$\mathrm{E}_{\mathrm{C}}^{(k,l)}(\theta) = l\mathrm{E}_{\mathrm{C}}^{(k,1)}(l\theta). \tag{6.16}$$

2. Admission-Control and Time-Slot Allocation

Let the kth user's statistical QoS requirement be denoted by $\{D_{\max}^{(k)}, \varepsilon^{(k)}\}$, where $D_{\max}^{(k)}$ is the delay-bound and $\varepsilon^{(k)}$ is the violation probability. Similar to the procedure described in Chapter II, the time-slot allocation algorithms can be designed in the following steps.

S1: Denote the effective bandwidth of the kth user's arrival-process by $\mathrm{E}_{\mathrm{B}}^{(k)}(\theta)$. Find the solution of the rate and QoS-exponent (δ_l, θ_l) such that $\mathrm{E}_{\mathrm{B}}^{(k)}(\theta_l) = \mathrm{E}_{\mathrm{C}}^{(k,l)}(\theta_l) = \delta_l$.

S2: Using $L^{(k)} = l$ number of time-slots, the delay-bound violation probability can be derived as

$$\Pr\{\text{Delay} > D_{\max}^{(k)}\} \approx \exp\left(-\theta_l \delta_l D_{\max}^{(k)}\right) \tag{6.17}$$

S3: The number $L^{(k)}$ is determined by

$$L^{(k)} = \min_{1 \le l \le L} \{\, l \,\}, \ \text{s.t.}\ \exp\left(-\theta_l \delta_l D_{\max}^{(k)}\right) \le \varepsilon^{(k)}. \tag{6.18}$$

For each real-time user, $L^{(k)}$ can be calculated using Eq. (6.18). Clearly, the total number of time-slots that are allocated to the real-time users needs to satisfy the following equation:

$$\sum_{k=1}^{K} L^{(k)} \le L. \tag{6.19}$$

When a new mobile real-time user applies to join the system, the admission-control algorithm examines if the number of available time-slot resources is sufficient to support the new real-time mobile user's statistical QoS. If yes, the new real-time mobile user is admitted to join the system; otherwise, this new real-time mobile user is rejected to join the system.

Table II. QoS Requirements for Audio and Video Services.

	BER $\mathrm{P}_{\mathrm{tgt}}$	Delay-bound $D_{\max}$	Violation Prob. ε
Audio	10^{-3}	50 ms	10^{-2}
Video	10^{-4}	150 ms	10^{-3}

3. Numerical and Simulation Results

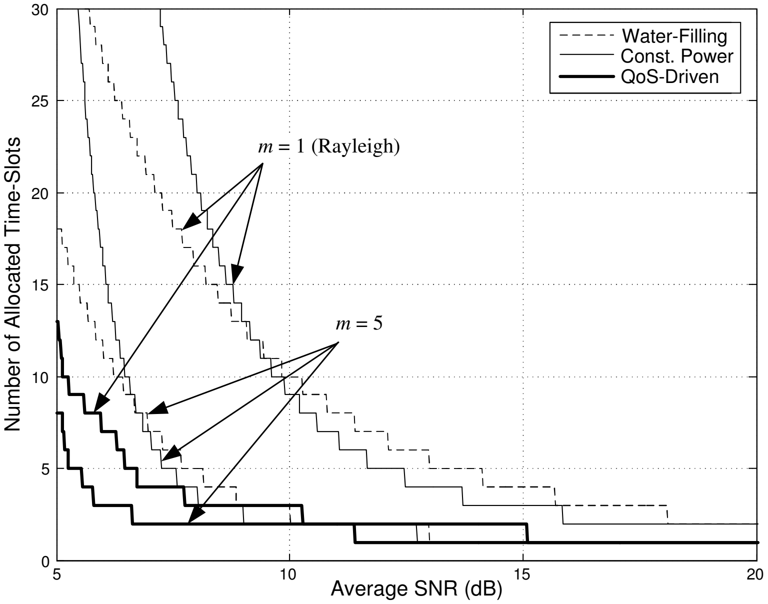
We evaluate the proposed time-slot allocation algorithms through numerical solutions and simulations. In the following, we set the number of adaptive-modulation modes $N = 8$, the total system spectral-bandwidth $B = 1000$ KHz, the data-link-layer frame time-duration $T_f = 2$ ms, the number of time-slots per frame $L = 100$, and the maximum Doppler frequency $f_d = 15$ Hz. Moreover, we generate two types of real-time services. The first type simulates the low speed audio service, where we model the arrival traffic by the well-known ON-OFF fluid model. The holding times in "ON" and "OFF" states are exponentially distributed with the mean equal to 8.9 ms and 8.4 ms, respectively. The "ON" state traffic is modeled as a constant-rate of 32 Kbps. The second one simulates a high-speed video traffic flow. We employ a first-order auto-regressive (AR) process to simulate video traffic characteristics [70], the bit-rate of which can be expressed as

$$\nu(t) = a\nu(t-1) + bw \tag{6.20}$$

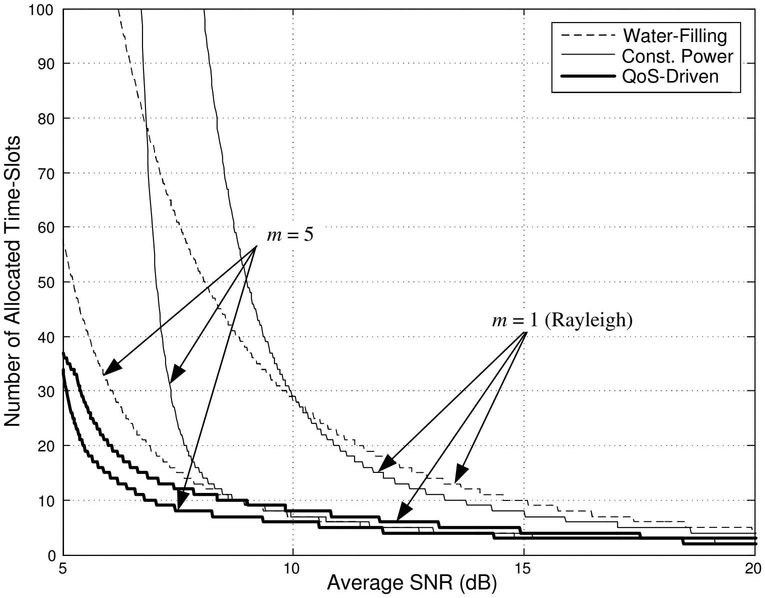
where $a = 0.8781$, $b = 0.1108$ [70] and w is a Gaussian random variable with the mean 80 Kbps and the standard deviation of 30 Kbps. The effective bandwidth of the audio and video traffic is derived according to [1] and [18], respectively. The QoS requirements of these two types of services are shown in Table II.

Using the time-slot allocation algorithm proposed in Section 2, Fig. 37 shows the numerical results of allocated time-slots for audio and video services as a function of the average SNR. As shown by Fig. 37, for both audio and video services, the required time-slots for supporting the QoS decreases as the average SNR increases. The better quality channel (fading parameter $m = 5$) needs the fewer number of time-slots than the Rayleigh fading channel (fading parameter $m = 1$). When the SNR is low, the time-slot allocation algorithms may not find the feasible solution of the $L^{(k)}$ to support the required QoS, since $L^{(k)}$ must satisfy $1 \leq L^{(k)} \leq L$. From Fig. 37 we can also observe that our proposed QoS-driven power control has significant superiorities over both the conventional water-filling scheme and constant power approach.

To evaluate whether the allocated time-slots can support the required statistical QoS, Fig. 38 plots the simulated delay-bound violation probabilities for video and audio services using our proposed QoS-driven power control. We can obverse from Fig. 38 that for both audio and video services the delay-bound violation probabilities are below the required upper-bounds ε's. The simulated delay-bound violation probability is lower than the designated delay-bound violation probability ε. Interestingly, Fig. 38 shows that the QoS-violation probability *fluctuates* according to the time-slot allocation outcomes, which is because our time-slot allocation results vary within a discrete set. For the conventional water-filling scheme and constant power approach, we can observe the similar delay-bound violation probability performance, which is omitted for lack of space. Note that the conventional power control schemes achieve the similar QoS violation performance by using much more resources (i.e., time-slots, see Fig. 37) than our proposed QoS-driven power control scheme.



(a) Audio time-slot allocation.



(b) Video time-slot allocation.

Fig. 37. The numerical time-slot allocation for audio and video services.

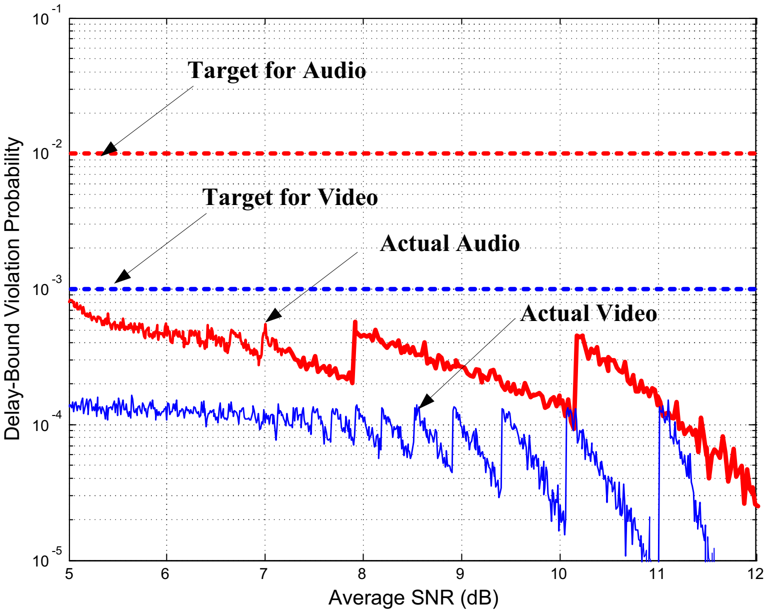


Fig. 38. Simulation results of the delay-bound violation probability for QoS-driven power control. The fading parameter $m = 1$ (Rayleigh fading channel).

D. Joint Power-Level and Time-Slot Allocation

1. Power-Level and Time-Slot Allocation Using Dynamic Programming

In previous sections, we assume that the average transmission power $\overline{P}$ at the basestation transmitter is fixed. In this section, we remove this constraint and let the average transmission power vary within a discrete set. In fact, setting the initial power-level has already been adopted in, e.g., UMTS 3GPP standard [96] for cellular networks. However, in [96] it does not mention how to adjust the power-level to guarantee the QoS requirement. In this chapter, the idea of joint power-level and time-slot allocation can be described as follows. To guarantee the kth user's QoS requirement, the basestation may assign a larger number of time-slots while using a lower power-level; it is also possible to allocate a fewer number of time-slots while using a higher power-level. The goal of our proposed joint power-level and time-slot allocation algorithm is to assign each user with time-slots and power-levels such that the user's QoS requirement is guaranteed while minimizing the total transmission energy. Thus, when the number of users is large or the channel quality is poor, the basestation can increase its transmission power-level to admit more mobile users. On the other hand, when the number of mobile users is small or the channel quality is good enough, the basestation can decrease its transmission power-level while still guaranteeing the desired QoS requirements. In a multi-cell wireless networks, e.g., the cellular networks, this will not only save the power resources at the basestation, but also generate less interference to the other cells.

It is clear that under current problem formulation, we can also use different power-control policies *for each given power-level*. However, in this section, we will only focus on our proposed QoS-driven power control, since this scheme offers the optimal performance. Let the set of the discrete average power-levels be denoted by $\mathcal{P} = \{\overline{P}_1, \overline{P}_2, ..., \overline{P}_I\}$, where $0 < \overline{P}_1 < \overline{P}_2 < ... < \overline{P}_I$. Moreover, let $\overline{P}_k(L^{(k)})$ denote the minimum power-level that is required to guarantee the kth user's QoS requirement when allocating $L^{(k)}$ time-slots to the mobile user. Then, the problem of our dynamic resource-allocation can be formulated as follows:

$$\textbf{Objective:} \qquad \min\left\{\sum_{k=1}^{K} L^{(k)} P_k(L^{(k)})\right\} \tag{6.21}$$

subject to:

$$\begin{cases} 1 \leq L^{(k)} \leq L, \ \forall k \in \{1, 2, ..., K\} \\ \sum_{k=1}^{K} L^{(k)} \leq L \end{cases} \tag{6.22}$$

where

$$\overline{P}_k(L^{(k)}) = \min\left\{\overline{P} \in \mathcal{P} \,\middle|\, \exp\left(-\theta_{L^{(k)}}\delta_{L^{(k)}} D_{\max}^{(k)}\right) \leq \varepsilon^{(k)}\right\}. \tag{6.23}$$

To obtain the feasible solutions of the time-slots $L^{(k)}$ and the power-level $\overline{P}_k(L^{(k)})$, let us consider the procedure illustrated in Fig. 39. Given a time-slot-allocation table obtained from Section C (e.g., Fig. 37), we can partition the average-SNR range by a number of consecutive intervals, with each interval corresponding to a power-level. At the range where the average SNR is too low (as shown by the shaded-area in the left-hand-side of Fig. 39), there is no feasible solution of $L^{(k)}$ due to the constraint of Eq. (6.22) that $L^{(k)}$ must satisfy $L^{(k)} \leq L$. On the other hand, at the range where the average SNR is too large (as shown by the shaded-area on the right-hand-side of Fig. 39), there is no feasible solution of $\overline{P}_k(L^{(k)})$ due to the condition of Eq. (6.23) that $\overline{P}_k(L^{(k)})$ must satisfy $\overline{P}_k(L^{(k)}) \leq \overline{P}_I$. At the range in between, each average SNR-interval is achieved by using certain power-level $\overline{P}_i$, where $i \in \{1, 2, ..., I\}$. Then, for a given $L^{(k)}$, the required power-level $\overline{P}_k(L^{(k)})$ can be obtained by mapping $L^{(k)}$ into the corresponding SNR-interval. For example, for the case shown in Fig. 39, the power-level $\overline{P}_k(L^{(k)})$ falls into the SNR-interval belonging to $\overline{P}_2$ (as shown by the shaded-area in the middle of Fig. 39). Therefore, the required minimum power-level is $\overline{P}_k(L^{(k)}) = \overline{P}_2$.

Once $\overline{P}_k(L^{(k)})$ is attained, this minimization problem can be solved by the dynamic programming (DP) approach [48]. Let us define $u_k(l) \triangleq l\overline{P}_k(l)$, where $l = 1, 2, ..., L$. The cost function of the first mobile user, denoted by $\mathcal{J}_1(l)$, can be expressed as

$$\mathcal{J}_1(l) = u_1(l). \tag{6.24}$$

Then, the cost function for the kth mobile user can be derived iteratively as:

$$\mathcal{J}_k(l) = \min_{1 \leq t \leq l-1} \Big\{ u_k(t) + \mathcal{J}_{k-1}(l-t) \Big\}, \text{ for } k \leq l \leq L \tag{6.25}$$

where $k = 2, 3, ..., K$. The resource-allocation algorithm is executed every time when

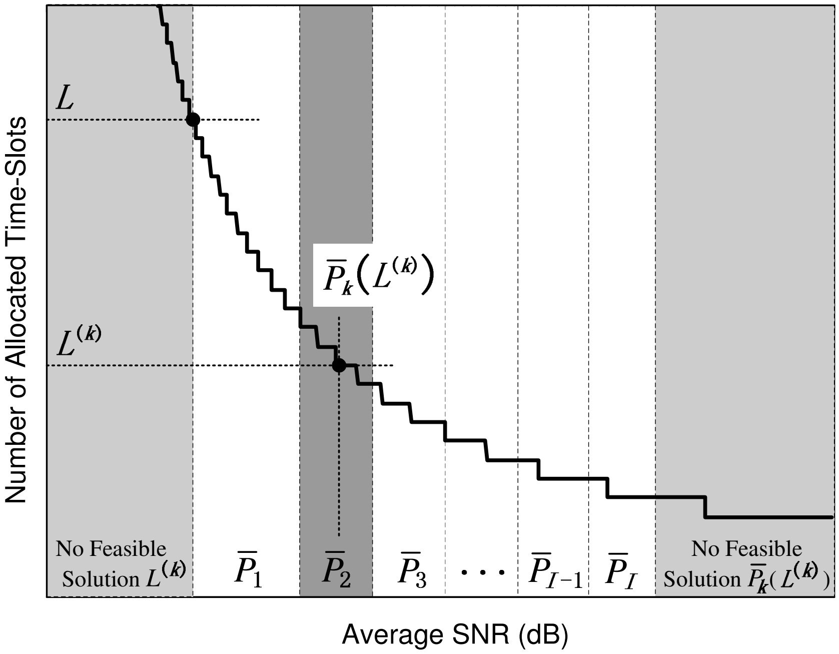


Fig. 39. The time-slot and power-level mapping relations.

the new mobile user arrives or the old mobile user leaves. In the case when the new user tries to join the network, it is possible that there is no feasible solution for the above problem. Thus, the basestation cannot support the QoS requirement for the admission-testing mobile user and therefore this mobile user is rejected to join the wireless networks. Otherwise, the new mobile user is assigned with certain time-slots and power-level for transmissions.

2. Complexity Discussions

In general, the complexity of finding the optimal power-level and time-slots for multiple mobile users *exponentially* increases with the dimension of the searching space. For example, for K users each being assigned with L time-slots and I power-levels, the complexity is approximately proportional to $(LI)^K$. In contrast, by using our proposed dynamic-programming based allocation scheme, the complexity is *linearly* increased with LK. The key reasons of this complexity decreasing include the followings. First, the employment of dynamic programming reduces the exponential complexity to linear complexity. Second, by using the power-level mapping procedure introduced in Section 1, the burden of finding the minimum power-level (with complexity proportional to I) is transferred to look up the "time-slot allocation table" as shown by Fig. 39. Therefore, the complexity of dynamic-programming is independent of I. In practical systems, this time-slot allocation table can be calculated off-line and stored at the basestation in advance, without costing run-time CPU resources.

3. Simulation Results

We also conduct simulations to evaluate our proposed joint power-level and time-slot allocation algorithms. In the simulations, the traffic types are randomly selected between audio and video services with probability of 50% for each type. We set the discrete average-power varying within a dynamic range of ±3 dB, with 7 discrete levels $\{-3\,\text{dB}, -2\,\text{dB}, ..., 2\,\text{dB}, 3\,\text{dB}\}$ relative to the central power-level (0 dB). Also, for a fair comparison with the results in previous sections, we let the SNR of each user be uniformly distributed between 5 dB and 25 dB when using the central power-level (0 dB). Note that in UMTS 3GPP standard [96], the power-level dynamic range is ±9 dB (normal condition) and ±12 dB (extreme condition), which is much larger than that used in our simulation. Therefore, our simulation results are still conservative in terms of performance improvements.

Fig. 40 plots the average energy consumption comparisons between the above three schemes, where the power is normalized by the central power-level (0 dB). We can also observe from Fig. 40 that the joint power-level and time-slot allocation has significant advantage over the water-filling and constant-power approaches. Fig. 41

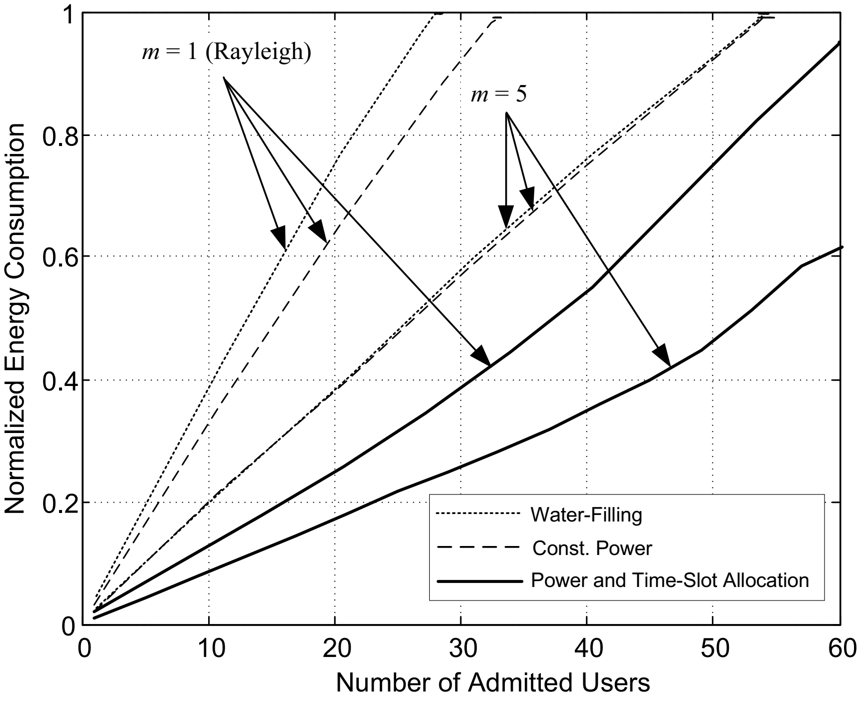


Fig. 40. The average energy consumption comparisons.

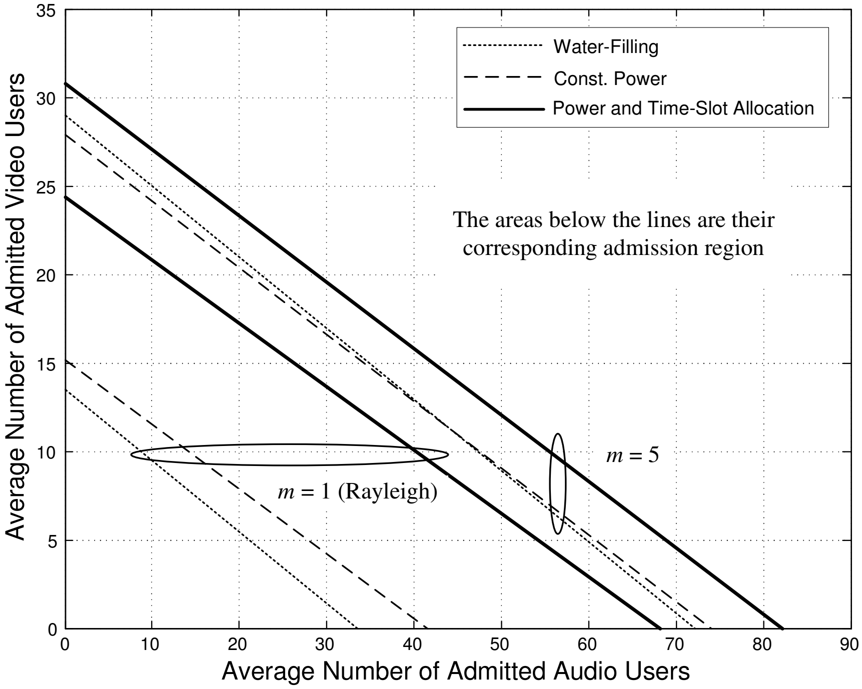


Fig. 41. The average admission region of the system.

depicts the simulation results of the average admission-regions for the video and audio users. As shown by Fig. 41, the averaged admission region can be enlarged by the dynamic-programming-based resource allocation. When the fading parameter $m = 5$, the improvement is not as significant as that in Rayleigh fading channel, which is due to the system capacity limit ($L = 100$). However, our simulations show that this admission region is achieved by using only 65% of the power as compared to that in Rayleigh channel.

E. The Impact of Feedback Delay

In previous sections, we assume that the CSI is reliably fed back to the transmitter without error and delay. However, in practice, this assumption hardly holds. In particular, the CSI feedback delay is un-avoidable in most situations. Without loss of generality, we discuss the impact of feedback delay on a single user and omit the user-index for simplicity.

In order to guarantee the reliability QoS, the system needs to maintain the same BER as that for the case without feedback delay. As a result, the boundary points $\{\Gamma_n\}_{n=1}^{N-1}$ for the adaptive modulation should be re-calculated. In [64], the authors analyzed the impact of CSI feedback delay on BER performance for the adaptive modulation. In [37] (also see in Chapter VIII for details), we also investigated the feedback delay issue for transmit-selection-combining (SC)/receive-maximal-ratio combining (MRC)-based multiple-input-multiple-output (MIMO) scheme from BER perspective. Using the similar approach to [37, 64], we study the impact of feedback delay on the system's delay-bound QoS performance for different power-control policies as follows.

1. QoS-Driven and Water-Filling Power Controls

We first investigate our proposed QoS-driven power control. When considering the feedback delay, the transmission procedure can be described as follows. The constellation M_n is determined based on the SNR γ at time t, but the constellation is transmitted at time $t+\tau$ with actual SNR denoted by $\widehat{\gamma}$. In order to achieve the actual BER of $\mathrm{P_{tgt}}$ as in the case without delay, the system needs to be designed to operate at a lower target BER, which is denoted by $\mathrm{P}'_{\mathrm{tgt}}$. According to Eq. (6.2), the instantaneous BER at time $t+\tau$, denoted by $\mathrm{BER}_n(\widehat{\gamma}|\gamma)$, is given by

$$\mathrm{BER}_n(\widehat{\gamma}|\gamma) = 0.2\exp\Big(-g_n\mu_n(\gamma)\widehat{\gamma}\Big) = 0.2\exp\left(\frac{\log(5\mathrm{P}'_{\mathrm{tgt}})\widehat{\gamma}}{\gamma}\right) \tag{6.26}$$

where $\mu_n(\gamma)$ is the QoS-driven power-control law given by Eq. (6.4), except that $\mathrm{P_{tgt}}$ in the parameter ν_n should be replaced by the new target BER $\mathrm{P}'_{\mathrm{tgt}}$. Then, we obtain the average BER with a given γ, denoted by $\mathrm{BER}_n(\gamma)$, as follows:

$$\mathrm{BER}_n(\gamma) \;\; = \;\; \int_0^\infty \mathrm{BER}_n(\widehat{\gamma}|\gamma)\; p_{\widehat{\Gamma}|\Gamma}\left(\widehat{\gamma}|\gamma\right) d\widehat{\gamma} \tag{6.27}$$

where $p_{\widehat{\Gamma}|\Gamma}\left(\widehat{\gamma}|\gamma\right)$ is the pdf of $\widehat{\gamma}$ conditioned on γ, which is given by [37]

$$p_{\widehat{\Gamma}|\Gamma}\left(\widehat{\gamma}|\gamma\right) = \frac{1}{(1-\rho)}\left(\frac{m}{\overline{\gamma}}\right)\left(\frac{\widehat{\gamma}}{\rho\gamma}\right)^{\frac{m-1}{2}}\exp\left(-\frac{m(\rho\gamma+\widehat{\gamma})}{(1-\rho)\overline{\gamma}}\right)I_{m-1}\left(\frac{2m\sqrt{\rho\gamma\widehat{\gamma}}}{(1-\rho)\overline{\gamma}}\right) \tag{6.28}$$

where $I_\nu(\cdot)$ denotes the modified Bessel function of the first kind with order ν and ρ represents the correlation coefficient between $\widehat{\gamma}$ and γ, which is given by $\rho = J_0^2(2\pi f_d\tau)$ [34] with $J_0(\cdot)$ denoting the zero-th-order Bessel function of the first kind. Omitting the tedious derivations for lack of space, we obtain $\mathrm{BER}_n(\gamma)$ in Eq. (6.27) as a closed-form as follows:

$$\mathrm{BER}_n(\gamma) = 0.2\left(\frac{m\gamma}{m\gamma-(1-\rho)\overline{\gamma}\log(5\mathrm{P}'_{\mathrm{tgt}})}\right)^m \exp\left(\frac{m\rho\log(5\mathrm{P}'_{\mathrm{tgt}})\gamma}{m\gamma-(1-\rho)\overline{\gamma}\log(5\mathrm{P}'_{\mathrm{tgt}})}\right) \tag{6.29}$$

Averaging Eq. (6.29) with respect to the pdf $p_\Gamma(\gamma)$ of γ given by Eq. (6.1), we can express the average BER, denoted by $\overline{\mathrm{BER}}_n$, when γ falls into the nth mode, as follows:

$$\overline{\mathrm{BER}}_n = \frac{1}{\pi_n}\int_{\Gamma_n}^{\Gamma_{n+1}} \mathrm{BER}_n(\gamma)p_\Gamma(\gamma)d\gamma \tag{6.30}$$

where π_n and Γ_n are given by Eqs. (6.3) and (6.6), respectively. It is hard to find the closed-form expression for Eq. (6.30). However, it can be solved by a single finite-integral as

$$\begin{aligned}\overline{\mathrm{BER}}_n &= \frac{0.2(1-\rho)^m[\log(5\mathrm{P}'_{\mathrm{tgt}})]^m}{\pi_n\Gamma(m)}\\ &\quad\cdot\int_{x_n}^{x_{n+1}} \exp\left(\frac{\log(5\mathrm{P}'_{\mathrm{tgt}})x(1-\rho x)}{1-x}\right)\frac{x^{2m-1}}{(1-x)^{m+1}}dx\end{aligned} \tag{6.31}$$

where $x_n = m\Gamma_n/\left[m\Gamma_n - (1-\rho)\overline{\gamma}\log(5\mathrm{P}'_{\mathrm{tgt}})\right]$ and $x_N = 1$. The numerical searching procedure is used to search for the new target BER $\mathrm{P}'_{\mathrm{tgt}}$ such that the actual BER after delay satisfies

$$\mathrm{P}_{\mathrm{tgt}} = \frac{\sum_{n=1}^{N-1} n\pi_n\overline{\mathrm{BER}}_n}{\sum_{n=1}^{N-1} n\pi_n}. \tag{6.32}$$

Once the new target BER $\mathrm{P}'_{\mathrm{tgt}}$ is obtained, we can find the new boundary points $\{\Gamma_n\}_{n=1}^{N-1}$ and thus reconstruct the FSMC of the service-process. Then, the resource-allocation algorithms can be re-executed based on the new FSMC. For water-filling power control, the procedure is the similar, but omitted for lack of space.

2. Constant Power-Control

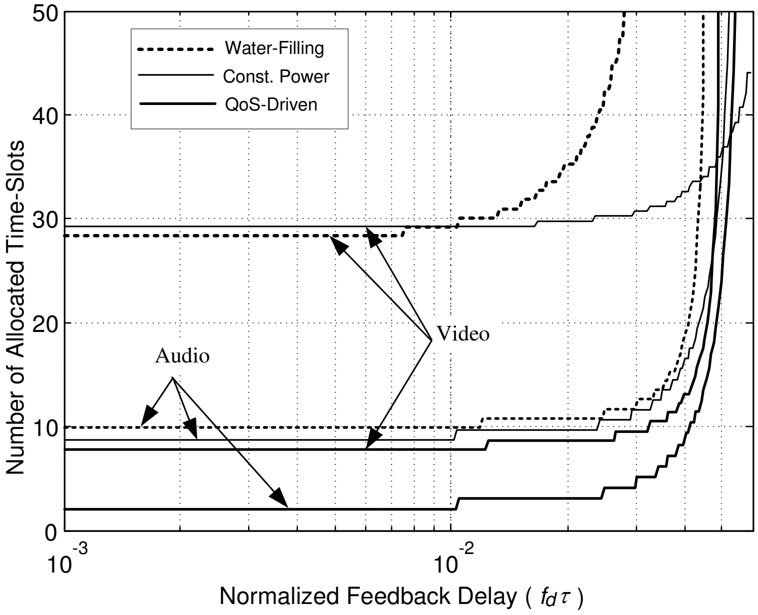
Based on the similar approach to Section 1, we can show that the average BER for the nth mode can be derived as

$$\begin{aligned}\overline{\mathrm{BER}}_n &= \frac{1}{\pi_n}\int_{\Gamma_n}^{\Gamma_{n+1}} \mathrm{BER}_n(\gamma)p_\Gamma(\gamma)d\gamma \\ &= \frac{0.2}{\pi_n\Gamma(m)}\left(\frac{m}{b'_n}\right)^m\left[\Gamma\left(m, \frac{b'_n\Gamma_n}{\overline{\gamma}}\right) - \Gamma\left(m, \frac{b'_n\Gamma_{n+1}}{\overline{\gamma}}\right)\right] \qquad (6.33)\end{aligned}$$

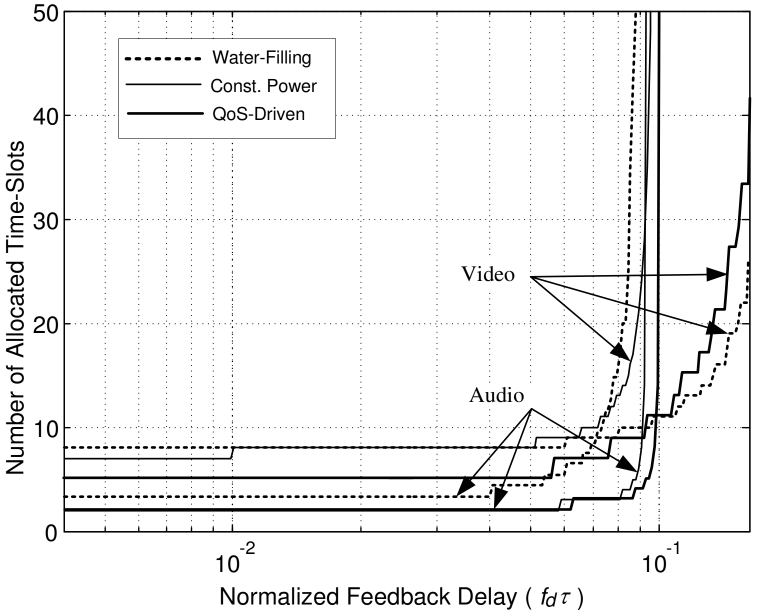
where Γ_n is given by Eq. (6.9), $b'_n = m(\overline{\gamma}g_n + m)/\zeta_n$, and $\zeta_n = m + (1-\rho)\overline{\gamma}g_n$. The searching procedure is also to find the new boundary points $\{\Gamma_n\}_{n=1}^{N-1}$ such that Eq. (6.32) is satisfied. Also, after the boundary points $\{\Gamma_n\}_{n=1}^{N-1}$ are determined, we can reconstruct the FSMC of the service-process and then we can re-execute the resource-allocation algorithms based on the new FSMC.

3. Numerical and Simulation Results

The above analyses are verified by the numerical and simulation results. In Fig. 42, we investigate the impact of feedback delay on time-slot allocations. We can see from Fig. 42 that the time-slot allocation results remain unchanged when the normalized feedback delay is below certain threshold. When $f_d\tau$ further increases, the number $L^{(k)}$ starts increasing in order to maintain the same statistical QoS requirements. From Fig. 42, we know that for all power-control policies, the better quality channel ($m = 5$) can tolerant larger feedback delay than the Rayleigh fading channel ($m = 1$). Specifically, the Rayleigh channel can only tolerant feedback delay with $f_d\tau \leq 0.01$, while the channel with $m = 5$ can tolerant the delay $f_d\tau \geq 0.04$. Note that in our system, we have $T_f \times f_d = 0.03$, implying that the channel with $m = 5$ can tolerant the feedback delay which is even larger than one frame's time duration. Thus, the proposed scheme provides sufficient robustness to the system in an in-door mobile



(a) $m = 1$.



(b) $m = 5$.

Fig. 42. The impact of CSI feedback delay on the time-slot allocation. The average SNR is set to $\overline{\gamma} = 10$ dB.

environment, e.g., the widely used WLAN.

F. Summary

We proposed and analyzed a cross-layer-model based adaptive resource-allocation scheme for diverse QoS guarantees over downlink mobile wireless networks. Our scheme jointly allocates power-levels and time-slots for real-time users to guarantee the diverse statistical delay-bound QoS requirements. We developed the admission-control and power/time-slot allocation algorithms by extending the effective capacity method. We also studied the impact of adaptive power control and CSI feedback delay at physical-layer on the QoS provisioning performance. Compared to the conventional water-filling and constant power approach, our proposed QoS-driven power adaptation shows significant advantages. The joint power/time-slot allocation scheme can significantly reduce the transmit power, or equivalently, increase the admission region. Also, in an in-door mobile environment, our proposed algorithm is shown to be robust to the CSI feedback delay.

The cellular wireless networks have been widely employed in real communication systems. In the next chapter, we will study a promising newly proposed network infrastructure, namely, the cooperative relay networks, and investigate its resource allocation strategy.

CHAPTER VII

RESOURCE ALLOCATION FOR COOPERATIVE RELAY NETWORKS

A. Introduction

With the explosive developments of wireless communications, quality-of-service (QoS) provisioning has become a critically important performance metric for the future wireless networks. Unlike wireline networks, in which QoS can be guaranteed by independent optimization within each layer in the open system interconnection (OSI) model, over wireless networks there is a strong interconnection between layers, which makes the layered design and optimization approach less efficient. For example, at the physical layer, a great deal of research focuses on techniques that can enhance the spectral efficiency of wireless systems. The framework used to evaluate these techniques is mainly based on information theory, using the concept of Shannon capacity [6, 7]. However, it is well known that Shannon capacity does not place any restrictions on complexity and delay [64]. As a result, the optimization merely at the physical layer may not lead to the desired delay QoS requested by the services at upper-protocol layers.

To deal with this problem, there have been increasing interests in design for wireless networks that rely on interactions between various layers of the protocol stack. This approach, called *cross-layer design and optimization*, has been widely recognized as a promising solution to provide diverse QoS provisioning in wireless multimedia communications [97]. The cross-layer approach relaxes the layering architecture of the conventional network model, which can result in a significant performance enhancement. However, such a design principle across different layers usually involves high complexity, which may cause the optimization problem intractable [98]. Consequently, how to develop efficient cross-layer approaches while minimizing the additional requested information exchanged between layers is an important issue from both theoretical and practical point-of-views.

On the other hand, relay communications have recently emerged as a powerful spatial diversity technique that can improve the performance over conventional point-to-point transmissions. The original work on relay communications was initialed by Cover and Gamal [99]. Since then, it has been extensively studied using different performance metrics [26, 100–107], especially when the concept of *user cooperation* was proposed [100, 101]. Clearly, combining the idea of cross-layer design with the relay network architecture, it is possible to significantly improve the system QoS provisioning performance. However, the research on how to efficiently employ the unique nature of relay architecture for designing the cross-layer protocols, and what is the impact of cross-layer resource allocation on supporting diverse QoS requirements over wireless relay networks, are still quite scarce [108].

To remedy the above deficiency, in this chapter we propose a cross-layer resource allocation scheme for relay networks with the target at delay QoS guarantees for wireless multimedia communications. Our proposed scheme aims at maximizing the relay network throughput subject to a given delay QoS constraint. Our work builds on the integration of information theoretic results with the theory of statistical QoS guarantees, in particular, the recently developed powerful concept termed *effective capacity* [20–23]. The theory of statistical QoS guarantees has been extensively studied in the early 90's with the emphasis on wired asynchronous transfer mode (ATM) networks [1,12–19]. This theory enables us to analyze network statistics such as queue distributions, buffer overflow probabilities, and delay-bound violation probabilities, which are all important delay QoS metrics. As a part of the statistical QoS theory, effective capacity is particularly convenient for analyzing the statistical QoS performance of wireless multimedia transmissions where the service process is driven by the time-varying wireless channel.

Specifically, our resource allocation scheme is across the physical and the datalink layers. Applying the effective-capacity based approach, we convert the original throughput maximization to effective capacity maximization, and characterize the delay constraint by the so-called *QoS exponent* θ, which is the only requested information exchanged between the physical layer and the datalink layer in our cross-layer scheme. In particular, the dynamics of θ corresponds to different delay QoS constraints. For instance, non-real-time services such as data disseminations aim at maximizing the throughput with a loose delay constraint ($\theta \to 0$). In contrast, the key QoS requirement for real-time multimedia services is the timely delivery with stringently upper-bounded delay ($\theta \to \infty$). There also exist some services falling in between, like paging and interactive web surfing, which are delay sensitive but the delay QoS requirements are not as stringent as those of real-time applications ($0 < \theta < \infty$).

We focus on simple half-duplex relay protocols proposed in [104], namely, amplify-and-forward (AF) and decode-and-forward (DF), and develop the associated dynamic resource-allocation algorithms, where the resource allocation policies are functions of both the network channel state information (CSI) and the QoS constraint θ. The resulting resource allocation policy in turn provides a guideline on how to design the relay protocol that can efficiently support stringent QoS constraints. For DF relay networks, we also study a fixed power allocation scheme and investigate its performance. The simulations and numerical results verify that our proposed cross-layer resource allocation can efficiently support diverse QoS requirements over wireless relay networks. Moreover, both AF and DF relays show significant superiorities over direct transmissions when the delay constraints are stringent. On the other hand, our results demonstrate the importance of deploying the dynamic resource allocation for stringent delay QoS guarantees.

The rest of the chapter is organized as follows. Section B describes our cross-layer relay network model. Sections C and D develop the cross-layer resource allocation policies for AF and DF relay networks, respectively. Section E investigates a fixed power allocation policy for DF relay networks. Section F presents simulations and numerical results to evaluate the performance of our proposed cross-layer resource allocation. The chapter concludes with Section G.

B. System Descriptions

1. Network Model

The cross-layer relay network model is shown in Fig. 43. We concentrate on a discrete-time system with a source node (S), a destination node (D), and a relay node (R), where the relay assists communications between the source and the destination without having its own data to send. As illustrated by Fig. 43, a first-in-first-out (FIFO) queue is implemented at the source node, which comprises the datalink-layer *packets* to be transmitted to the destination. At the physical (PHY)-layer, the datalink-layer packets are divided into *frames*, which form the data units through wireless transmissions. The frame duration is denoted by T_f, which is assumed to be less than the fading coherence time, but sufficiently long so that the information-theoretic assumption of infinite code-block length is meaningful. Based on a given QoS constraint θ requested by the service and CSI fed back from the corresponding receivers, the source and the relay need to find an optimal resource allocation strategy that can maximize the throughput subject to the QoS constraint θ. At the relay node, the transmission only involves the physical layer, as shown by Fig. 43. In this chapter, we also make the following assumptions.

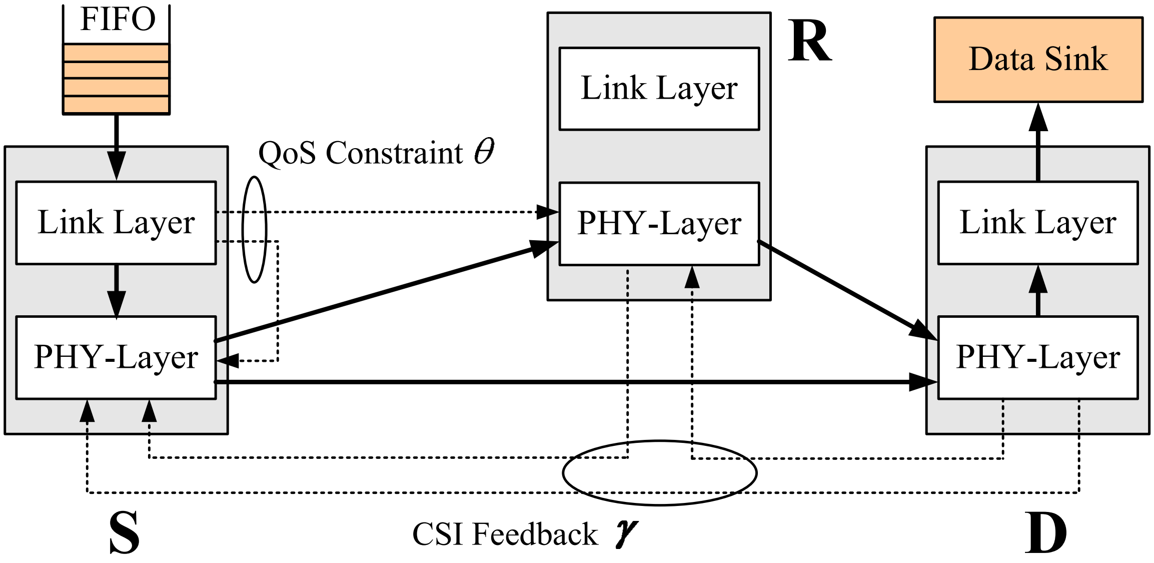


Fig. 43. The cross-layer relay network model.

A1: The discrete-time channel is assumed to be block fading. The path gains are invariant within a frame's duration T_f, but vary independently from one frame to another. The block fading channel model is commonly used in literatures, which can also greatly simplify our analyses. Moreover, through the study in Chapter III we observe that there exists a simple and efficient approach to convert the resource allocation policy obtained in block fading channels to that over correlated fading channels, making the investigation of block fading channel more applicable.

A2: We assume that CSI is perfectly estimated at the corresponding receivers and reliably fed back to the source and the relay without delay. The assumption that the feedback is reliable can be (at least approximately) satisfied by using heavily coding feedback channels. On the other hand, the feedback delay can be compensated by channel prediction [92].

A3: We further assume that for a given instantaneous channel gain, the physical-layer codewords adaptively operates at the instantaneous achievable rate of the relay

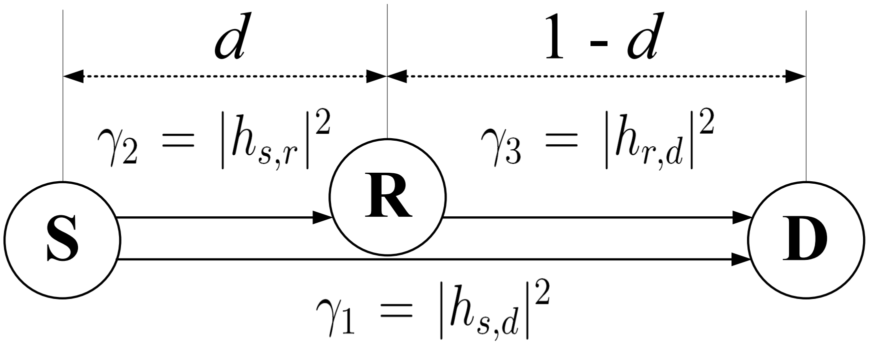


Fig. 44. The relay channel model.

protocol. This assumption implies that an ideal adaptive modulation and coding scheme is implemented.

2. Channel Model

The relay channel model is shown in Fig. 44. We assume a flat fading channel model. The instantaneous channel coefficient between sender i and receiver j is denoted by $\{h_{i,j}\}$, where $i \in \{s, r\}$ and $j \in \{r, d\}$ with $i \neq j$, and s, r, d represent the source, relay, and destination, respectively. The magnitudes of these channel coefficients are assumed to follow an independent Rayleigh distribution, with the mean determined by the large-scale path loss. At each receiver, the additive noise is modeled as independent zero-mean, circularly symmetric complex white Gaussian with *unit* variance.

In the following discussions, we denote the channel gain $\gamma_1 = |h_{s,d}|^2$, $\gamma_2 = |h_{s,r}|^2$, and $\gamma_3 = |h_{r,d}|^2$, where γ_i follows an exponential distribution with parameter λ_i, $i \in \{1, 2, 3\}$. To study the impact of relay location on network performance, we normalize the distance between the source and the destination to one, and let the relay be located in a line between the source and the destination. The source-relay distance and the relay-destination distance are denoted by d and $1 - d$, respectively, where $0 < d < 1$. Based on the channel model and network topology described above, the network CSI is determined by a 3-tuple $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \gamma_3)$, which follows independent exponential distribution with parameter $\lambda_1 = 1$, $\lambda_2 = d^\alpha$, and $\lambda_3 = (1-d)^\alpha$, respectively, where α denotes the path loss exponent. A typical value of α lies in the range of $(2, 5)$. In the simulations and numerical results presented in Section F, we will assume $\alpha = 4$.

3. Relay Protocols

Different relay protocols have been investigated in literatures [26, 102–107]. In this chapter, we mainly focus on the simple half-duplex relay protocols proposed in [104], namely, amplify-and-forward (AF) and decode-and-forward (DF). As compared to full-duplex relay, half-duplex relay restricts that terminals cannot transmit and receive simultaneously at the same frequency band, which enjoys much lower implementation complexity than the full-duplex relay. Moreover, the orthogonal transmission strategies of AF and DF protocols in [104] eliminates the potential interference between the source and the relay.

Specifically, for both AF and DF relays, each frame duration T_f is divided into two equal portions. During the first half period of frame duration, the source transmits to both the relay and the destination. In the second half period, the relay forwards the message to the destination, where the forward strategy depends on specific relay protocol used.

a. AF Protocol

In AF mode, the relay simply amplifies and then forwards what it receives during the first half to the destination. This strategy is also called non-regenerative relay or analog relay protocol [106, 107]. Let P_s and P_r denote the average transmit power assigned to the source and the relay, respectively. Then, the achievable rate of AF protocol, denoted by R_{AF}, can be expressed as [104]

$$R_{AF} = \left(\frac{T_f B}{2}\right) \log_2 \left(1 + 2\gamma_1 P_s + \frac{4\gamma_2 P_s \gamma_3 P_r}{1 + 2\gamma_2 P_s + 2\gamma_3 P_r}\right) \tag{7.1}$$

where B denotes the system spectral bandwidth. Note that in Eq. (7.1), since each transmitter sends data only for half of the frame duration, the source uses power $2P_s$ during the first half and the relay uses power $2P_r$ during the second half, which results in a total average power of $P_s + P_r$ per frame.

b. DF Protocol

In DF mode, the relay forwards the message to the destination if it decodes successfully. Correspondingly, this strategy is also called regenerative relay or digital relay [106, 107]. The achievable rate of DF protocol, denoted by R_{DF}, can be expressed as [104]

$$R_{DF} = \left(\frac{T_f B}{2}\right) \min\left\{\log_2\left(1 + 2\gamma_2 P_s\right), \log_2\left(1 + 2\gamma_1 P_s + 2\gamma_3 P_r\right)\right\}. \tag{7.2}$$

Throughout this chapter, we further assume that the relay network has a mean total transmit power constraint, denoted by $\overline{P}$. Thus, the transmit power P_s and P_r need to satisfy

$$\mathbb{E}[P_s + P_r] \leq \overline{P} \tag{7.3}$$

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

C. Dynamic Resource Allocation for AF Relay Networks

1. Problem Formulation

Conventionally, the resource allocation policy can be expressed as a function of the instantaneous network CSI $\boldsymbol{\gamma}$. In contrast, our resource allocation policy is a function of not only the instantaneous CSI $\boldsymbol{\gamma}$, but also the QoS exponent θ. Correspondingly, let us define $\boldsymbol{\nu} \triangleq (\theta, \boldsymbol{\gamma})$ as *network state information* (NSI). Then, we can rewrite Eq. (7.1) as

$$R_{AF}(\boldsymbol{\nu}) = \left(\frac{T_f B}{2}\right) \log_2 \left(1 + 2\gamma_1 P_s(\boldsymbol{\nu}) + \frac{4\gamma_2 P_s(\boldsymbol{\nu})\gamma_3 P_r(\boldsymbol{\nu})}{1 + 2\gamma_2 P_s(\boldsymbol{\nu}) + 2\gamma_3 P_r(\boldsymbol{\nu})}\right) \tag{7.4}$$

where both the achievable rate $R_{AF}(\boldsymbol{\nu})$ and the average transmit power $P_s(\boldsymbol{\nu})$ and $P_r(\boldsymbol{\nu})$ are functions of the NSI $\boldsymbol{\nu}$. For a given QoS constraint specified by θ, in order to find the optimal resource allocation policy that maximizes the effective capacity of Eq. (2.15), we can formulate our maximization problem as follows:

$$(P1) \quad \arg\max_{\mathbf{P}(\boldsymbol{\nu})} \left\{ -\frac{1}{\theta} \log \left(\mathbb{E}_{\boldsymbol{\gamma}} \left[\exp\Big(-\theta R_{AF}(\boldsymbol{\nu}) \Big) \right] \right) \right\} \tag{7.5}$$

subject to the following power constraints:

$$\begin{cases} \mathbb{E}_{\boldsymbol{\gamma}} \left[P_s(\boldsymbol{\nu}) + P_r(\boldsymbol{\nu}) \right] \leq \overline{P} \\ P_s(\boldsymbol{\nu}) \geq 0 \text{ and } P_r(\boldsymbol{\nu}) \geq 0 \end{cases} \tag{7.6}$$

where we define $\mathbf{P}(\boldsymbol{\nu}) \triangleq (P_s(\boldsymbol{\nu}), P_r(\boldsymbol{\nu}))$ as network power allocation policy, and $\mathbb{E}_{\boldsymbol{\gamma}}[\cdot]$ emphasizes that the expectation is with respect to $\boldsymbol{\gamma}$. Since $\log(\cdot)$ is a monotonically increasing function, for each given QoS constraint $\theta > 0$, the maximization problem $(P1)$ is equivalent to the following minimization problem:

$$(P1') \quad \arg\min_{\mathbf{P}(\boldsymbol{\nu})} \left\{ \mathbb{E}_{\boldsymbol{\gamma}} \left[\exp\Big(-\theta R_{AF}(\boldsymbol{\nu}) \Big) \right] \right\} \tag{7.7}$$

subject to the same constraints given by Eq. (7.6). Problem $(P1')$ is still not easy to solve since the objective is not convex. To simplify the problem, we approximate the rate $R_{AF}(\boldsymbol{\nu})$ by $\widetilde{R}_{AF}(\boldsymbol{\nu})$ as

$$\widetilde{R}_{AF}(\boldsymbol{\nu}) = \left(\frac{T_f B}{2}\right) \log_2 \left(1 + 2\gamma_1 P_s(\boldsymbol{\nu}) + \frac{2\gamma_2 P_s(\boldsymbol{\nu})\gamma_3 P_r(\boldsymbol{\nu})}{\gamma_2 P_s(\boldsymbol{\nu}) + \gamma_3 P_r(\boldsymbol{\nu})}\right). \tag{7.8}$$

The approximation $\widetilde{R}_{AF}(\boldsymbol{\nu})$ in Eq. (7.8) takes the advantages of mathematical tractability over $R_{AF}(\boldsymbol{\nu})$ in Eq. (7.4). In particular, $\widetilde{R}_{AF}(\boldsymbol{\nu})$ is strictly concave on the space spanned by $(P_s(\boldsymbol{\nu}), P_r(\boldsymbol{\nu}))$, which makes the related optimization much easier than the original problem $(P1')$. Furthermore, $\widetilde{R}_{AF}(\boldsymbol{\nu})$ serves as a tight upper-bound for $R_{AF}(\boldsymbol{\nu})$, especially at the high signal-to-noise (SNR) regime [106]. As a result, the effective capacity derived by using $\widetilde{R}_{AF}(\boldsymbol{\nu})$ is also a tight upper-bound for the effective capacity derived by using $R_{AF}(\boldsymbol{\nu})$.

Replacing $R_{AF}(\boldsymbol{\nu})$ by $\widetilde{R}_{AF}(\boldsymbol{\nu})$ in Eq. (7.7), we get the following optimization problem ready to be solved:

$$\begin{aligned}(P1'') \qquad & \arg\min_{\mathbf{P}(\boldsymbol{\nu})} \left\{ \mathbb{E}_{\boldsymbol{\gamma}} \left[\exp\left(-\theta \widetilde{R}_{AF}(\boldsymbol{\nu})\right)\right]\right\} \\ & = \arg\min_{\mathbf{P}(\boldsymbol{\nu})} \left\{ \mathbb{E}_{\boldsymbol{\gamma}} \left[\left(1 + 2\gamma_1 P_s(\boldsymbol{\nu}) + \frac{2\gamma_2 P_s(\boldsymbol{\nu})\gamma_3 P_r(\boldsymbol{\nu})}{\gamma_2 P_s(\boldsymbol{\nu}) + \gamma_3 P_r(\boldsymbol{\nu})}\right)^{-\frac{\beta}{2}}\right]\right\} \end{aligned} \tag{7.9}$$

subject to the constraints given by Eq. (7.6), where we define

$$\beta \triangleq \frac{\theta T_f B}{\log 2}$$

as the *normalized* QoS exponent.

2. Resource Allocation Policy

The following theorem solves the optimization problem $(P1'')$ derived in the above.

Theorem 6. *The optimal resource allocation policy* $\mathbf{P}(\boldsymbol{\nu})$ *that solves* $(P1'')$ *is deter-*

mined by

$$\begin{cases} P_s(\boldsymbol{\nu}) = uP_r(\boldsymbol{\nu}) \\ P_r(\boldsymbol{\nu}) = \dfrac{1}{v}\left(\left[\left(\dfrac{\gamma_0}{\gamma_3}\right)\left(\dfrac{\gamma_3+c}{\gamma_1+c}\right)^2\right]^{-\frac{2}{\beta+2}} - 1\right), \end{cases} \tag{7.10}$$

if both $P_s(\boldsymbol{\nu}) > 0$ and $P_r(\boldsymbol{\nu}) > 0$ in Eq. (7.10), where the parameters u and v are defined by

$$\begin{cases} u = \frac{\gamma_3(\gamma_1+c)}{(\gamma_3-\gamma_1)\gamma_2} \\ v = \frac{2c\gamma_3(\gamma_1+c)^2}{(\gamma_3-\gamma_1)\gamma_2(\gamma_3+c)}, \end{cases} \tag{7.11}$$

with $c = \sqrt{\gamma_1\gamma_3 + \gamma_2\gamma_3 - \gamma_1\gamma_2}$, and γ_0 is a cutoff threshold determined by the mean total network power constraint.

Otherwise, the policy reduces to direct transmission and $\mathbf{P}(\boldsymbol{\nu})$ is determined by

$$\begin{cases} P_s(\boldsymbol{\nu}) = \dfrac{1}{2}\left[\left(\gamma_0^{\frac{2}{\beta+2}}\gamma_1^{\frac{\beta}{\beta+2}}\right)^{-1} - \gamma_1{}^{-1}\right]^+ \\ P_r(\boldsymbol{\nu}) = 0 \end{cases} \tag{7.12}$$

where $[x]^+ \triangleq \max\{x, 0\}$.

Proof. The proof is provided in Appendix M. □

As mentioned in Section 1, the above solution for the problem $(P1'')$ serves as a tight upper-bound for the optimal effective capacity. On the other hand, by applying the above solution directly to the original problem $(P1)$, we can obtain a lower-bound for the optimal effective capacity (since it is an achievable effective capacity). We will see by the numerical examples later that the upper-bound and lower-bound are very close, especially at the high SNR regime.

Since a necessary condition for the resource allocation to take the form of Eq. (7.10) is $P_s(\boldsymbol{\nu}) > 0$, which in turn requests $u > 0$, implying $\gamma_3 > \gamma_1$, we therefore have the following corollary.

Corollary 2. *For AF protocol, if $\gamma_1 \geq \gamma_3$ (the direct S-D link is better than the relay R-D link), then the optimal resource allocation reduces to the direct transmission, no matter what the QoS constraint θ is.*

3. Limiting Resource Allocation Policies

The dynamics of the QoS exponent θ characterizes how stringent the QoS requirement is. We showed in Chapter III that as the QoS exponent $\theta \to 0$, the optimal effective capacity approaches the ergodic capacity of the system. On the other hand, as the QoS exponent $\theta \to \infty$, the optimal effective capacity approaches the zero-outage capacity of the system. Making use of these properties, we can obtain the resource allocation policy that characterizes the upper- and lower-bounds for the ergodic and zero-outage capacity of the AF relay protocol.

Proposition 7. *The resource allocation policy that can upper- and lower-bound the ergodic capacity of the AF relay protocol is determined by*

$$\begin{cases} P_s(\boldsymbol{\nu}) = uP_r(\boldsymbol{\nu}) \\ P_r(\boldsymbol{\nu}) = \dfrac{1}{v}\left[\dfrac{\gamma_3}{\gamma_0}\left(\dfrac{\gamma_1 + c}{\gamma_3 + c}\right)^2 - 1\right], \end{cases} \tag{7.13}$$

if both $P_s(\boldsymbol{\nu}) > 0$ and $P_r(\boldsymbol{\nu}) > 0$ in Eq. (7.13); otherwise, it reduces to the direct transmission (water-filling) as

$$\begin{cases} P_s(\boldsymbol{\nu}) = \dfrac{1}{2}\left[\dfrac{1}{\gamma_0} - \dfrac{1}{\gamma_1}\right]^+ \\ P_r(\boldsymbol{\nu}) = 0. \end{cases} \tag{7.14}$$

Proof. Letting $\theta \to 0$ in Eqs. (7.10) and (7.12), we can obtain the desired results. □

Proposition 8. *The resource allocation policy that can upper- and lower-bound the zero-outage capacity of the AF relay protocol is given by*

$$\begin{cases} P_s(\boldsymbol{\nu}) = uP_r(\boldsymbol{\nu})\mathcal{I}\{\gamma_3 > \gamma_1\} + \frac{\sigma}{2\gamma_1}\mathcal{I}\{\gamma_3 \le \gamma_1\} \\ P_r(\boldsymbol{\nu}) = \dfrac{\sigma}{v}\mathcal{I}\{\gamma_3 > \gamma_1\} \end{cases} \tag{7.15}$$

where $\mathcal{I}\{\cdot\}$ denotes the indicator function, and σ is a constant such that the mean total network power constraint is satisfied. Under such a policy, the transmission maintains a constant service rate $(T_f B/2)\log_2\bigl(1+\sigma\bigr)$, no matter what the channel realization $\boldsymbol{\gamma}$ is.

Proof. Letting $\theta \to \infty$ in Eqs. (7.10) and (7.12), we can obtain the desired results, where the constant σ is determined by $\sigma = \lim_{\theta\to\infty} \gamma_0^{-\frac{2}{\beta+2}} - 1$. □

D. Dynamic Resource Allocation for DF Relay Networks

1. Resource Allocation for Original DF Protocol

Similar to the AF case, we first re-write Eq. (7.2) as a function of the NSI $\boldsymbol{\nu}$ as follows:

$$R_{DF}(\boldsymbol{\nu}) = \left(\frac{T_f B}{2}\right)\min\Bigl\{\log_2\Bigl(1+2\gamma_2 P_s(\boldsymbol{\nu})\Bigr), \log_2\Bigl(1+2\gamma_1 P_s(\boldsymbol{\nu}) + 2\gamma_3 P_r(\boldsymbol{\nu})\Bigr)\Bigr\}. \tag{7.16}$$

Then, the optimization problem can be formulated as

$$(P2) \quad \arg\max_{\mathbf{P}(\boldsymbol{\nu})}\left\{-\frac{1}{\theta}\log\left(\mathbb{E}_{\boldsymbol{\gamma}}\left[\exp\Bigl(-\theta R_{DF}(\boldsymbol{\nu})\Bigr)\right]\right)\right\} \tag{7.17}$$

subject to the constraint given by Eq. (7.6). Again, the above maximization problem $(P2)$ is equivalent to the following minimization problem:

$$(P2') \quad \arg\min_{\mathbf{P}(\boldsymbol{\nu})}\Bigl\{\mathbb{E}_{\boldsymbol{\gamma}}\Bigl[\max\bigl\{\mathcal{F}_1(\boldsymbol{\nu}), \mathcal{F}_2(\boldsymbol{\nu})\bigr\}\Bigr]\Bigr\} \tag{7.18}$$

where

$$\mathcal{F}_1(\boldsymbol{\nu}) = \Big(1 + 2\gamma_2 P_s(\boldsymbol{\nu})\Big)^{-\frac{\beta}{2}} \tag{7.19}$$

and

$$\mathcal{F}_2(\boldsymbol{\nu}) = \Big(1 + 2\gamma_1 P_s(\boldsymbol{\nu}) + 2\gamma_3 P_r(\boldsymbol{\nu})\Big)^{-\frac{\beta}{2}}. \tag{7.20}$$

It is easy to show that $(P2')$ is a strictly convex optimization problem and thus has the unique optimal solution. To solve $(P2')$, we consider the following two scenarios.

Scenario-1: $\gamma_2 < \gamma_1$.

If $\gamma_2 < \gamma_1$, then $\mathcal{F}_1(\boldsymbol{\nu}) > \mathcal{F}_2(\boldsymbol{\nu})$ always holds, no matter what the value of $P_r(\boldsymbol{\nu})$ is. To save the transmit power, the optimal resource allocation strategy must satisfy $P_r(\boldsymbol{\nu}) = 0$. As a result, problem $(P2')$ becomes:

$$\arg\min_{\mathbf{P}(\boldsymbol{\nu})} \Big\{ \mathbb{E}_{\boldsymbol{\gamma}}\big[\mathcal{F}_1(\boldsymbol{\nu})\big] \Big\} \tag{7.21}$$

subject to $\mathbb{E}_{\boldsymbol{\gamma}}[P_s(\boldsymbol{\nu})] = \overline{P}$. This is equivalent to a direct transmission problem, where the transmission link is from the source to the relay. The above problem has been solved by Chapter III. The optimal resource allocation policy is determined by:

$$\begin{cases} P_s(\boldsymbol{\nu}) = \dfrac{1}{2}\left[\left(\gamma_0^{\frac{2}{\beta+2}} \gamma_2^{\frac{\beta}{\beta+2}}\right)^{-1} - \gamma_2{}^{-1}\right]^+ \\ P_r(\boldsymbol{\nu}) = 0. \end{cases} \tag{7.22}$$

Scenario-2: $\gamma_2 \geq \gamma_1$.

If $\gamma_2 \geq \gamma_1$, then we can find appropriate $P_s(\boldsymbol{\nu})$ and $P_r(\boldsymbol{\nu})$ such that

$$\mathcal{F}_1(\boldsymbol{\nu}) = \mathcal{F}_2(\boldsymbol{\nu}), \tag{7.23}$$

which in turn gives

$$P_r(\boldsymbol{\nu}) = \tilde{u} P_s(\boldsymbol{\nu}) \tag{7.24}$$

where $\tilde{u} = (\gamma_2 - \gamma_1)/\gamma_3$. In this case, the objective function of the problem $(P2')$ is the same as that given in Eq. (7.21), but subject to the constraints given by Eqs. (7.6) and (7.24). Thus, we can construct the following Lagrangian problem as

$$\begin{aligned} \mathcal{J}_2 &= \mathbb{E}_{\boldsymbol{\gamma}}\left[\mathcal{F}_1(\boldsymbol{\nu})\right] + \lambda\Big(\mathbb{E}_{\boldsymbol{\gamma}}\left[P_s(\boldsymbol{\nu}) + P_r(\boldsymbol{\nu})\right] - \overline{P}\Big) \\ &= \mathbb{E}_{\boldsymbol{\gamma}}\left[\Big(1 + 2\gamma_2 P_s(\boldsymbol{\nu})\Big)^{-\frac{\beta}{2}}\right] + \lambda\Big(\mathbb{E}_{\boldsymbol{\gamma}}\left[(1+\tilde{u})P_s(\boldsymbol{\nu})\right] - \overline{P}\Big). \end{aligned} \tag{7.25}$$

Solving the above Lagrangian problem, we obtain the resource allocation policy under the condition of $\gamma_2 \geq \gamma_1$ as

$$\begin{cases} P_s(\boldsymbol{\nu}) = \dfrac{1}{2}\left[\left(\big[(1+\tilde{u})\gamma_0\big]^{\frac{2}{\beta+2}} \gamma_2^{\frac{\beta}{\beta+2}}\right)^{-1} - \gamma_2^{-1}\right]^+ \\ P_r(\boldsymbol{\nu}) = \tilde{u} P_s(\boldsymbol{\nu}). \end{cases} \tag{7.26}$$

In summary, the optimal resource allocation policy for the original DF protocol is given by either Eq. (7.22) or Eq. (7.26), depending on whether $\gamma_2 < \gamma_1$ or not.

To study the zero-outage capacity of DF relay networks, we let $\theta \to \infty$, and the corresponding resource allocation policy can be expressed as

$$\begin{cases} P_s(\boldsymbol{\nu}) = \dfrac{\sigma}{2\gamma_2} \\ P_r(\boldsymbol{\nu}) = \tilde{u} P_s(\boldsymbol{\nu}) \mathcal{I}\{\gamma_2 \geq \gamma_1\}. \end{cases} \tag{7.27}$$

Similar to the AF case, under such a policy, the transmission maintains a constant service rate $(T_f B/2)\log_2\big(1+\sigma\big)$, no matter what the channel realization $\boldsymbol{\gamma}$ is.

It is important to notice that when $\theta \to 0$, the corresponding resource allocation policy does not lead to the ergodic capacity of DF relay networks, since the ergodic capacity of DF relay networks is determined by

$$\mathcal{C}=\left(\frac{T_fB}{2}\right)\max_{\mathbf{P}(\boldsymbol{\nu})}\min\left\{\mathbb{E}_{\boldsymbol{\gamma}}\left[\log_2\Big(1+2\gamma_2P_s\Big)\right],\mathbb{E}_{\boldsymbol{\gamma}}\left[\log_2\Big(1+2\gamma_1P_s+2\gamma_3P_r\Big)\right]\right\}, \tag{7.28}$$

but the optimal effective capacity of our scheme at $\theta\to 0$ is given by

$$\mathcal{C}'=\left(\frac{T_fB}{2}\right)\max_{\mathbf{P}(\boldsymbol{\nu})}\mathbb{E}_{\boldsymbol{\gamma}}\left[\min\left\{\log_2\Big(1+2\gamma_2P_s\Big),\log_2\Big(1+2\gamma_1P_s+2\gamma_3P_r\Big)\right\}\right]. \tag{7.29}$$

By Jensen's inequality, $\mathcal{C}\geq\mathcal{C}'$ always holds. From an implementation perspective, Eqs. (7.28) and (7.29) correspond to two different transmission strategies for the relay node [105]. On one hand, it can immediately transmit a received and decoded frame to the destination whenever it receives the frame from the source, which corresponds to Eq. (7.29). On the other hand, it can also queue data and then transmit the queued contents when the channel is favorable, which corresponds to Eq. (7.28). Since in our system model, the relay strategy falls into the first category, our obtained effective capacity is no greater than the ergodic capacity. However, this strategy is more practical because it results in shorter delay. The readers are referred to [105] for detailed discussions about resource allocation to achieve the ergodic capacity of relay networks.

2. Improved DF Protocol for Stringent QoS Guarantees

One major drawback of the original DF relay protocol, which we will call the protocol ($R0$) hereafter, is that it cannot support stringent QoS requirement for any non-zero arrival process. We have the following proposition to formally characterize this problem.

Proposition 9. *As the QoS exponent $\theta\to\infty$, the optimal effective capacity for* *protocol* ($R0$) *approaches zero, no matter how much spectral bandwidth and power resources are assigned for the transmission.*

Proof. The proof is provided in Appendix N. □

To be more specific, Proposition 9 states that the zero-outage capacity of protocol ($R0$) is zero. Intuitively, from Eq. (7.2) we can observe that the performance of the protocol ($R0$) is upper-bounded by the direct transmission from the source to the relay. However, it is well known that when each terminal has a single antenna, direct transmission cannot achieve zero outage probability with a finite average power limitation. Therefore, in order to improve the performance of DF relay for stringent QoS guarantees, we need modify the original protocol ($R0$).

a. Protocol ($R1$)

A straight-forward idea of modification is to improve the performance of DF relay under the case of $\gamma_1 > \gamma_2$ (i.e., **Scenario-1**). Since in this case, the performance of DF relay is always worse than the direct transmission, we can use direct transmission instead of relay.

The optimal resource allocation policy for protocol ($R1$) can be described as follows.

- If $\gamma_1 > \gamma_2$, then the resource allocation is determined by Eq. (7.12).
- Otherwise, the resource allocation is determined by Eq. (7.26).

Unfortunately, even under this revision, the resulting DF protocol still cannot support stringent QoS requirement as $\theta \to \infty$, as proved by Appendix N.

b. Protocol ($R2$)

The major reason that protocol ($R1$) still cannot support stringent QoS is that it does not provide *diversity* for the link from relay to the destination. To overcome this problem, we need to use the direct transmission instead of relay when either γ_2 or γ_3 is less than γ_1. This strategy in fact provide *selection diversity* to the R-D link. The optimal resource allocation policy for protocol ($R2$) can be described as follows.

- If $\gamma_1 > \gamma_2$ or $\gamma_1 > \gamma_3$, then the resource allocation is determined by Eq. (7.12).

- Otherwise, the resource allocation is determined by Eq. (7.26).

Proposition 10. *As the QoS exponent $\theta \to \infty$, the optimal effective capacity for protocol ($R2$) approaches a non-zero constant rate with a finite mean total power constraint given by Eq. (7.3).*

Proof. The proof is provided in Appendix N. □

A similar protocol is called "opportunistic cooperative" in [26], where the focus is mainly on the outage probability minimization.

E. Fixed Power Allocation for DF Relay Networks

In previous sections, we assume that the source and the relay nodes can dynamically allocate the transmit power under a mean total power constraint. In this section, we study the case where they do not have the ability to perform temporal power allocation. Let $P_s = \kappa\overline{P}$ and $P_r = (1-\kappa)\overline{P}$ be the power assigned to source and relay, respectively, where $\kappa \in (0, 1)$. Then, our goal is to find an optimal κ that can maximize the effective capacity under the constraint of a given QoS exponent θ.

Rewrite the rate in Eq. (7.2) for the DF protocol as follows:

$$\begin{aligned}R_{DF}(\kappa,\boldsymbol{\gamma}) &= \left(\frac{T_fB}{2}\right)\min\Big\{\log_2\Big(1+2\gamma_2\overline{P}\kappa\Big),\log_2\Big(1+2\gamma_1\overline{P}\kappa+2\gamma_3\overline{P}(1-\kappa)\Big)\Big\}\\ &= \left(\frac{T_fB}{2}\right)\log_2\Big(1+2\overline{P}\min\Big\{\gamma_2\kappa,\gamma_1\kappa+\gamma_3(1-\kappa)\Big\}\Big)\\ &= \left(\frac{T_fB}{2}\right)\log_2\Big(1+\widetilde{\gamma}\Big) \qquad (7.30)\end{aligned}$$

where we define $\widetilde{\gamma}=2\overline{P}\min\{\gamma_2\kappa,\gamma_1\kappa+\gamma_3(1-\kappa)\}$. Following the similar idea to the previous sections, we can formulate the optimization problem as

$$\begin{aligned}(P3)\qquad &\max_{\kappa\in(0,1)}\left\{-\frac{1}{\theta}\log\bigg(\mathbb{E}_{\boldsymbol{\gamma}}\Big[\exp\Big(-\theta R_{DF}(\kappa,\boldsymbol{\gamma})\Big)\Big]\bigg)\right\}\\ &=\max_{\kappa\in(0,1)}\left\{-\frac{1}{\theta}\log\Big(\mathcal{G}\left(\kappa,\theta,\boldsymbol{\lambda},\overline{P}\right)\Big)\right\}. \qquad (7.31)\end{aligned}$$

In Eq. (7.31), we define

$$\mathcal{G}\left(\kappa,\theta,\boldsymbol{\lambda},\overline{P}\right)=\int_0^{+\infty}\Big(1+\widetilde{\gamma}\Big)^{-\frac{\beta}{2}}p_{\widetilde{\Gamma}}(\widetilde{\gamma})d\widetilde{\gamma} \qquad (7.32)$$

where $\boldsymbol{\lambda}=(\lambda_1,\lambda_2,\lambda_3)$ with λ_i denoting the parameter of the exponential distribution for the channel gain γ_i, and $p_{\widetilde{\Gamma}}(\widetilde{\gamma})$ denotes the probability density function (pdf) of $\widetilde{\gamma}$, which can be derived by using the following proposition.

Proposition 11. *Let $X\triangleq\min\{X_2,X_1+X_3\}$, where X_i is an independent exponential distributed random variable with parameter μ_i for $i\in\{1,2,3\}$. Then the pdf of X, denoted by $p_X(x)$, can be expressed as follows:*

$$p_X(x)=\begin{cases}\frac{\mu_3(\mu_1+\mu_2)}{\mu_3-\mu_1}e^{-(\mu_1+\mu_2)x}+\frac{\mu_1(\mu_2+\mu_3)}{\mu_1-\mu_3}e^{-(\mu_2+\mu_3)x}, & \mu_1\neq\mu_3\\ \left[\mu_2+\mu_1(\mu_1+\mu_2)x\right]e^{-(\mu_1+\mu_2)x}, & \mu_1=\mu_3.\end{cases} \qquad (7.33)$$

Proof. The proof follows by the direct derivations. □

Corollary 3. *The pdf of $\widetilde{\gamma}$, denoted by $p_{\widetilde{\Gamma}}(\widetilde{\gamma})$, can be obtained directly by letting*

$\mu_1 = \lambda_1/(2\overline{P}\kappa)$, $\mu_2 = \lambda_2/(2\overline{P}\kappa)$, and $\mu_3 = \lambda_3/[2\overline{P}(1-\kappa)]$ in Eq. (7.33).

The integral in Eq. (7.32) can be calculated by using the results given in [74, Sec. 3.383.5]. After some algebraic manipulations, we can obtain the closed-form expression for $\mathcal{G}\left(\kappa,\theta,\boldsymbol{\lambda},\overline{P}\right)$ given in Eq. (7.32) as follows:

$$\begin{aligned}\mathcal{G}\left(\kappa,\theta,\boldsymbol{\lambda},\overline{P}\right) &= -\frac{\lambda_1[\lambda_2(1-\kappa)+\lambda_3\kappa]}{2\overline{P}\kappa[\lambda_3\kappa-\lambda_1(1-\kappa)]}\exp\left(\frac{\lambda_2(1-\kappa)+\lambda_3\kappa}{2\overline{P}\kappa(1-\kappa)}\right)\\ &\quad\cdot E_{\frac{\beta}{2}}\left(\frac{\lambda_2(1-\kappa)+\lambda_3\kappa}{2\overline{P}\kappa(1-\kappa)}\right)\\ &\quad+\frac{(\lambda_1+\lambda_2)\lambda_3}{2\overline{P}[\lambda_3\kappa-\lambda_1(1-\kappa)]}\exp\left(\frac{\lambda_1+\lambda_2}{2\overline{P}\kappa}\right)E_{\frac{\beta}{2}}\left(\frac{\lambda_1+\lambda_2}{2\overline{P}\kappa}\right)\end{aligned} \tag{7.34}$$

if $(1-\kappa)\lambda_1 \neq \kappa\lambda_3$. Otherwise, if $(1-\kappa)\lambda_1 = \kappa\lambda_3$, we get

$$\begin{aligned}\mathcal{G}\left(\kappa,\theta,\boldsymbol{\lambda},\overline{P}\right) &= \frac{\lambda_2}{2\overline{P}\kappa}\exp\left(\frac{\lambda_1+\lambda_2}{2\overline{P}\kappa}\right)E_{\frac{\beta}{2}}\left(\frac{\lambda_1+\lambda_2}{2\overline{P}\kappa}\right)+\frac{\lambda_1(\lambda_1+\lambda_2)}{4\overline{P}^2\kappa^2}\left[\frac{\Gamma\left(\frac{\beta}{2}-2\right)}{\Gamma\left(\frac{\beta}{2}\right)}\right.\\ &\cdot {}_1F_1\left(2,3-\frac{\beta}{2},\frac{\lambda_1+\lambda_2}{2\overline{P}\kappa}\right)+\Gamma\left(2-\frac{\beta}{2}\right)\left(\frac{\lambda_1+\lambda_2}{2\overline{P}\kappa}\right)^{\frac{\beta}{2}-2}{}_1F_1\left(\frac{\beta}{2},\frac{\beta}{2}-1,\frac{\lambda_1+\lambda_2}{2\overline{P}\kappa}\right)\left.\vphantom{\frac{\Gamma}{\Gamma}}\right]\end{aligned} \tag{7.35}$$

where $\Gamma(\cdot)$ denotes the Gamma function, $E_\nu(\cdot)$ denotes the νth order exponential integral function, and ${}_1F_1(\cdot,\cdot,\cdot)$ denotes the confluent hypergeometric function.

At the high SNR regime, using the Taylor expansion $e^x = 1 + x + x^2/2 + \cdots$, and the asymptotic property of the exponential integral function [109]

$$\lim_{z\to 0} E_\nu(z) = \Gamma(1-\nu)z^{\nu-1} - \frac{1}{1-\nu} \tag{7.36}$$

it turns out that Eq. (7.34) can be simplified as

$$\begin{aligned}\lim_{\overline{P}\to\infty}\mathcal{G}\left(\kappa,\theta,\boldsymbol{\lambda},\overline{P}\right) &= \frac{\lambda_2}{\overline{P}\kappa(\beta-2)}+\frac{\kappa(1-\kappa)\Gamma\left(1-\frac{\beta}{2}\right)}{\lambda_3\kappa-\lambda_1(1-\kappa)}\\ &\quad\cdot\left[\frac{\lambda_3}{1-\kappa}\left(\frac{\lambda_1+\lambda_2}{2\overline{P}\kappa}\right)^{\frac{\beta}{2}}-\frac{\lambda_1}{\kappa}\left(\frac{\lambda_2(1-\kappa)+\lambda_3\kappa}{2\overline{P}\kappa(1-\kappa)}\right)^{\frac{\beta}{2}}\right].\end{aligned} \tag{7.37}$$

However, it is still difficult to solve the optimal κ^* explicitly, due to the intractability

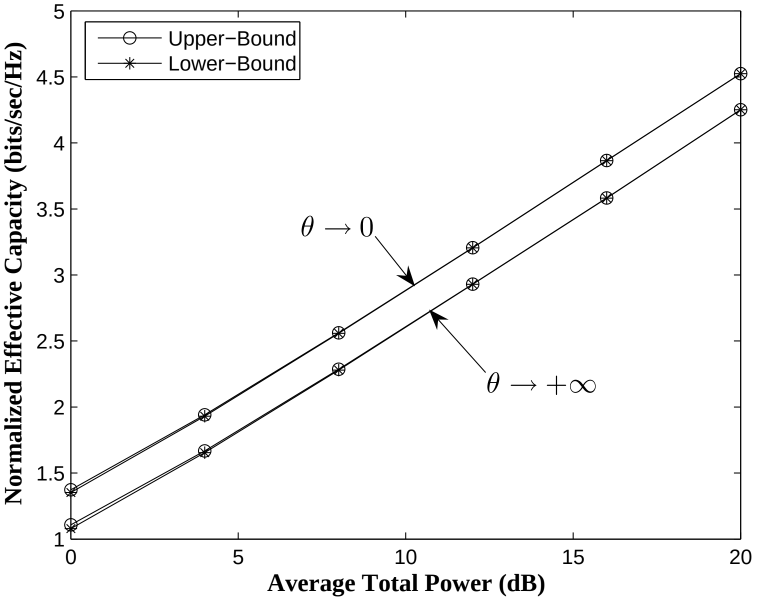


Fig. 45. The effective capacity upper-bound and lower-bound for AF protocol under optimal resource allocation policies. The S-R distance is set to $d = 0.5$.

of $\mathcal{G}\left(\kappa, \theta, \boldsymbol{\lambda}, \overline{P}\right)$. Therefore, we will use the numerical search to obtain the optimal $\kappa^* \in (0, 1)$.

F. Simulations and Numerical Evaluations

We evaluate the performance of our proposed cross-layer resource allocation scheme over wireless relay networks by simulations and numerical analyses. In the following, we set the product of frame duration and spectral bandwidth $T_f B / \log(2) = 1$ such that $\beta = \theta$. The other system parameters are detailed respectively in each of the figures.

Using Propositions 7 and 8, Fig. 45 plots the effective capacity upper- and lower-bounds for the AF protocol by using our proposed resource allocation policy. We can observe from Fig. 45 that for both the loose QoS constraint ($\theta \to 0$) and stringent

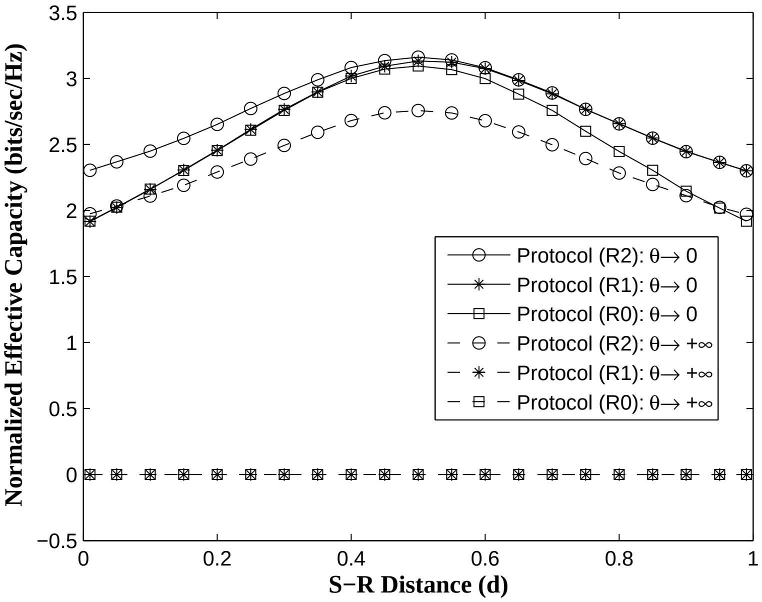
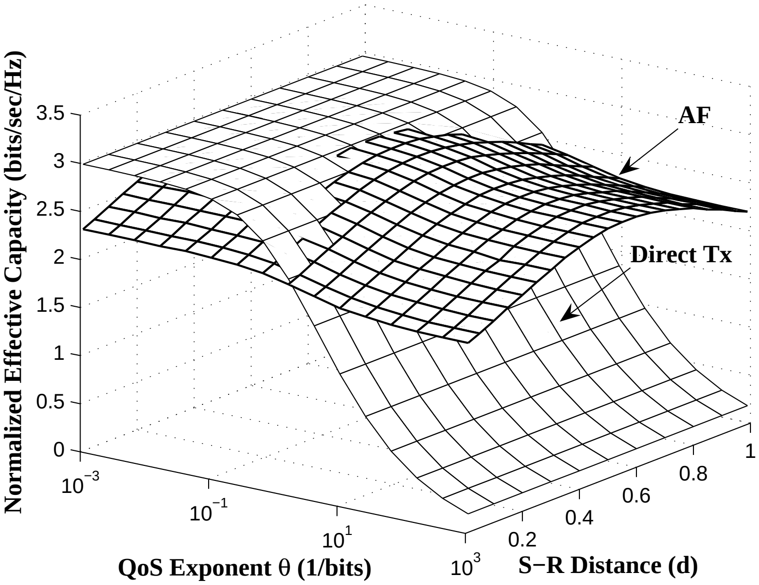


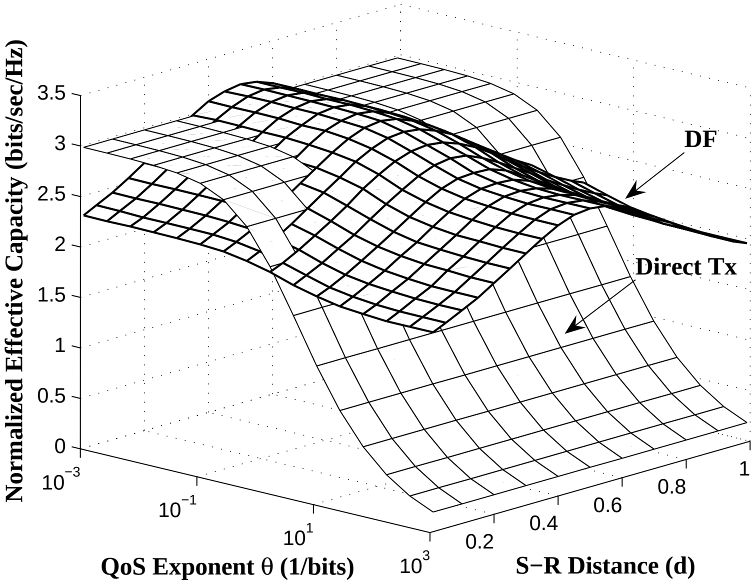
Fig. 46. The effective capacity of different DF protocols under optimal resource allocation. The average total power is 10 dB.

QoS constraint ($\theta \to \infty$), the upper- and lower-bounds are very close to each other. In particular, at the high SNR regime, the upper- and lower-bounds are indistinguishable, indicating that the deployed resource allocation policy essentially achieves the optimality. Thus, in the following, we will only plot the lower-bound of the effective capacity for AF relay protocol for simplicity.

Fig. 46 plots the optimal effective capacities for different DF relay protocols. As verified by Fig. 46, when the QoS is loose ($\theta \to 0$), protocol ($R2$) achieves the best effective capacity and the original protocol ($R0$) attains the worst performance among the three protocols. The performance of protocol ($R1$) approaches protocol ($R0$) when the relay is close to the source and approaches protocol ($R2$) when the relay is close to the destination. On the other hand, as the QoS constraint becomes stringent ($\theta \to \infty$), only protocol ($R2$) can achieve a nonzero effective capacity. For both protocols ($R0$) and ($R1$), the optimal effective capacity approaches zero, as pointed



(a) AF protocol.



(b) DF protocol.

Fig. 47. The effective capacity of AF and DF schemes under optimal resource allocation policies. The average total power is 10 dB.

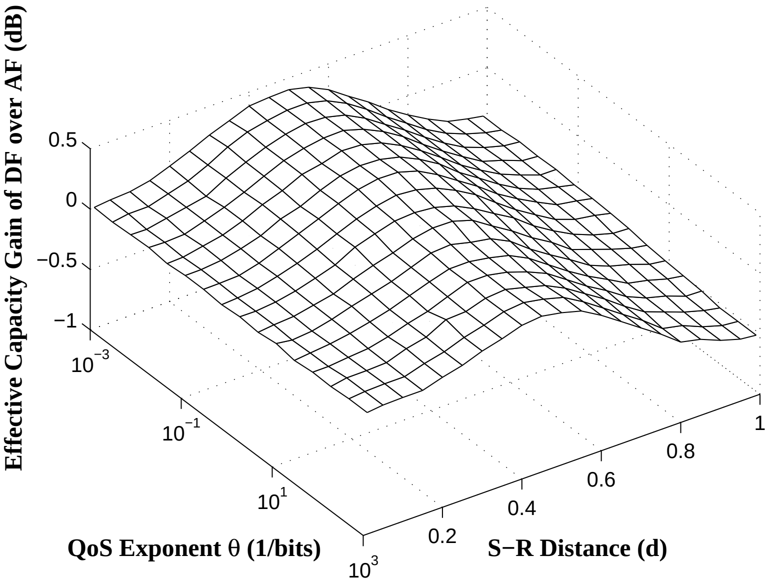


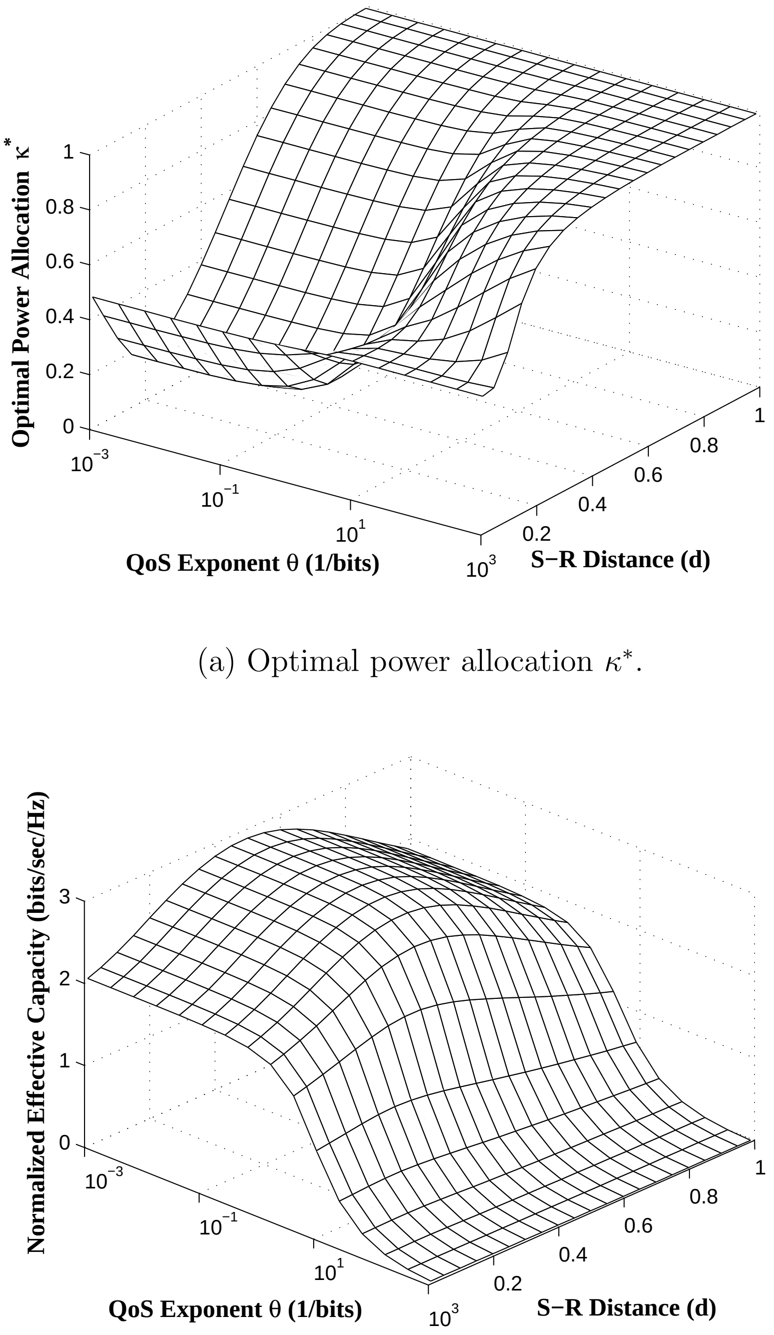
Fig. 48. The effective capacity gain ratio of DF protocol over AF protocol under optimal resource allocation. The average total power is set equal to 10 dB.

out by our analytical analyses. Thus, in the following, we will only plot the effective capacity by using protocol $(R2)$.

Fig. 47 shows the optimal effective capacities for AF and DF protocols as a function of the QoS exponent θ and the S-R distance d. For comparison purpose, we also plot the optimal effective capacity obtained by using direct transmission proposed in Chapter III. Since there is no relay node for the direct transmission scheme, the performance of direct transmission is independent of parameter d. We can observe from Fig. 47 that when the QoS constraint is loose, the effective capacities of AF and DF are close to that of direct transmission. However, as the QoS constraint become more stringent, the two relay protocols both show significant advantages over direct transmission. The performance comparisons of AF and DF relay are plotted in Fig. 48 in terms of the ratio of the effective capacity for DF relay to that for AF relay. We can observe that the two protocols show the similar performances. DF relay performs relatively better when the relay is close to the source, while AF relay performs better when the relay is close to the destination. Note that when the QoS constraint is loose, the direct transmission may outperform relay transmission, which is due to the fact that both relay protocols operate in half-duplex mode and only utilize half of the degree of freedom. When more powerful relay protocols are employed, e.g., the relay protocols proposed in [26, 105], the performance of relay transmission is expected to show significant gain over direct transmission for both loose and stringent QoS provisioning.

Fig. 49(a) numerically plots the optimal power assignment κ^* for DF relay protocol under fixed power allocation policy. By applying the optimal power assignment, the resulting effective capacity is shown in Fig. 49(b). We can observe from Fig. 49(b) that even using the optimal power assignment, the effective capacity also converges to zero as the QoS exponent $\theta \to \infty$. However, this is the best that the fixed power allocation can do to maximize the effective capacity. This implies that no matter how much power and spectral bandwidth resource are assigned and no matter how elegant coding/modulation is employed, the fixed power allocation cannot support stringent QoS over Rayleigh flat-fading channels.

To compare the performance of dynamic and static resource allocations, the effective capacity gain of protocol $(R2)$ over fixed power assignment is shown in Fig. 50. We can see from Fig. 50 that for loose QoS constraint, the performance of the two strategies are similar. The effective capacity by using dynamic resource allocation is only slightly better than that by using fixed power assignment. However, as the QoS constraint becomes stringent, the dynamic resource allocation significantly outperforms the fixed power allocation, which confirms the importance of employing the dynamic resource allocation for stringent QoS provisioning.



(a) Optimal power allocation κ^*.

(b) Optimal effective capacity.

Fig. 49. The effective capacity of DF relay with optimal fixed power allocation. The average total power is equal to 10 dB.

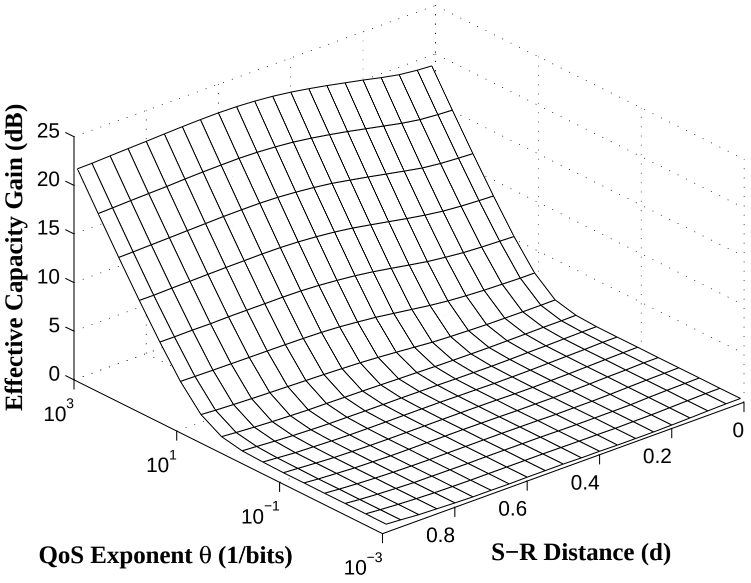


Fig. 50. The effective capacity gain of dynamic resource allocation over fixed power allocation for DF protocol. The average total power is 10 dB.

G. Summary

We proposed and analyzed the cross-layer resource allocation scheme for relay networks to guarantee diverse QoS requirements. By integrating information theory with the effective capacity, our proposed cross-layer scheme characterizes the delay QoS constraint by the QoS exponent, which turns out to be a simple and efficient approach in the cross-layer design and optimization. Over both AF and DF wireless relay networks, we developed the associated resource allocation algorithms. The simulation and numerical results verified that our proposed cross-layer resource allocation can efficiently support diverse QoS requirements. On the other hand, even for simple AF and DF protocols, the relay transmission shows significant advantages over direct transmissions when the delay constraints become stringent.

In this chapter, our focus is mainly on how to apply the effective-capacity-based approach to wireless relay networks as an efficient cross-layer design strategy, whereas the problem of what is the optimal relay protocol is beyond the scope of this chapter. It is worth noting that the performances of different relay protocols are significantly different. When employing more powerful relay protocols, e.g., those proposed and studied in [26,105], the network performance can be much better. However, the cross-layer resource allocation scheme developed in this chapter can be readily extended to the other scenarios using the more powerful relay protocols.

Chapters II–VIII form the first part of this dissertation. Specifically, we investigated QoS-driven resource allocation schemes of single channel, multichannel, cellular networks, and cooperative relay networks, respectively, for statistical QoS guarantees. From the next chapter, we will turn our attention to the second part of this dissertation, and study adaptive antenna selection for MIMO communication systems and networks.

CHAPTER VIII

SELECTION DIVERSITY WITH MRC FOR MC DS-CDMA WIRELESS NETWORKS

A. Introduction

Recently, multicarrier (MC) direct-sequence (DS) code-division multiple access (CDMA) that integrates the advantages of orthogonal frequency division multiplexing (OFDM) with DS-CDMA has emerged as a promising technique for the next generation wireless communications and networks [110–114]. MC DS-CDMA can significantly alleviate the impacts of frequency-selective fading by mapping the serial data flow into a number of low-rate parallel substreams and transmitting the time-domain spread signals over multiple orthogonal subcarriers. The authors in [110] showed that when appropriately selecting the system parameters and using the *antenna diversity*, MC DS-CDMA is capable of supporting ubiquitous broadband wireless services over diverse propagation environments.

The antenna diversity technique, on the other hand, making use of multiple antennas at transmitter and/or receiver, is another effective approach to combat the time-varying fading channels. Among the large number of antenna diversity schemes, the selection diversity (SD)/maximal-ratio combining (MRC) offers a good tradeoff between complexity and performance and thus received a great deal of research attentions [38–42]. It is natural to consider integrating SD/MRC with MC DS-CDMA to further improve the system performance. However, how to combine these two tech-

©2006 IEEE. Reprinted, with permission, from "Transmit selection diversity with maximal-ratio combining for multicarrier DS-CDMA wireless networks over Nakagami-m fading channels" by J. Tang and X. Zhang, published in *IEEE Journal on Selected Areas in Communications*, vol. 24, no. 1, pp. 104–112, Jan. 2006.

niques in the most efficient way, and how the combination can impact the wireless network performance have not been thoroughly studied.

The SD/MRC technique can be considered as a special case of the more general hybrid-selection/maximal-ratio combining (H-S/MRC) schemes [43–49], where a subset of the available antennas are selected at transmitter and/or receiver. The H-S/MRC technique over independent identical fading environment has been studied in [38–47], followed by some more complicated models such as the unequal fading and power models [48] and the correlated fading models [49]. In [44], the authors developed the novel approach termed as *virtual branch* to analyze the performance of H-S/MRC scheme, which transforms the mutually *dependent* order statistics to *independent* virtual branches and thus significantly simplify the analyses. In [47], the authors further studied the impact of channel state information (CSI) estimation error on H-S/MRC scheme. Most previous works in this area mainly focused on H-S/MRC employed at the *receiver* side only [39, 43, 44, 47–49] such that the CSI *feedback* from the receiver to the transmitter *is not needed* to make the antenna selection decision. In contrast, when the selection is applied at the transmitter side, the CSI feedback is necessary and thus the imperfectness of CSI feedback will impair the performance of the H-S/MRC scheme. While the powerful *virtual branch* technique [44] particularly focuses on the *order statistics*, this technique is hard to be extended to our imperfect CSI feedback analyses, specifically, for the delayed-CSI feedback analysis, because our delayed-feedback analysis involves the *induced order statistics* [50]. In [38, 42], the authors investigated the impact of feedback delay on SD/MRC scheme. Also, the authors in [46] studied the impact of feedback error on space-time block coding (STBC)-based antenna-selection scheme proposed in [45]. However, no closed-form symbol-error rate (SER) expressions were obtained in [38, 42, 46].

To make the analyses tractable, in this chapter we focus on the SD-based scheme employed over the independent identical fading channels. We first generalize the SD scheme into MIMO MC DS-CDMA systems. Based on the unique infrastructure of MC DS-CDMA systems, the SD is employed in both spatial and frequency domains, such that the joint *optimal subcarrier-and-antenna pair* is selected for each substream to transmit data. This two-dimensional selection diversity, referred as transmit selection diversity (TSD), offers not only the higher order of selection diversity, but also the lower peak-to-average power ratio (PAPR), which is the main drawback imposed by MC DS-CDMA communications systems [112]. Taking the feedback delays and errors into considerations, we then develop the unified framework to analyze the SER of the proposed TSD/MRC-based MC DS-CDMA scheme over frequency-selective Nakagami-m fading channels. Following the excellent work in [34, 115, 116], we derive the SER's as closed-form expressions. Applying this developed analytical framework, we finally analyze the impact of TSD on MC DS-CDMA systems in various wireless networks. The diverse network architectures we have considered include downlink cellular networks, uplink cellular networks, and Ad Hoc wireless networks. We also compare the performance of TSD/MRC with conventional SD/MRC and STBC/MRC-based MC DS-CDMA systems, through which we gain the insights about how severe the feedback imperfectness can impair the performance of various schemes. Our analyses show that in a wide variation of CSI-feedback imperfectness, TSD/MRC-based scheme has significant advantages over SD/MRC-based and STBC/MRC-based schemes for both downlink cellular networks and Ad Hoc wireless networks. However, our analytical findings also indicate that the TSD/MRC-based scheme cannot always outperform SD/MRC and STBC/MRC for uplink cellular networks even when the perfect CSI feedbacks are available.

The rest of the chapter is organized as follows. Sections B and C derive the SER with perfect and imperfect feedbacks, respectively. Section D generalizes and applies the obtained results to various multiuser MC DS-CDMA wireless networks. Section E numerically evaluates the proposed scheme and compares it with other existing schemes. The chapter concludes with Section F.

B. System Model

1. Transmitter Model

We first consider a point-to-point wireless link with N_t antennas at the transmitter and N_r antennas at the receiver. We denote the average power by P_t, the total number of subcarriers by $U = P \times Q$, where the parameters P and Q will be detailed below, and the central frequencies of these subcarriers by $\{f_0, f_1, ..., f_{U-1}\}$. First, a block of P symbols with each duration of T'_s are converted to P parallel *substreams* using the serial-to-parallel (S/P) converter. The signal over the pth substream is expressed as $s_p(t) = \sum_{i=-\infty}^{+\infty} s_p[i] P_{T_s}(t - iT_s)$, where $p \in [0, P-1]$ is the index of substream, i is the discrete time-index, $s_p[i]$ denotes the transmitted symbol at time i, $T_s = PT'_s$ represents the symbol duration after S/P conversion, and $P_T(\cdot)$ stands for the unit rectangular-pulse function of duration T. Each substream is then multiplied by the spread-code $c(t) = \sum_{j=-\infty}^{+\infty} c[j] P_{T_c}(t - jT_c)$, where $c[j]$ takes the value of $\{+1, -1\}$, and T_c denotes the chip duration which follows $T_c = T_s/N_s$ with N_s the spreading gain.

For conventional MC DS-CDMA systems [110], the P substreams are then further copied to Q parallel *branches* (also known as the identical-bit subcarriers [113]) and transmitted simultaneously. In contrast, in our proposed system, the transmitter *jointly* selects the optimal branch-and-antenna pair in frequency and spatial domains for each substream. Specifically, for the pth substream, the transmitter will select the optimal antenna among N_t antennas and the optimal subcarrier among Q branches

$\{f_{qP+p}\}_{q=0}^{Q-1}$ to maximize the received power, where $q \in [0, Q-1]$ denotes the index of the branch. Finally, the modulation can be implemented by the U-points inverse fast Fourier transform (IFFT).

2. Channel Model

In this chapter, we consider the generalized frequency-selective Nakagami-m fading channels because this model is very general and often best fits the land-mobile and indoor-mobile multipath propagations [34, 43, 115–117]. We assume that the system parameters T_c and P are designed such that $T_M < T_c \leq PT_m$ [110], where T_m and T_M denote the minimum and the maximum delay spreads of the channel, respectively. Under such an assumption, on one hand, each subcarrier signal is guaranteed to experience the flat-fading; on the other hand, the Q-branches of the same substream are ensured to experience the independent fading such that the frequency diversity is achieved.

The channel impulse response function, denoted by $h_{ij}^u(t)$, between the ith transmit antenna ($i \in [1, N_t]$) and the jth receive antenna ($j \in [1, N_r]$) at the uth subcarrier ($u \in [0, U-1]$) can be expressed as

$$h_{ij}^u(t) = \alpha_{ij}^u \delta(t - t_0) \exp\left(-\jmath\, \phi_{ij}^u\right) \tag{8.1}$$

where $\jmath \overset{\triangle}{=} \sqrt{-1}$, t_0 is the path-delay, $\{\alpha_{ij}^u\}$ is the set of path-envelopes, and $\{\phi_{ij}^u\}$ is the set of path-phases. We assume that $\{\phi_{ij}^u\}$ are independent identically distributed (i.i.d.) random variables (r.v.'s) uniformly distributed between $[0, 2\pi)$, and $\{\alpha_{ij}^u\}$ are *i.i.d.* r.v.'s. The common probability density function (pdf) of $\{\alpha_{ij}^u\}$, denoted by $p_\alpha(\alpha)$, follows the Nakagami-m distribution specified by

$$p_\alpha(\alpha) = 2\left(\frac{m}{\Omega}\right)^m \frac{\alpha^{2m-1}}{\Gamma(m)} \exp\left(-m\frac{\alpha^2}{\Omega}\right) \tag{8.2}$$

where $\Omega = E\left\{[\alpha_{ij}^u]^2\right\}$ is the average path-gain and $\Gamma(x)$ denotes the Gamma function. Let $\lambda_{ij}^u \triangleq \left(\alpha_{ij}^u\right)^2$ denote the corresponding path-power. When employing TSD, the transmitter will select, for the pth substream, the optimal branch-and-antenna pair, indexed by $\{q^*, i^*\}$, that maximizes the total received power. We denote the subcarrier central frequency, the path-power, the path-envelope, and the path-phase corresponding to the selected pth substream by $f_{(p)}$, $\lambda_{ij}^{(p)}$, $\alpha_{ij}^{(p)}$ and $\phi_{ij}^{(p)}$, respectively. Thus, the received signal, denoted by $r_j(t)$, at the jth antenna of the receiver is given by

$$r_j(t) \;=\; \sqrt{\frac{P_t}{P}} \sum_{p=0}^{P-1} \alpha_{ij}^{(p)} s_p(t-t_0) c(t-t_0) \exp\left(\jmath\left[2\pi f_{(p)} t + \varphi_{ij}^{(p)}\right]\right) + n_j(t) \quad (8.3)$$

where $\varphi_{ij}^{(p)} = -2\pi f_{(p)} t_0 - \phi_{ij}^{(p)}$, and $n_j(t)$ denotes the complex additive white Gaussian noise (AWGN) with zero-mean and double-sided power spectral density of $N_0/2$ per dimension.

3. Receiver Model

We assume that the receiver has the perfect knowledge of CSI. The received signal $r_j(t)$ first correlates with the referenced waveform $c(t-t_0)$, the output of which, denoted by $Y_{ij}^{(p)}$, can be expressed as

$$Y_{ij}^{(p)} \;=\; \sqrt{\frac{1}{T_s}} \int_{t_0}^{t_0+T_s} r_j(t)\, \alpha_{ij}^{(p)}\, c(t-t_0) \exp\left(-\jmath\left[2\pi f_{(p)} t + \varphi_{ij}^{(p)}\right]\right) dt. \quad (8.4)$$

Then, the outputs $\{Y_{ij}^{(p)}\}_{j=1}^{N_r}$ of the correlators from different receive antennas are combined together, i.e., $Y_p \triangleq \sum_{j=1}^{N_r} Y_{ij}^{(p)}$. Let the pth substream's SNR at the MRC combiner output be denoted by γ_p, then $\gamma_p = (E_s/N_0) \sum_{j=1}^{N_r} \lambda_{ij}^{(p)}$, where $E_s = P_t T_s'$ denotes the average transmission energy per symbol. It is well known that when selection diversity is not employed, the pdf and CDF of the SNR at MRC combiner output, denoted by $p_\Gamma(\gamma)$ and $P_\Gamma(\gamma)$, respectively, are specified by

$$\begin{cases} p_\Gamma(\gamma) = \frac{\gamma^{L_r-1}}{\Gamma(L_r)} \left(\frac{m}{\overline{\gamma}}\right)^{L_r} \exp\left(-\frac{m\gamma}{\overline{\gamma}}\right) \\ P_\Gamma(\gamma) = \frac{\gamma\left(L_r, \frac{m\gamma}{\overline{\gamma}}\right)}{\Gamma(L_r)} \end{cases} \tag{8.5}$$

where $L_r \triangleq mN_r$, $\gamma(x,z) = \int_0^z t^{x-1}e^{-t}dt$ denotes the lower incomplete Gamma function [118], and $\overline{\gamma} = E_s\Omega/N_0$ represents the average SNR per receive antenna.

4. PAPR Discussions for MC DS-CDMA Systems

One of the main problems inherently associated with MC DS-CDMA communications is the high PAPR at the output signal [112]. It is well known that the peak-factor of the multicarrier signal is proportional to the number of used subcarriers [119]. Furthermore, in conventional MC DS-CDMA systems, the frequency-repeated signals transmitted on Q identical-bit subcarriers are not independent with each other, which will result in the even severer PAPR problem [112].

In our proposed scheme, although it has N_t transmit antennas and $U = P \times Q$ subcarriers, the transmitter actually sends the signal by P different paths. In average, the number of used subcarriers at each antenna is equal to P/N_t. In the worst case, the maximum number of used subcarriers at one antenna is P, where all substreams are coincidentally transmitted by this antenna. Furthermore, by using the frequency-domain selection, the proposed scheme avoids sending duplicated Q signals on identical-bit subcarriers. In contrast, the conventional MIMO MC DS-CDMA schemes, such as STBC-based transmitter, always send signals by all $U = P \times Q$ subcarriers at each antenna, with Q copies of each substream signal [110]. On the other hand, the conventional SD[1] schemes cannot avoid the possibility that the same

[1] In this chapter, SD refers to the scheme that employs antenna selection at each subcarrier, with Q identical-bit subcarriers for the same substream.

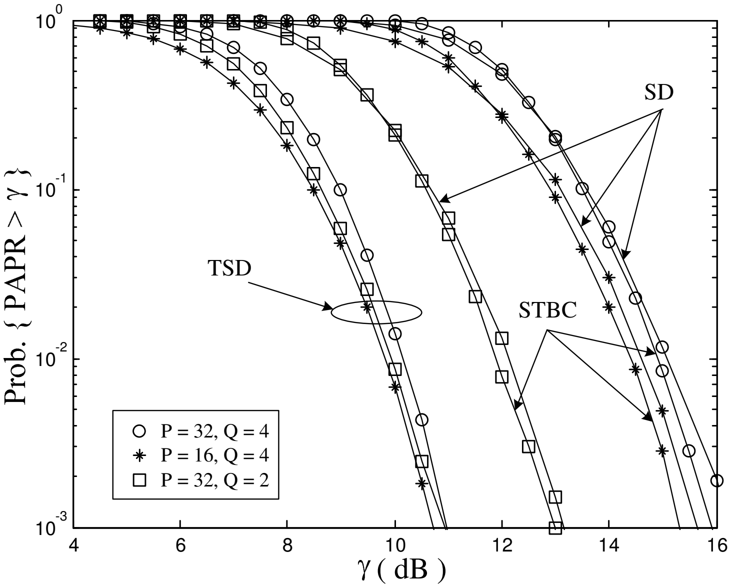


Fig. 51. The PAPR comparisons between our proposed TSD-based MC DS-CDMA scheme and the conventional SD-based and STBC-based MC DS-CDMA schemes using QPSK modulations. The number of transmit antennas is set equal to $N_t = 2$.

transmit antenna sends several branches of the same substream. In the worst case, one transmit antenna has to send the signals for all U subcarriers, which leads to high PAPR. Thus, our proposed transmit selection diversity scheme can significantly decrease the PAPR imposed by the multicarrier DS-CDMA communications systems.

Fig. 51 plots the simulated PAPR comparisons between our proposed TSD/MRC-based scheme and the conventional schemes, including STBC/MRC-based and also SD/MRC-based schemes. We can observe that the PAPR's of our proposed scheme are significantly lower than those of the other two conventional schemes. As shown in Fig. 51, the larger the number $U = P \times Q$ of subcarriers causes the larger PAPR for both conventional and the proposed systems, which is consistent with the well known results given in [119]. However, the PAPR increasing-rate of our proposed TSD/MRC-based scheme is much lower than those of the two conventional schemes.

Furthermore, increasing the frequency-repeating branches from $Q = 2$ to $Q = 4$ dramatically degrades the PAPR-performance for the two conventional schemes. In contrast, the higher order Q of the frequency-diversity guarantees the better PAPR-performance for our proposed scheme.

C. SER With Perfect CSI

The performance of the selection diversity under perfect CSI feedbacks has been extensively studied in literatures, e.g., [41–44]. In this section, we derive the SER within the context of our proposed TSD/MRC-based scheme. For the integer $L_r \triangleq mN_r$, the common pdf of the SNR's $\{\gamma_p\}_{p=0}^{P-1}$, denoted by $f_\Gamma(\gamma)$, follows the ordered Gamma distribution, which can be derived by the similar approach in [43] as

$$f_\Gamma(\gamma) = \frac{L_t}{(L_r-1)!}\sum_{i=0}^{L_t-1}(-1)^i\binom{L_t-1}{i}\exp\left(-(i+1)\frac{m\gamma}{\overline{\gamma}}\right)\sum_{j=0}^{i(L_r-1)}\xi_{ji}\left(\frac{m}{\overline{\gamma}}\right)^{j+L_r}\gamma^{j+L_r-1} \tag{8.6}$$

where $L_t \triangleq QN_t$ and ξ_{ji} is the multinomial coefficients by $\xi_{ji} = \sum_{x=a}^{b}\xi_{x(i-1)}/(j-x)!$, with $a = \max\{0, j-(L_r-1)\}$, $b = \min\{j, (i-1)(L_r-1)\}$, $\xi_{j0} = \xi_{0i} = 1$, $\xi_{j1} = 1/(j!)$, and $\xi_{1i} = i$. Then, using the alternative representation of the Gaussian Q-function [34, 115, 116], the SER, denoted by P_M, for a set of commonly used signal constellations, can be expressed as follows:

$$P_M = \int_0^{+\infty}\left[\frac{\kappa}{\pi}\int_0^{\Theta}\exp\left(-\frac{\beta\gamma}{\sin^2\theta}\right)d\theta\right]f_\Gamma(\gamma)\,d\gamma = \frac{\kappa}{\pi}\int_0^{\Theta}I_{[L_t]}(\overline{\gamma},\beta,\theta)d\theta \tag{8.7}$$

where we define

$$I_{[L_t]}(\overline{\gamma},\beta,\theta) \triangleq \int_0^{+\infty}\exp\left(-\frac{\beta\gamma}{\sin^2\theta}\right)f_\Gamma(\gamma)d\gamma. \tag{8.8}$$

In Eq. (8.7), κ, β, and Θ are constellation-dependent parameters (see [34] for details). Solving $I_{[L_t]}(\overline{\gamma}, \beta, \theta)$ given in Eq. (8.8) using the approach proposed in [115], we obtain its closed-form expression as follows:

$$\begin{aligned} I_{[L_t]}(\overline{\gamma}, \beta, \theta) &= \frac{L_t}{(L_r-1)!} \sum_{i=0}^{L_t-1} (-1)^i \binom{L_t-1}{i} \sum_{j=0}^{i(L_r-1)} \xi_{ji} \\ & \quad \times (j+L_r-1)! \left[\frac{m \sin^2\theta}{\beta\overline{\gamma} + (i+1)m\sin^2\theta} \right]^{j+L_r} . \end{aligned} \tag{8.9}$$

Then, the closed-form expression for P_M given in Eq. (8.7) can be derived using [34, Appendix 5A], which is omitted here for lack of space.

D. SER With Imperfect CSI

1. The SER With Time-Delayed CSI Feedbacks

The impact of feedback delay on the selection diversity has been studied in [42]. However, the authors in [42] did not obtain any closed-form SER expressions. We assume in this section that the feedback a transmitter receives is error-free, but experiences a time-delay, denoted by τ. Due to the time-varying nature of the wireless channel, the current optimal SNR γ_p may have changed already at the moment when the transmitter receives the feedback after the delay τ. Let $\widetilde{\gamma}_p$ denote the time-delayed version of SNR for the original optimal γ_p. According to the order statistics, $\widetilde{\gamma}_p$ is called the *induced order statistics* (or the *concomitant*) [50] of the original ordered γ_p. Thus, the pdf of $\widetilde{\gamma}_p$, denoted by $f_{\widetilde{\Gamma}}(\widetilde{\gamma})$, is given by

$$f_{\widetilde{\Gamma}}(\widetilde{\gamma}) = \int_0^{+\infty} p_{\widetilde{\Gamma}|\Gamma}\left(\widetilde{\gamma}|\gamma\right) f_{\Gamma}(\gamma) d\gamma \tag{8.10}$$

where $p_{\widetilde{\Gamma}|\Gamma}\left(\widetilde{\gamma}|\gamma\right)$ denotes the pdf of $\widetilde{\gamma}$ conditioned on γ, where $\widetilde{\gamma}$ and γ are two correlated Gamma-distribution r.v.'s. According to [43], the conditional pdf $p_{\widetilde{\Gamma}|\Gamma}\left(\widetilde{\gamma}|\gamma\right)$

can be expressed as

$$p_{\widetilde{\Gamma}|\Gamma}(\widetilde{\gamma}|\gamma) = \frac{1}{(1-\rho)}\left(\frac{m}{\overline{\gamma}}\right)\left(\frac{\widetilde{\gamma}}{\rho\gamma}\right)^{\frac{L_r-1}{2}} \exp\left(-\frac{m(\rho\gamma+\widetilde{\gamma})}{(1-\rho)\overline{\gamma}}\right) I_{L_r-1}\!\left(\frac{2m\sqrt{\rho\gamma\widetilde{\gamma}}}{(1-\rho)\overline{\gamma}}\right) \quad (8.11)$$

where $I_\nu(\cdot)$ denotes the modified Bessel function of the first kind with the order of ν, the correlation coefficient ρ is determined by $\rho = J_0^2(2\pi f_d \tau)$ [39] with $J_0(\cdot)$ denoting the zeroth-order Bessel function of the first kind, and f_d representing the Doppler frequency. Thus, considering CSI feedback delays, we obtain the time-delay impacted SER, denoted by $P_M^{(d)}$, as follows:

$$P_M^{(d)} = \int_0^{+\infty}\left[\frac{\kappa}{\pi}\int_0^{\Theta}\exp\left(-\frac{\beta\widetilde{\gamma}}{\sin^2\theta}\right)d\theta\right] f_{\widetilde{\Gamma}}(\widetilde{\gamma})\, d\widetilde{\gamma} = \frac{\kappa}{\pi}\int_0^{\Theta}\widetilde{I}_{[L_t]}(\overline{\gamma},\beta,\theta)d\theta \quad (8.12)$$

where we define the time-delayed version of $I_{[L_t]}(\overline{\gamma},\beta,\theta)$, denoted by $\widetilde{I}_{[L_t]}(\overline{\gamma},\beta,\theta)$, as follows:

$$\widetilde{I}_{[L_t]}(\overline{\gamma},\beta,\theta) \triangleq \int_0^{+\infty}\exp\left(-\frac{\beta\widetilde{\gamma}}{\sin^2\theta}\right) f_{\widetilde{\Gamma}}(\widetilde{\gamma})d\widetilde{\gamma}. \quad (8.13)$$

As derived in the Appendix O, we obtain the closed-form expression of Eq. (8.13) as follows:

$$\begin{aligned}\widetilde{I}_{[L_t]}(\overline{\gamma},\beta,\theta) \;&=\; \frac{L_t}{(L_r-1)!}\sum_{i=0}^{L_t-1}(-1)^i\binom{L_t-1}{i}\sum_{j=0}^{i(L_r-1)}\Bigg\{\xi_{ji}(j+L_r-1)!\sum_{k=0}^{j}\binom{j}{k} \\ &\times\frac{\rho^k(1-\rho)^{j-k}}{[i(1-\rho)+1]^j}\left[\frac{m\sin^2\theta}{(i+1)m\sin^2\theta+[i(1-\rho)+1]\beta\overline{\gamma}}\right]^{k+L_r}\Bigg\}. \end{aligned} \quad (8.14)$$

Similar to Section C, we can also obtain the closed-form expression for the SER $P_M^{(d)}$ by [34, Appendix 5A]. Note that when $\rho = 1$, $\widetilde{I}_{[L_t]}(\overline{\gamma},\beta,\theta)$ derived in Eq. (8.14) reduces to $I_{[L_t]}(\overline{\gamma},\beta,\theta)$ derived in Eq. (8.9), which are expected since $\rho = 1$ corresponds to $\tau = 0$ (i.e., the perfect CSI feedbacks without delay [39]). It is also worth noting that when $L_t = 1$, we have

$$\widetilde{I}_{[1]}(\overline{\gamma}, \beta, \theta) = I_{[1]}(\overline{\gamma}, \beta, \theta) = \left(\frac{m \sin^2 \theta}{\beta \overline{\gamma} + m \sin^2 \theta} \right)^{L_r} \tag{8.15}$$

which is also expected since if no selection diversity is employed ($L_t = 1$), the feedback delay will not affect the performance of our proposed TSD/MRC-based MC DS-CDMA scheme.

2. The SER With Erroneous CSI Feedbacks

The impact of CSI feedback error on antenna selection has been studied in [46], where two out of three transmit antennas are selected to employ STBC. However, no closed-form SER expressions are obtained in [46]. In our scheme, for each substream, the indices of the optimal branch-and-antenna pair $\{q^*, i^*\}$ are decided and fed back to the transmitter by the receiver. In this section, we assume that this feedback is sent to the transmitter without delay, but may be received in error due to the unreliable feedback channel. Let the feedback-frame for each substream be carried by a binary vector with $B = \lceil \log_2(QN_t) \rceil = \lceil \log_2(L_t) \rceil$ bits, representing the index of the optimal subcarrier-and-antenna pair. Thus, a total number of $B \times P$ bits of feedback are needed when using our scheme at each subcarrier.

Let P_e denote the probability of a single-bit feedback error. Under the assumption that the feedback is uncoded and the feedback bit-errors are independent, the probability ε of the feedback-frame error, i.e., the transmitter erroneously selects a certain nonoptimal branch-and-antenna pair for data transmission, is determined by $\varepsilon = 1 - (1 - P_e)^B$.[2] Thus, with the probability of $(1 - \varepsilon)$, the pth substream's feedback

[2] In this chapter, we focus on the system with *uncoded* CSI feedback for derivation convenience. For the system with *coded* CSI feedback, the analysis in this section also holds except that the expression of frame-error rate ε should be changed.

is correct. Then, the transmitter will select the optimal branch-and-antenna pair with the maximal SNR γ_p. However, with the probability of ε, the pth substream's feedback is not correct. Then, the transmitter will select an arbitrary branch-and-antenna pair with its SNR being the nonmaximum. The set of those nonmaximum SNR's is the *complementary* set of the maximal SNR. Using the order statistics [50], the pdf of a *random sample* within those SNR's, denoted by $f_{\Gamma_C}(\gamma)$, can be derived as follows:

$$\begin{aligned} f_{\Gamma_C}(\gamma) &= \left(\frac{1}{L_t-1}\right)\sum_{l=1}^{L_t-1}\frac{L_t!\,p_\Gamma(\gamma)}{(l-1)!(L_t-l)!}\big[P_\Gamma(\gamma)\big]^{l-1}\big[1-P_\Gamma(\gamma)\big]^{L_t-l} \\ &= \frac{L_t p_\Gamma(\gamma)-f_\Gamma(\gamma)}{L_t-1} \end{aligned} \tag{8.16}$$

where $p_\Gamma(\gamma)$ and $P_\Gamma(\gamma)$ are given by Eq. (8.5), and $f_\Gamma(\gamma)$ is given by Eq. (8.6). Thus, we obtain the pdf, denoted by $f_\Gamma^{(e)}(\gamma)$, for the combined SNR when taking the erroneous CSI feedbacks into account, as follows:

$$f_\Gamma^{(e)}(\gamma) = (1-\varepsilon)f_\Gamma(\gamma)+\varepsilon f_{\Gamma_C}(\gamma) = (1-\chi)f_\Gamma(\gamma)+\chi p_\Gamma(\gamma) \tag{8.17}$$

where we further define $\chi \stackrel{\triangle}{=} L_t\,\varepsilon/(L_t-1)$. Thus, considering the CSI feedback error, we get the corresponding SER, denoted by $P_M^{(e)}$, as follows:

$$P_M^{(e)} = \int_0^{+\infty}\left[\frac{\kappa}{\pi}\int_0^{\Theta}\exp\left(-\frac{\beta\gamma}{\sin^2\theta}\right)d\theta\right]f_\Gamma^{(e)}(\gamma)\,d\gamma = \frac{\kappa}{\pi}\int_0^{\Theta} I_{[L_t]}^{(e)}(\overline{\gamma},\beta,\theta)d\theta \tag{8.18}$$

where we define

$$I_{[L_t]}^{(e)}(\overline{\gamma},\beta,\theta) \stackrel{\triangle}{=} \int_0^{+\infty}\exp\left(-\frac{\beta\gamma}{\sin^2\theta}\right)f_\Gamma^{(e)}(\gamma)d\gamma(1-\chi)I_{[L_t]}(\overline{\gamma},\beta,\theta)+\chi I_{[1]}(\overline{\gamma},\beta,\theta) \tag{8.19}$$

where $I_{[L_t]}(\overline{\gamma},\beta,\theta)$ is given by Eq. (8.9) and $I_{[1]}(\overline{\gamma},\beta,\theta)$ is given by Eq. (8.15).

3. The SER With Both Time-Delayed and Erroneous CSI Feedbacks

Applying the analyses above, we derive the SER when *jointly* considering feedback delays and feedback errors. The scenario we consider is as follows. With the probability of $(1-\varepsilon)$, the pth substream's feedback is correct. Since the feedback is also delayed by τ, the transmitter will select the branch-and-antenna pair, which has the delayed version of the maximum-SNR γ_p, i.e., the *concomitant* $\widetilde{\gamma}_p$, to transmit data. On the other hand, with the probability of ε, the pth substream's feedback is received in error. Also considering the delay τ, the transmitter will select an arbitrary branch-and-antenna pair with the delayed version of the nonmaximum SNR, to transmit data. Using the induced order statistics [50], we derive the pdf of the delayed version of the nonmaximum SNR's, denoted by $f_{\widetilde{\Gamma}_C}(\widetilde{\gamma})$, as follows:

$$f_{\widetilde{\Gamma}_C}(\widetilde{\gamma}) = \int_0^{+\infty} p_{\widetilde{\Gamma}|\Gamma}(\widetilde{\gamma}|\gamma) f_{\Gamma_C}(\gamma) d\gamma = \frac{L_t p_{\widetilde{\Gamma}}(\widetilde{\gamma}) - f_{\widetilde{\Gamma}}(\widetilde{\gamma})}{L_t - 1} \tag{8.20}$$

where $p_{\widetilde{\Gamma}}(\widetilde{\gamma})$ is the special case of $f_{\widetilde{\Gamma}}(\widetilde{\gamma})$ with $L_t = 1$. Similar to Section 2, the pdf, denoted by $f_{\widetilde{\Gamma}}^{(e)}(\widetilde{\gamma})$, for the combined SNR when jointly taking both feedback errors and delays into account, is given by

$$f_{\widetilde{\Gamma}}^{(e)}(\widetilde{\gamma}) \overset{\triangle}{=} (1-\varepsilon) f_{\widetilde{\Gamma}}(\widetilde{\gamma}) + \varepsilon f_{\widetilde{\Gamma}_C}(\widetilde{\gamma}) = (1-\chi) f_{\widetilde{\Gamma}}(\widetilde{\gamma}) + \chi p_{\widetilde{\Gamma}}(\widetilde{\gamma}) \tag{8.21}$$

where χ is defined in Eq. (8.17). Using Eq. (8.21), we derive their corresponding SER, denoted by $\widetilde{P}_M^{(e)}$, as

$$\widetilde{P}_M^{(e)} = \int_0^{+\infty} \left[\frac{\kappa}{\pi} \int_0^{\Theta} \exp\left(-\frac{\beta \widetilde{\gamma}}{\sin^2\theta} \right) d\theta \right] f_{\widetilde{\Gamma}}^{(e)}(\widetilde{\gamma})\, d\widetilde{\gamma} = \frac{\kappa}{\pi} \int_0^{\Theta} \widetilde{I}_{[L_t]}^{(e)}(\overline{\gamma}, \beta, \theta) d\theta \tag{8.22}$$

where we define

$$\widetilde{I}_{[L_t]}^{(e)}(\overline{\gamma}, \beta, \theta) \overset{\triangle}{=} \int_0^{+\infty} \exp\left(-\frac{\beta \widetilde{\gamma}}{\sin^2\theta} \right) f_{\widetilde{\Gamma}}^{(e)}(\widetilde{\gamma}) d\widetilde{\gamma} = (1-\chi) \widetilde{I}_{[L_t]}(\overline{\gamma}, \beta, \theta) + \chi \widetilde{I}_{[1]}(\overline{\gamma}, \beta, \theta) \tag{8.23}$$

where $\widetilde{I}_{[L_t]}(\overline{\gamma}, \beta, \theta)$ is given by Eq. (8.14) and $\widetilde{I}_{[1]}(\overline{\gamma}, \beta, \theta)$ is given by Eq. (8.15).

E. Applications to Different Multiuser MC DS-CDMA Wireless Networks

1. Downlink Cellular Networks

In this section, we generalize and apply the proposed TSD/MRC-based MC DS-CDMA scheme to multiuser communications over different types of wireless networks. We first consider the *synchronous* downlink of the cellular networks. We assume that the basestation synchronously transmits signals to K mobile users, and the spread-codes $\{c_k(t)\}_{k=1}^{K}$ assigned to different mobile users are all orthogonal with each other. Since each subcarrier signal experiences the flat-fading, the orthogonality between different spread-codes can still be guaranteed. Therefore, the downlink cellular network has the *near-single-user* performance [110]. As a result, the analytical analyses derived in the previous sections can be directly applied and extended to the downlink cellular networks.[3]

Note that certain nonideal factors, such as the nonzero interferences from neighbor cells and the Doppler frequency-drifts impairing the orthogonality between different subcarriers, may degrade the system performance. In regarding to the first problem, the neighbor-cell interferences can be modeled as part of AWGN using the standard Gaussian approximation, provided that the number of interference terms is sufficiently large. For the second problem, the performance degradation is neglectable, as long as the mobilities of the users are not too high [110].

[3] The results in the previous sections can be considered as the lower-bound of SER's for the downlink cellular networks.

2. Uplink Cellular Networks

We consider the applications of our proposed scheme to *asynchronous* uplink of the cellular networks, where K mobile users asynchronously transmit signals to the basestation. Under the assumption of perfect power control, the received signal at the jth antenna at the basestation can be expressed as

$$r_j(t) = \sqrt{\frac{P_t}{P}} \sum_{k=1}^{K} \sum_{p=0}^{P-1} \alpha_{ij}^{k,(p)} s_{kp}(t - t_0^k) c_k(t - t_0^k) \exp\left(\jmath \left[2\pi f_{k,(p)} t + \varphi_{ij}^{k,(p)}\right]\right) + n_j(t) \tag{8.24}$$

where we use the notations similar to those used in Section B except that the subscript or superscript k is added to distinguish different users. We further assume that no multiuser detection (MUD) [77] techniques are employed. By using the framework developed in [111] and the standard Gaussian approximation, the combined SNR for the kth mobile user of the pth substream, denoted by $\gamma_{k,p}$, can be derived as $\gamma_{k,p} = (E_s/N_e) \sum_{j=1}^{N_r} \lambda_{ij}^{k,(p)}$, where N_e denotes the *effective* AWGN for each substream, which is defined by

$$\frac{N_e}{2} \triangleq \left(\frac{K-1}{Q}\right) \frac{E_s \Omega_e}{N_s} \left(\frac{1}{3} + \frac{1}{U} \sum_{u=0}^{U-1} \sum_{\substack{v=0 \\ v \neq u}}^{U-1} \frac{1}{2\pi^2 (v-u)^2} \right) + \frac{N_0}{2} \tag{8.25}$$

where Ω_e denotes the *effective* average path-gain for each substream at each receive antenna, which is determined by the general expression as follows:

$$\Omega_e = \frac{N_0}{E_s N_r} \int_0^{+\infty} \widehat{\gamma} f_{\widehat{\Gamma}}(\widehat{\gamma}) d\widehat{\gamma} \tag{8.26}$$

where $f_{\widehat{\Gamma}}(\widehat{\gamma})$ needs to be substituted by the specific pdf expressions, depending on either perfect or imperfect CSI feedback being considered. Substituting the appropriate equation into Eq. (8.26) and solving the integral, we obtain the different effective average path-gains Ω_e's. By substituting the appropriate Ω_e into N_e given by Eq. (8.25) and letting the new average signal-to-interference-and-noise Ratio (SINR) be $\overline{\gamma} = E_s\Omega/N_e$, the results derived in Sections C and D can be directly applied to the uplink cellular networks. Applying our proposed scheme over uplink cellular networks, all mobile users select the optimal branch-and-antenna pair transmitting signals to the basestation. As a result, at the basestation, although the received power of the useful signals for each user is enhanced, the strength of interferences is also increased. Noting that the capacity of the cellular networks is interference-limited, the overall uplink performance may degrade when employing the TSD/MRC-based MC DS-CDMA scheme due to the strong interference.

3. Ad Hoc Wireless Networks

In this section, we consider our proposed scheme applied in asynchronous Ad Hoc wireless networks, where K *pairs* of mobile users communicate with each other in each pair asynchronously and independently. Without loss of generality, we focus on the first pair of mobile users in the following analyses. We call the kth pair's transmitter and receiver by "the kth transmitter and kth receiver", respectively. Then, the received signal at the jth antenna of the first receiver, denoted by $r_j^{(1)}(t)$, can be expressed as

$$\begin{aligned} r_j^{(1)}(t) \;&=\; \sqrt{\frac{P_t}{P}} \sum_{k=1}^{K}\sum_{p=0}^{P-1} \alpha_{ij}^{(k,1),(p)}\, s_{kp}(t-t_0^k)c_k(t-t_0^k) \\ &\quad \times \exp\left(\jmath\left[2\pi f_{k,(p)}t + \varphi_{i^*j}^{(k,1),u}\right]\right) + n_j(t) \end{aligned} \qquad (8.27)$$

where superscript $^{(k,1)}$ denotes the channel between the kth transmitter and the first receiver. Comparing Eq. (8.27) with that at uplink cellular networks in Eq. (8.24), we can find that the useful signal parts and the noise terms have the same structure, while the interference terms are different. Noting that although the channel between the first transmitter and the first receiver (i.e., the useful signal channel) is optimal, the channels between the other transmitters and the first receiver (i.e., the interference channels) are *random* and *independent*. Thus, unlike what happening in the uplink cellular networks, the strong interferences will *not be cumulated* at the receiver in Ad Hoc wireless networks. Then, the SNR at the first receiver, denoted by $\gamma_p^{(1)}$, can be derived as $\gamma_p^{(1)} = (E_s/N_e)\sum_{j=1}^{N_r}\lambda_{ij}^{(1,1),(p)}$, where $\lambda_{ij}^{(1,1),(p)}$ denotes the path-power between the first transmitter and the first receiver, and N_e also represents the effective AWGN given by

$$\frac{N_e}{2} \triangleq \left(\frac{K-1}{Q}\right)\frac{E_s\Omega}{N_s}\left(\frac{1}{3}+\frac{1}{U}\sum_{u=0}^{U-1}\sum_{\substack{v=0\\ v\neq u}}^{U-1}\frac{1}{2\pi^2(v-u)^2}\right)+\frac{N_0}{2} \tag{8.28}$$

which differs from Eq. (8.25) only in that the effective average path-gain is Ω in Eq. (8.28), instead of Ω_e used in Eq. (8.25).

F. Numerical Results for Performance Evaluations

Without loss of generality, we evaluate the performance of our proposed scheme using BPSK modulation (with $\kappa = \beta = 1$ and $\Theta = \pi/2$) over Rayleigh fading channels ($m = 1$). The number P of substreams is set to $P = 32$ and the spreading gain N_s is set equal to $N_s = 128$. For comparison and illustration purposes, we plot the SER's of MRC, STBC/MRC-, and SD/MRC-based MC DS-CDMA schemes whenever necessary. Note that both STBC/MRC and MRC are independent of feedbacks' imperfectness.

Fig. 52 plots the exact SER's and the corresponding Chernoff-bounds versus the average SNR $\overline{\gamma}$, where we also derive the closed-form Chernoff-bounds for the

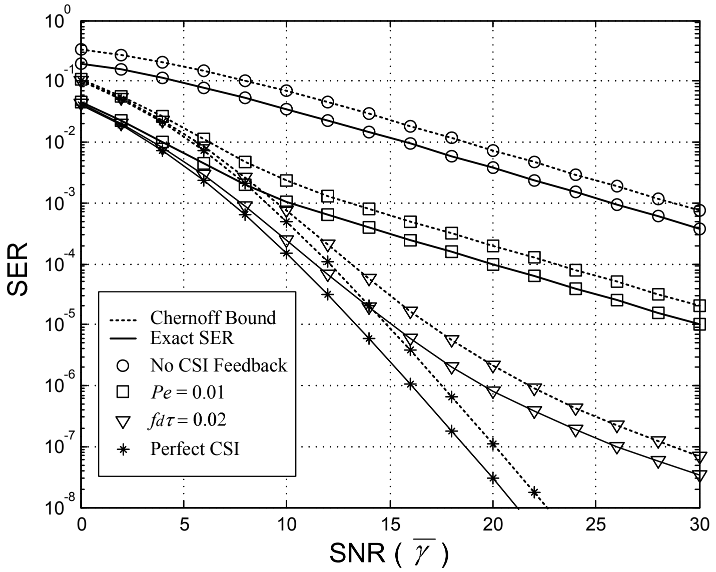
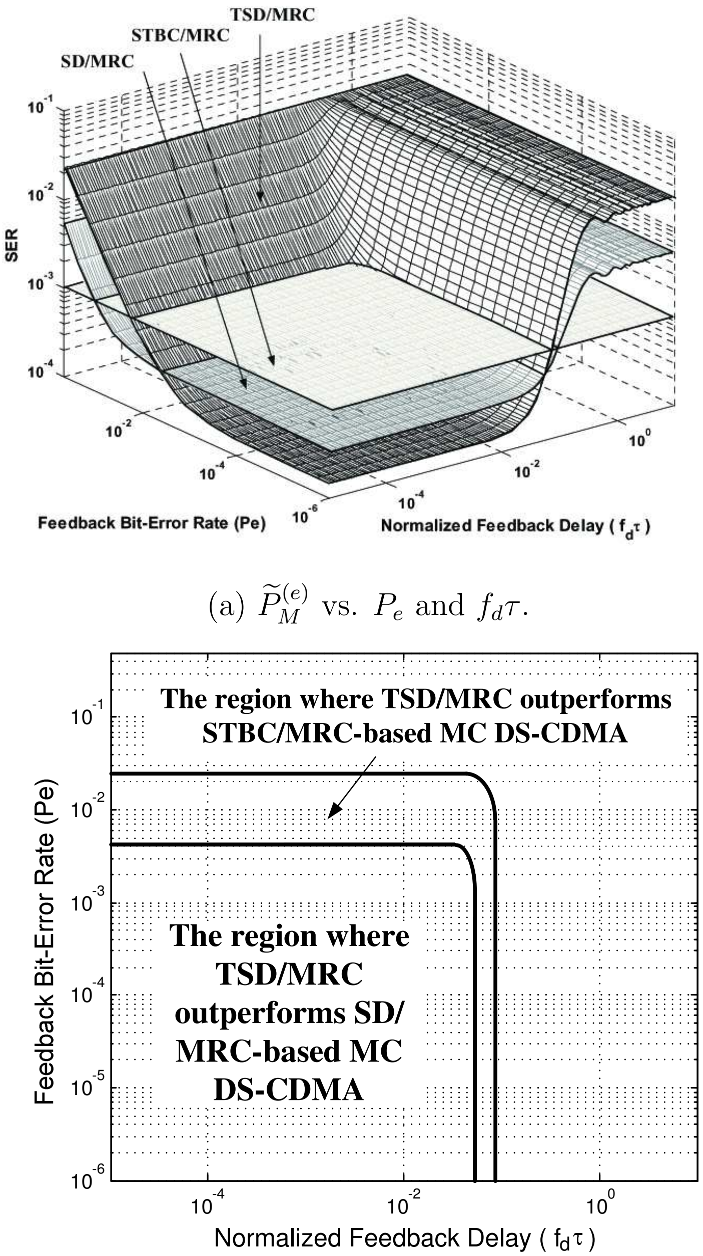


Fig. 52. The SER versus the average SNR $\overline{\gamma}$ with perfect and imperfect CSI feedbacks for our proposed TSD/MRC-based MC DS-CDMA scheme. The number of transmit and receive antennas are set to $N_t = 2$ and $N_r = 1$, respectively. The number of branches per substream is $Q = 2$.

SER's, but omit them for lack of space. As shown in Fig. 52, the feedback delay and the feedback error can significantly impact the SER of TSD/MRC-based scheme, especially for the high SNR's. The plots marked with "no CSI feedback" means that the transmitter selects the antennas and subcarriers arbitrarily, which corresponds to the worst SER performance of the proposed scheme.

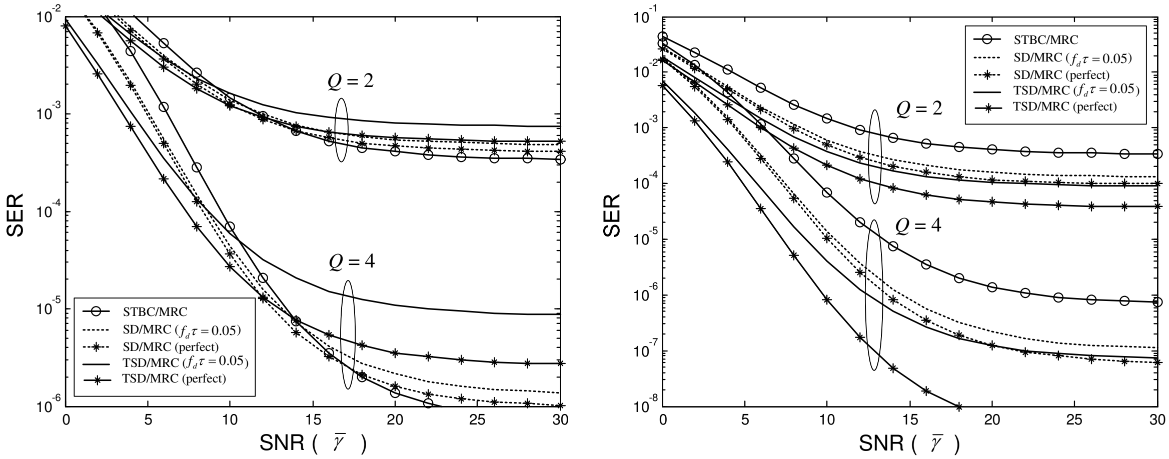
Fig. 53(a) plots the SER $\widetilde{P}_M^{(e)}$ of TSD/MRC scheme when both feedback delays $f_d\tau$ and feedback errors P_e vary, characterizing $\widetilde{P}_M^{(e)}$'s general dynamics. Fig. 53(b) plots the projections of the two intersecting lines where the SER of TSD/MRC scheme is equal to SER's of SD/MRC and STBC/MRC schemes, respectively. The regions within the projected lines determine the variation region of the tolerable imperfect feedbacks that ensure TSD/MRC outperforming SD/MRC and STBC/MRC, respectively. We can also see from Fig. 53(b) that in a wide range of feedback imperfectness,



(a) $\widetilde{P}_M^{(e)}$ vs. P_e and $f_d\tau$.

(b) Projection of SER onto $(P_e, f_d\tau)$.

Fig. 53. The SER $\widetilde{P}_M^{(e)}$ performance of our proposed TSD/MRC MC DS-CDMA scheme when jointly considering feedback delays and errors. $N_t = 2$, $N_r = 1$, and $Q = 2$.



(a) Uplink wireless networks.

(b) Ad Hoc wireless networks.

Fig. 54. The SER performance of TSD/MRC-based MC DS-CDMA scheme over uplink and Ad Hoc wireless networks. $N_t = 2$, $N_r = 2$, and $K = 60$.

the TSD/MRC scheme outperforms the SD/MRC and STBC/MRC schemes.

The results above can also be considered as the approximations of SER's over the downlink cellular networks. In Fig 54, we evaluate the performance of TSD/MRC in uplink cellular networks and Ad Hoc wireless networks, respectively. We assume that we can upper-bound the imperfectness of the feedback errors to make the SER's virtually unchanged with P_e, such that only the feedback delay can impact the SER performance. Note that we set the feedback delay equal to $f_d\tau = 0.05$, which can significantly deteriorate the SER performance of TSD/MRC scheme.

Fig. 54(a) plots the SER against average SNR $\overline{\gamma}$ in uplink cellular networks. We can see from Fig. 54(a) that TSD/MRC-based schemes cannot guarantee the performance superiority over SD/MRC-based and STBC/MRC-based schemes. The feedback delay further degrades SER of TSD/MRC-based and SD/MRC-based schemes. In contrast, for the SER performance over Ad Hoc wireless networks as shown in Fig. 54(b), TSD/MRC-based schemes always have the better SER performance than the corresponding SD/MRC-based and STBC/MRC-based schemes, respectively, when the feedback is perfect. Even under the large feedback delays, the SER of TSD/MRC still performs the best among the three schemes. The larger the order of the transmit diversity, the higher the superiority of TSD/MRC over the other schemes.

G. Summary

We proposed the scheme that integrates TSD/MRC with MC DS-CDMA for diverse wireless networks. We also developed the analytical framework to analyze the SER's of the proposed scheme over Nakagami-m fading channels when taking feedback delays and errors into considerations. The proposed scheme can significantly decrease the PAPR that is inherently associated with MC DS-CDMA communications systems. The resultant SER's are compared with those of SD/MRC-based and STBC/MRC-based MC DS-CDMA schemes in different wireless-network scenarios. Our analyses showed that in a wide variation of feedback imperfectness, the proposed TSD/MRC-based MC DS-CDMA scheme is better applicable to both downlink cellular networks and Ad Hoc wireless networks. However, the analyses also indicated that TSD/MRC-based MC DS-CDMA scheme cannot always outperform SD/MRC-based and STBC/MRC-based MC DS-CDMA schemes in uplink cellular networks due to the imposed stronger interference.

In this chapter, we focus on only selecting the optimal *one* antenna for MIMO communications. In the next chapter, we will relax this constraint and study an STBC based antenna selection scheme.

CHAPTER IX

ALAMOUTI SCHEME WITH SELECTION DIVERSITY

A. Introduction

Recently, the selection diversity (SD)-based multiple-input-multiple-output (MIMO) systems received a great deal of research attentions [37, 41–49, 120]. Using the SD-based technique, a *subset* of the available antennas are selected at transmitter and/or receiver for high-efficient wireless transmissions, which achieves a good tradeoff between cost and performance.

A number of interesting transmit SD schemes are developed and investigated in literatures [37, 41, 42, 45, 46, 120]. In particular, the authors of [45] proposed the approach that integrates the SD with Alamouti transmit diversity, where two out of all transmit antennas are selected to transmit data using space-time block coding (STBC) [35, 36]. In this chapter, we refer this scheme as selection diversity (SD)-STBC. In [120], the authors derived the closed-form expression of bit-error rate (BER) for SD-STBC scheme when assuming perfect channel-state information (CSI) feedbacks. In [46], the authors studied the impact of CSI feedback error on SD-STBC scheme, but the BER is expressed by infinite integrals with no closed-form equations obtained. In [42] and [37], the authors investigated the simpler SD scheme (i.e., no STBC is involved) when assuming perfect and delayed CSI feedbacks. However, applying SD scheme, the transmitter can only select the *single* optimal antenna, which is easier to analyze than the SD-STBC scheme.

©2005 IEEE. Reprinted, with permission, from "Alamouti scheme with joint antenna selection and power allocation over Rayleigh fading channels in wireless networks" by J. Tang and X. Zhang, published in *Proc. IEEE GLOBECOM*, St. Louis, MO, Nov. 2005, pp. 3319–3323.

In [44], the authors developed the approach termed as *virtual branch* to analyze the error performance of SD scheme. While the virtual branch technique using the *order statistics* can be employed to obtain the error performance for SD-STBC with perfect CSI feedbacks, this technique is hard to be extended to the analyses with delayed CSI feedbacks, because delayed feedback analyses involve the *induced order statistics* [50]. It is also worth noting that in [121], the authors did excellent work in analyzing the error performance of closed-loop transmit diversity. However, the focus of [121] was not selection-diversity based closed-loop transmit diversity.

To make the analyses tractable, in this chapter, we focus on the SD-STBC employed over the independent identical fading channels. We investigate the error performance of the SD-STBC scheme while jointly taking the power allocation into account, such that the transmitter adaptively allocates transmit power among the selected antennas to minimize the symbol-error rate (SER). We derive the SER as either the closed-form expressions or the single-fold finite integral when assuming perfect and delayed CSI feedbacks, respectively. Our results show that when the CSI feedback is perfect, the optimal power allocation is to assign all power to the single optimal antenna, such that the SD-STBC reduces to the SD. On the other hand, when taking the CSI feedback delay into account, the SD-STBC scheme with dynamic power allocation ensures the better SER performance than the conventional SD scheme and SD-STBC scheme with equal-power (EP) allocation.

The rest of the chapter is organized as follows. Section B describes the system model. Section C derives the SER when the feedback is perfect. Section D derives the SER with the delayed feedbacks. Section E presents the numerical results of the SER and discusses the optimal power allocation strategy. The chapter concludes with Section F.

B. System Model

We consider a point-to-point wireless link over flat-fading Rayleigh channel with N ($N \geq 2$) antennas at the transmitter and L ($L \geq 1$) antennas at the receiver. The complex channel gain between the ith transmit antenna and the jth receive antenna is denoted by $h_{ij}[n]$, where $i \in [1, N]$, $j \in [1, L]$, and n is the discrete time-index. The channel gains are modeled as stationary and ergodic random processes, the marginal distributions of which are assumed to follow the independent identically distributed (i.i.d.) Gaussian with zero-mean and variance of $\Omega/2$ per dimension. The receiver is assumed to have perfect knowledge of the CSI. Thus, the indices of the optimal two transmit antennas, which maximize the total received power, can be obtained and fed back to the transmitter. Based on this partial CSI feedback, the transmitter selects the optimal two antennas out of N candidates to transmit data using STBC. Since our discussion focuses on the level of the symbol duration, we omit the time-index n in the rest of the chapter for simplicity.

Let $\gamma_{(i)}$ denote the SNR at the maximum-ratio combining (MRC) output by only using the ith transmit antenna, which can be expressed as

$$\gamma_{(i)} = \left(\frac{E_s}{N_0}\right) \sum_{j=1}^{N_r} \left|h_{ij}\right|^2 \tag{9.1}$$

where E_s denotes the average energy per symbol. Sort $\{\gamma_{(i)}\}_{i=1}^{N}$ from the highest SNR to the lowest SNR, as follows:

$$\gamma_1 \geq \gamma_2 \geq ... \geq \gamma_N. \tag{9.2}$$

When employing SD-STBC, the transmitter selects the two antenna which can maximize the SNR. Thus, the SNR at MRC output, denoted by $\gamma_{\mathrm{SD-STBC}}$, can be expressed as

$$\gamma_{\mathrm{SD-STBC}} = \alpha\gamma_1 + \beta\gamma_2 \tag{9.3}$$

where $0 \leq \alpha \leq 1$ denotes the percentage of power that allocates to the optimal antenna, and $\beta \overset{\triangle}{=} 1 - \alpha$ is the percentage of power that assigns to the suboptimal antenna.

C. SER With Perfect CSI

1. The Derivations of Joint Ordered PDF

In this section, we assume that the CSI feedback is perfect, i.e., there is no feedback delay considered. Over the flat-fading Rayleigh channel, the SNR $\gamma_{(i)}$ follows the central χ^2 distribution with degree of freedom equal to $2L$. The probability density function (pdf) and cumulative density function (CDF) of $\gamma_{(i)}$, denoted by $p_\Gamma(\gamma)$ and $P_\Gamma(\gamma)$, respectively, can be expressed as

$$p_\Gamma(\gamma) = \frac{\lambda^L}{(L-1)!}\gamma^{L-1}e^{-\lambda\gamma} \tag{9.4}$$

and

$$P_\Gamma(\gamma) = 1 - e^{-\lambda\gamma}\sum_{i=0}^{L-1}\frac{(\lambda\gamma)^i}{i!}, \tag{9.5}$$

respectively, where $\lambda = 1/\overline{\gamma}$ with $\overline{\gamma}$ denoting the average SNR per receive antenna. Using the order statistics [50], the joint pdf of γ_1 and γ_2, denoted by $f_{\Gamma_1,\Gamma_2}(\gamma_1,\gamma_2)$ can be expressed as

$$
\begin{aligned}
f_{\Gamma_1,\Gamma_2}(\gamma_1,\gamma_2) &= 2\binom{N}{2} p_\Gamma(\gamma_1) p_\Gamma(\gamma_2) \left[P_\Gamma(\gamma_2)\right]^{N-2} \\
&= \frac{N(N-1)}{\left[(L-1)!\right]^2} \gamma_1^{L-1} e^{-\lambda\gamma_1} \sum_{i=0}^{N-2} (-1)^i e^{-(i+1)\lambda\gamma_2} \binom{N-2}{i} \sum_{j=0}^{i(L-1)} \xi_{ji} \lambda^{j+2L} \gamma_2^{j+L-1}
\end{aligned}
\tag{9.6}
$$

where ξ_{ji} is specified by $\xi_{ji} = \sum_{k=a}^{b} \frac{1}{(j-k)!} \xi_{k(i-1)}$, with $a = \max\{0, j-(L-1)\}$, $b = \min\{j, (i-1)(L-1)\}$, $\xi_{j0} = \xi_{0i} = 1$, $\xi_{j1} = 1/j!$, and $\xi_{1i} = i$ [43].

2. SER Derivations

Using the alternative representation of the Gaussian Q-function [34], the SER, denoted by P_M, for a set of commonly used signal constellations, can be expressed as follows:

$$
P_M = \frac{\kappa}{\pi} \int_0^{\Theta} \int_0^{\infty} \int_0^{\gamma_1} f_{\Gamma_1,\Gamma_2}(\gamma_1,\gamma_2) \exp\left(-\frac{g(\alpha\gamma_1+\beta\gamma_2)}{\sin^2\theta}\right) d\gamma_2 d\gamma_1\, d\theta \tag{9.7}
$$

where κ, g, and Θ are constellation-dependent parameters (see [34] for details). Plugging Eq. (9.6) into Eq. (9.7), the SER P_M seems to be complicated since it involves three folds of integral. However, as shown in the Appendix P, the inner two-fold integrals of Eq. (9.7) can be derived analytically. Thus, the SER P_M with perfect CSI feedbacks can be expressed as

$$
\begin{aligned}
P_M = \frac{\kappa N(N-1)}{\pi\left[(L-1)!\right]^2} \sum_{i=0}^{N-2} \Bigg\{ \binom{N-2}{i} \sum_{j=0}^{i(L-1)} \xi_{ji} \sum_{k=0}^{L-1} (-1)^{i+k} \binom{L-1}{k} \\
\cdot \frac{(j+2L-1)!}{(j+k+L)} \int_0^{\Theta} \Phi_{ijk}(\lambda \sin^2\theta) d\theta \Bigg\}
\end{aligned}
\tag{9.8}
$$

where

$$\Phi_{ijk}(x) = \frac{x^{j+2L}[(i+1)x + g\beta]^k}{(x+g\alpha)^L[(i+2)x + g(\alpha+\beta)]^{j+k+L}}. \tag{9.9}$$

Using Eq. (9.8), the SER P_M can be easily evaluated using numerical solutions. Furthermore, by adjusting the power ratio α and β, we can investigate the optimal power allocation strategy to minimize the SER of SD-STBC scheme. Intuitively, since $\alpha\gamma_1 + \beta\gamma_2 \leq \gamma_1$, equality holds if $\alpha = 1$. Thus, the optimal power allocation is to assign all power to the single optimal antenna, such that the SD-STBC reduces to the simpler SD scheme.

D. SER With Delayed CSI

We assume in this section that the feedback a transmitter receives experiences a time-delay, denoted by τ. Due to the time-varying nature of the wireless channel, the current optimal SNR γ_1 and suboptimal SNR γ_2 may have changed at the moment when the transmitter receives the feedback after the delay τ. Let $\widetilde{\gamma}_1$ and $\widetilde{\gamma}_2$ denote the time-delayed version of SNR's for the original optimal γ_1 and suboptimal γ_2, respectively. According to the order statistics, $\widetilde{\gamma}_1$ and $\widetilde{\gamma}_2$ are called the *induced order statistics* (or the *concomitant*) [50] of the original ordered γ_1 and γ_2. The joint pdf of $\widetilde{\gamma}_1$ and $\widetilde{\gamma}_2$, denoted by $f_{\widetilde{\Gamma}_1,\widetilde{\Gamma}_2}(\widetilde{\gamma}_1, \widetilde{\gamma}_2)$, is determined by

$$f_{\widetilde{\Gamma}_1,\widetilde{\Gamma}_2}(\widetilde{\gamma}_1, \widetilde{\gamma}_2) = \int_0^\infty \int_0^{\gamma_1} \prod_{i=1}^{2} p_{\widetilde{\Gamma}|\Gamma}(\widetilde{\gamma}_i|\gamma_i)\, f_{\Gamma_1,\Gamma_2}(\gamma_1,\gamma_2) d\gamma_2 d\gamma_1 \tag{9.10}$$

where $p_{\widetilde{\Gamma}|\Gamma}(\widetilde{\gamma}|\gamma)$ denotes the pdf of $\widetilde{\gamma}$ conditioned on γ, where $\widetilde{\gamma}$ and γ are two unordered SNR's. According to [117], the conditional pdf $p_{\widetilde{\Gamma}|\Gamma}(\widetilde{\gamma}|\gamma)$ is determined by

$$p_{\widetilde{\Gamma}|\Gamma}(\widetilde{\gamma}|\gamma) \;=\; \frac{\lambda}{(1-\rho)}\left(\frac{\widetilde{\gamma}}{\rho\gamma}\right)^{\frac{L-1}{2}} \exp\left(-\frac{\lambda(\rho\gamma+\widetilde{\gamma})}{1-\rho}\right) I_{L-1}\!\left(\frac{2\lambda\sqrt{\rho\gamma\widetilde{\gamma}}}{1-\rho}\right) \tag{9.11}$$

where $I_\nu(\cdot)$ denotes the modified Bessel function of the first kind with the order of ν, the correlation coefficient ρ is determined by $\rho = J_0^2(2\pi f_d \tau)$ [39] with $J_0(\cdot)$ denoting the zeroth-order Bessel function of the first kind, and f_d representing the Doppler frequency. Similar to Section C, using the alternative representation of the Gaussian Q-function, we obtain the SER when considering CSI feedback delays, denoted by $P_M^{(d)}$, as:

$$P_M^{(d)} = \frac{\kappa}{\pi} \int_0^\Theta \int_0^\infty \int_0^\infty f_{\widetilde{\Gamma}_1,\widetilde{\Gamma}_2}(\widetilde{\gamma}_1, \widetilde{\gamma}_2) \exp\left(-\frac{g(\alpha\widetilde{\gamma}_1 + \beta\widetilde{\gamma}_2)}{\sin^2\theta}\right) d\widetilde{\gamma}_2 d\widetilde{\gamma}_1 \, d\theta. \quad (9.12)$$

Substituting Eq. (9.10) into Eq. (9.12), the SER $P_M^{(d)}$ becomes even more complicated than Eq. (9.7) since it involves five-folds of integrals. However, as shown in the Appendix Q, the inner four-fold integrals can be derived analytically. Thus, the SER $P_M^{(d)}$ can be derived as

$$\begin{aligned} P_M^{(d)} = & \frac{\kappa N(N-1)}{\pi[(L-1)!]^2} \sum_{i=0}^{N-2} \Bigg\{ (-1)^i \binom{N-2}{i} \sum_{j=0}^{i(L-1)} \xi_{ji}(1-\rho)^{j+2L} \\ & \sum_{n=0}^{\infty} \sum_{k=0}^{\infty} \frac{\rho^{k+n}(j+k+n+2L-1)!}{k!n![i(1-\rho)+2]^{j+k+n+2L}(j+k+L)} \\ & \cdot {}_2F_1\left(1, j+k+n+2L; j+k+L+1; \frac{i(1-\rho)+1}{i(1-\rho)+2}\right) \int_0^\Theta \widetilde{\Phi}_{kn}(\lambda \sin^2\theta) d\theta \Bigg\} \end{aligned} \quad (9.13)$$

where

$$\widetilde{\Phi}_{kn}(x) = \left[\frac{x}{x + g\alpha(1-\rho)}\right]^{n+L} \left[\frac{x}{x + g\beta(1-\rho)}\right]^{k+L} \quad (9.14)$$

and ${}_2F_1(\cdot, \cdot; \cdot; \cdot)$ denotes the hypergeometric function [74]. Furthermore, if $\Theta = \pi/2$, which is the commonly used value of Θ for most constellations, the SER $P_M^{(d)}$ can be derived as a closed-form expression using [34, Appendix 5A.7], which is omitted here for lack of space.

Note that Eq. (9.13) involves double-fold infinite summation. However, it can be calculated numerically using the popular softwares such as Matlab and Mathematica. Also similar to Section C, based Eq. (9.13), we can investigate the optimal power allocation strategy to minimize the SER when considering CSI feedback delays.

E. Numerical Results

Without loss of generality, we evaluate the performance of SD-STBC scheme using BPSK modulation. Fig.55 plot the SER performance of the SD-STBC based scheme assuming perfect CSI feedbacks. We can see from Fig. 55 that the SER is monotonically decreasing when increasing the percentage α of power assigned to the optimal antenna. The optimal power allocation is to assign all power to the optimal antenna, such that the SD-STBC scheme reduces to the conventional SD scheme, which is expected since SD concentrates all its power to the optimal antenna while SD-STBC distributes its power among optimal and suboptimal antennas, which will impair its SER performance.

The SD scheme secures the better SER performance than EP-based SD-STBC when the CSI feedback is perfect. On the other hand, the EP-based SD-STBC shows the better robustness than SD, which is illustrated in Fig. 56. As the normalized delay $f_d\tau$ increases, the SER of SD converges to the scheme with no diversity ($N = 1$ and $L = 1$) while SD-STBC converges to that of STBC scheme ($N = 2$ and $L = 1$) since the SD-STBC scheme can at least guarantee the open-loop transmit diversity.

To take the advantages of both SD and SD-STBC, a SD-STBC scheme with adaptive power allocation can be employed, where the transmitter dynamically adjust the power ratio α based on current CSI feedback delay (or equivalently, Doppler frequency) to minimize the SER. A simple approach is to employ the threshold-

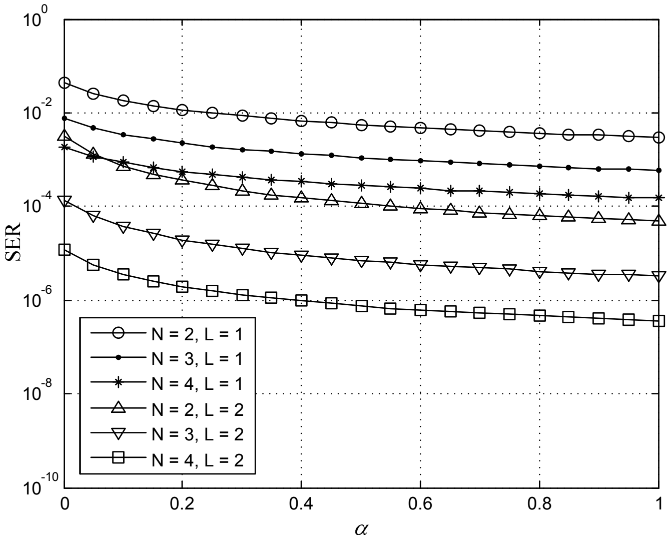


Fig. 55. SER of SD-STBC scheme with power allocations when the CSI feedback is perfect. The average SNR $\overline{\gamma}$ is set equal to 10 dB.

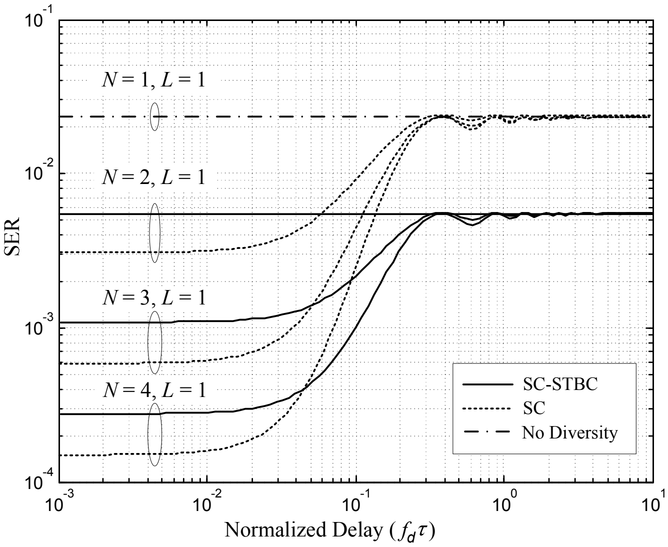


Fig. 56. SER comparison between SD-STBC and SD. The average SNR $\overline{\gamma}$ is set equal to 10 dB. The power is equally distributed ($\alpha = 0.5$).

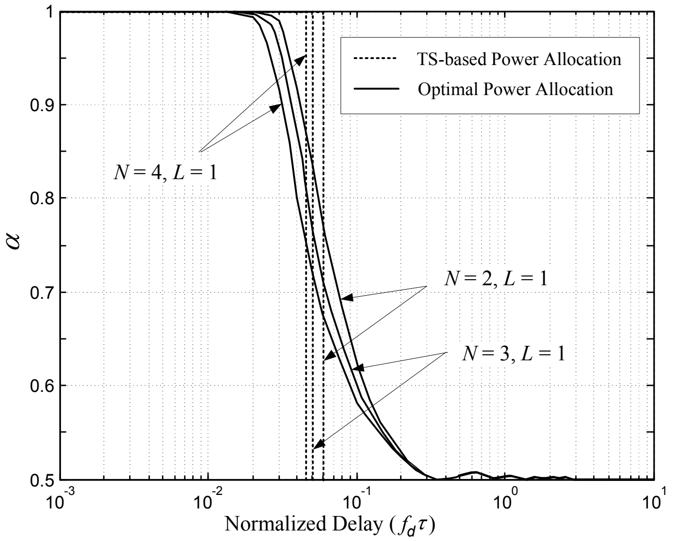


Fig. 57. The power allocation strategies for SD-STBC with the different numbers N's of transmit antennas. The average SNR $\overline{\gamma}$ is set equal to 10 dB.

switching (TS)-based power allocation, the strategy can be described as

$$\alpha = \begin{cases} 1, & \text{SER of SD is better,} \\ 0.5, & \text{SER of EP-based SD-STBC is better.} \end{cases} \tag{9.15}$$

On the other hand, the optimal power allocation, which continuously adjusts the power ratio α, can achieve the optimal SER performance, but requires much higher computational complexity. The TS-based power-allocation and the optimal power-allocation strategies discussed above are summarized and illustrated in Fig. 57, and their corresponding performances in terms of SER's are shown in Fig. 58. We can see from Fig. 58 that the optimal power allocation produces the lower SER than other schemes while the TS-based power allocation performs near optimal in a wide variations of feedback delay values. Thus, it shows a good tradeoff between performance and complexity.

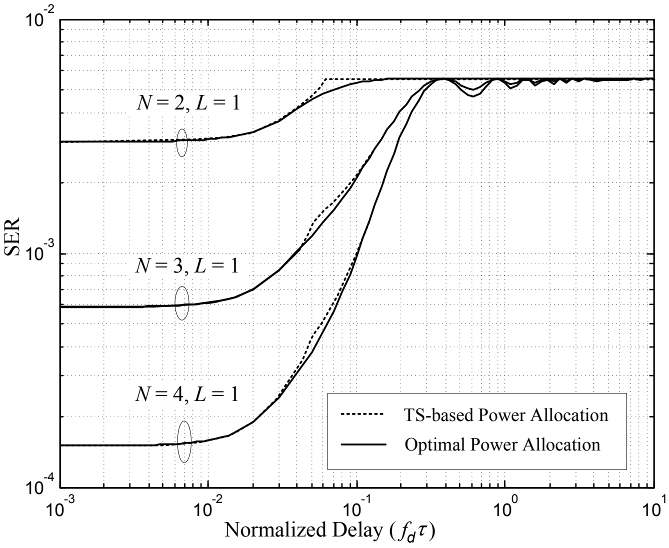


Fig. 58. SER using different power allocation schemes. The average SNR $\overline{\gamma}$ is set equal to 10 dB.

F. Summary

We presented the Alamouti scheme with joint antenna selection and power allocation. We also developed the framework to analyze the error performance of the scheme. For the cases of both perfect and delayed CSI feedbacks, we derived the SER's as either the closed-form expression or the single-fold finite integral. Our results show that when the CSI feedback is perfect, the optimal SD-STBC reduces to SD. When taking the CSI feedback delay into account, the SD-STBC scheme with adaptive power allocation performs better than the conventional SD scheme and EP-based SD-STBC scheme. The analytical optimal power-allocation is currently under our investigations.

CHAPTER X

CONCLUSION

A. Summary of the Dissertation

We considered the problems of the resource allocations for wireless communications and networks. In Chapter I, we introduced and motivated the problems. In Chapter II, we proposed the cross-layer design approach to study the interactions between PHY-layer AMC and MIMO-diversity and higher-protocol-layer on the statistical QoS performance of the mobile wireless networks. We identified the critical relationships between effective bandwidth and effective capacity and analytically obtained the effective capacity function in our proposed system configurations. Our numerical results showed that the AMC and MIMO-diversity employed at physical-layer have significant impact on the statistical QoS performance at upper-protocol-layers. The proposed cross-layer modeling accurately characterize the influence of physical-layer infrastructure on statistical QoS performance at higher-protocol layers.

While in Chapter II we only investigate the single user QoS provisioning, our developed cross-layer modeling technique can be readily extended to the scenarios with multiple users sharing the wireless media in, e.g., dynamic TDMA-based wireless networks. More importantly, our developed cross-layer modeling technique also offers the practical and effective approach to develop the highly-efficient admission-control, packet scheduling, and adaptive resource-allocation schemes to guarantee the QoS for real-time multimedia traffics over mobile wireless networks.

In Chapter III, we proposed and analyzed the QoS-driven resource allocation policies by applying the concept of effective capacity. Our analyses in block fading channel identified the key fact that there exists a fundamental tradeoff between spectral efficiency and QoS provisioning. Depending on the specific QoS requirements, the optimal power-adaptation policy dynamically changes between water-filling and channel inversion. For the more practical adaptive MQAM modulation-based systems, we also developed the corresponding optimal power and rate adaptation scheme. When taking the channel correlation into consideration, we proposed the simple, but efficient, power-control scheme for Markov chain modeled fading channels. The simulation results verified that such an approach can also be applied to the more general channel models.

In Chapter IV, we proposed and analyzed the QoS-driven resource allocation schemes for diversity and multiplexing systems. The proposed resource allocation policies are general and applicable to different fading channel distributions. Our results showed that as the QoS exponent increases from zero to infinity, the optimal effective capacity decreases accordingly from the ergodic capacity to the zero-outage capacity. Moreover, the multichannel transmission provides a significant advantage over single channel transmission for the stringent delay-QoS guarantees. Compared to the single channel transmission which has to deal with the tradeoff between throughputs and delay, the multichannel transmissions can achieve high throughput and stringent QoS at the same time.

In Chapter V, we proposed and analyzed QoS-driven resource allocation over parallel fading channels by taking the imperfect channel estimations into consideration. Solving the original non-convex problem by a 2-dimensional convex optimization approach, we developed the power allocation algorithms for different QoS and power constraints in a general system setting. As the QoS exponent θ increases from zero to infinity, the optimal effective capacity function connects the ergodic capacity with the zero-outage capacity, which is consistent with our previous work in the case of the perfect CSI. Our analyses indicate that the imperfect channel estimations have a significant impact on QoS provisioning, especially when the delay constraint is stringent. In particular, the positive zero-outage capacity is unattainable in the presence of channel estimation errors. On the other hand, our simulation results for the MIMO systems also suggest that a larger number of parallel channels can provide higher throughput and support more stringent QoS, while offering better robustness against the channel estimation errors.

In Chapter VI, we proposed and analyzed a cross-layer-model based adaptive resource-allocation scheme for diverse QoS guarantees over downlink cellular wireless networks. Our scheme jointly allocates power-levels and time-slots for real-time users to guarantee the diverse statistical delay-bound QoS requirements. We developed the admission-control and power/time-slot allocation algorithms. We also studied the impact of adaptive power control and CSI feedback delay at physical-layer on the QoS provisioning performance. Compared to the conventional water-filling and constant power approach, our proposed QoS-driven power adaptation shows significant advantages. The joint power/time-slot allocation scheme can significantly reduce the transmit power, or equivalently, increase the admission region. Also, in an indoor mobile environment, our proposed algorithm is shown to be robust to the CSI feedback delay.

In Chapter VII, we proposed and analyzed the resource allocation scheme for the relay networks to guarantee diverse QoS requirements. Over both AF and DF wireless relay networks, we developed the associated resource allocation algorithms. The simulation and numerical results verified that our proposed cross-layer resource allocation can efficiently support diverse QoS requirements. On the other hand, even for simple AF and DF protocols, the relay transmission shows the significant advantages over direct transmissions when the delay constraints become stringent. While in this chapter, our focus is mainly on how to apply the effective-capacity-based approach to wireless relay networks as an efficient cross-layer design strategy, the problem of what is the optimal relay protocol is beyond the scope of this chapter. It is worth noting that the performances of different relay protocols are significantly different. When employing more powerful relay protocols, the network performance can be much better. However, the cross-layer resource allocation scheme developed in this chapter can be readily extended to the other scenarios using the more powerful relay protocols.

In Chapter VIII, we proposed the scheme that integrates TSD/MRC with MC DS-CDMA for diverse wireless networks. We also developed the analytical framework to analyze the SER's of the proposed scheme over Nakagami-m fading channels when taking feedback delays and errors into considerations. The proposed scheme can significantly decrease the PAPR that is inherently associated with MC DS-CDMA communications systems. The resultant SER's are compared with those of SD/MRC-based and STBC/MRC-based MC DS-CDMA schemes in different wireless-network scenarios. Our analyses showed that in a wide variation of feedback imperfectness, the proposed TSD/MRC-based MC DS-CDMA scheme is better applicable to both downlink cellular networks and ad hoc wireless networks. However, the analyses also indicated that TSD/MRC-based MC DS-CDMA scheme cannot always outperform SD/MRC-based and STBC/MRC-based MC DS-CDMA schemes in the uplink cellular networks due to the imposed stronger interference.

Finally, in Chapter IX, we presented the Alamouti scheme with joint antenna selection and power allocation. We also developed the framework to analyze the error performance of the scheme. For the cases of both perfect and delayed CSI feedbacks, we derived the SER's as either the closed-form expression or the single-fold finite integral. Our results show that when the CSI feedback is perfect, the optimal SD-STBC reduces to SD. When taking the CSI feedback delay into account, the SD-

STBC scheme with adaptive power allocation performs better than the conventional SD scheme and EP-based SD-STBC scheme.

B. Future Work

1. Resource Allocation for Statistical QoS Guarantees

a. QoS Guarantees Over MAC and BC

In this dissertation, we mainly focus on resource allocation over point-to-point communications (i.e., Chapters II–V). For the downlink cellular networks in Chapter VI and the cooperative relay networks discussed in Chapter VII, the key communication component is still a point-to-point model. Under current framework, it is difficult to derive the *optimal* way of resource allocation in the general multiuser communication networks. From information theoretic point-of-view, there exists huge amount of work investigating multiuser information theory [2–5, 122–127], where either multiple access channel (MAC) or broadcast channel (BC) is considered. However, the results in information theory may not be directly applied for our case, where effective capacity is the target objective function, instead of mutual information. In particular, the sum of the effective capacity (compared to the sum of information capacity) is more difficult to deal with. New approach may be required to solve the optimization problem.

b. QoS Guarantees Over Multihop Networks

In addition to the problem of QoS provisioning for multiuser MAC or BC networks, the QoS provisioning over multihop wireless networks (e.g., ad hoc and sensor networks, as well as user cooperation based networks) is also an interesting and challenging direction. In [20], Wu provided the first attempt to solve QoS provisioning problem over tandem networks, which may be used as a basis for the future research. However, the framework developed in [20] is complicated that makes the analytical analysis difficult. In the future work, simplified model may be required in this direction.

c. QoS Guarantees Over Correlated Channels

In this dissertation, we mainly focus on i.i.d. block fading channel model mainly for the analytical convenience. Although we provide an approximation approach to study the correlated block fading channel, it still may not be accurate enough. More accurate channel model can be applied to study the QoS performance of wireless transmissions. In [14], Chang studied the effective bandwidth of various statistical traffic models. Due to the duality between effective bandwidth and effective capacity, the results can be applied for our effective capacity problem. In [128], we use a continuous-time Markov ON-OFF process to model the coded wireless transmission. Future work on this direction may be the employment of more comprehensive Markov model or automatic regression (AR) model. It is worth noting that the more complicated the channel model is, the more difficult to derive any analytical results.

2. Adaptive Antenna Selection

In Chapter IX, our results are expressed as double-folded infinite summation. Although based on the derived expressions, current mathematical softwares can calculate the SER efficiently, no insight can be observed directly from the expression. In the future research, the necessary simplification may be conducted for those complicated expressions. If directly simplification is not possible, tight upper-bound or lower-bound may be derived to not only simplify the expressions, but also provide more insight. Moreover, when jointly considering antenna selection and power allocation, we use the numerical approach to get the optimal resource allocation policy. The analytical optimal power-allocation is currently under our investigations. Also, our previous works only consider i.i.d. fading channel model. The performance of selection diversity over correlated fading channel is another promising area.

REFERENCES

[1] C.-S. Chang, "Stability, queue length, and delay of deterministic and stochastic queueing networks," *IEEE Transactions on Automatic Control*, vol. 39, no. 5, pp. 913–931, May 1994.

[2] D. N. Tse and S. V. Hanly, "Multi-access, fading channels: Part I: Polymatroidal structure, Optimal resource allocation and throughput capacities," *IEEE Transactions on Information Theory*, vol. 44, no. 7, pp. 2796–2815, Nov. 1998.

[3] S. V. Hanly and D. N. Tse, "Multi-access, fading channels: Part II: Delay limited capacities," *IEEE Transactions on Information Theory*, vol. 44, no. 7, pp. 2816–2831, Nov. 1998.

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

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

[6] R. G. Gallager, *Information Theory and Reliable Communication*, New York: Wiley, 1968.

[7] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, New York: Wiley-InterScience, 1991.

[8] I. E. Telatar, "Capacity of multi-antenna gaussian channels," *European Transactions on Telecommunications*, vol. 10, no. 6, pp. 585–595, Nov/Dec 1999.

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

[10] G. Caire, G. Taricco, and E. Biglieri, "Optimum power control over fading channels," *IEEE Transactions on Information Theory*, vol. 45, no. 5, pp. 1468–1489, Jul. 1999.

[11] E. Biglieri, G. Caire, and G. Taricco, "Limiting performance of block-fading channels with mulitple antennas," *IEEE Transactions on Information Theory*, vol. 47, no. 4, pp. 1273–1289, May 2001.

[12] G. Kesidis, J. Walrand, and C.-S. Chang, "Effective bandwidths for multiclass Markov fluids and other ATM sources," *IEEE/ACM Transactions on Networking*, vol. 1, no. 4, pp. 424–428, Aug. 1993.

[13] C.-S. Chang and J. A. Thomas, "Effective bandwidth in high-speed digital networks," *IEEE Journal on Selected Areas in Communications*, vol. 13, pp. 1091–1100, Aug. 1995.

[14] C.-S. Chang, *Performance Guarantees in Communication Networks*, Berlin, Germany: Springer-Verlag, 2000.

[15] F. Kelly, S. Zachary, and I. Ziedins, *Stochastic Networks: Theory and Applications*, vol. 4 of *Royal Statistical Society Lecture Notes Series*, Oxford: Oxford University Press, U.K., 1996.

[16] J. G. Kim and M. Krunz, "Bandwidth allocation in wireless networks with gauranteed packet-loss performance," *IEEE/ACM Transactions on Networking*, vol. 8, no. 3, pp. 337–349, Jun. 2000.

[17] M. Krunz and J. G. Kim, "Fluid analysis of delay and packet discard performance for QoS support in wireless networks," *IEEE Journal on Selected Areas in Communications*, vol. 19, no. 2, pp. 384–394, Feb. 2001.

[18] C. Courcoubetis and R. Weber, "Effective bandwidth for stationary sources," *Probability in Engineering and Information Sciences*, vol. 9, no. 2, pp. 285–294, 1995.

[19] A. I. Elwalid and D. Mitra, "Effective bandwidth of general Markovian traffic sources and admission control of high speed networks," *IEEE/ACM Transactions on Networking*, vol. 1, no. 3, pp. 329–343, Jun. 1993.

[20] D. Wu, "Providing quality of service guarantees in wireless networks," Ph.D. dissertation, Carnegie Mellon University, Pittsburgh, PA, 2003.

[21] D. Wu and R. Negi, "Effective capacity: a wireless link model for support of quality of service," *IEEE Transactions on Wireless Communications*, vol. 2, no. 4, pp. 630–643, July 2003.

[22] D. Wu and R. Negi, "Downlink scheduling in a cellular network for quality-of-service assurance," *IEEE Transactions on Vehicle Technology*, vol. 53, no. 5, pp. 1547–1557, Sep. 2004.

[23] D. Wu and R. Negi, "Utilizing multiuser diversity for efficient support of quality of service over a fading channel," *IEEE Transactions on Vehicle Technology*, vol. 54, no. 3, pp. 1198–1206, May 2005.

[24] G. L. Ghoudhury, D. M. Lucantoni, and W. Whitt, "Squeezing the most out of ATM," *IEEE Transactions on Communications*, vol. 44, no. 2, pp. 203–217, Feb. 1996.

[25] T. Yoo and A. Goldsmith, "Capacity and power allocation for fading MIMO channels with channel estimation error," *IEEE Transactions on Information Theory*, vol. 52, no. 5, pp. 2203–2214, May 2006.

[26] D. Gunduz and E. Erkip, "Opportunistic cooperation by dynamic resource allocation," *IEEE Transactions on Wireless Communications*, accepted.

[27] Q. Liu, S. Zhou, and G. B. Giannakis, "Cross-layer combining of adaptive modulation and coding with truncated ARQ over wireless links," *IEEE Transactions on Wireless Communications*, vol. 3, no. 5, pp. 1746–1755, September 2004.

[28] Q. Liu, S. Zhou, and G. B. Giannakis, "Queuing with adaptive modulation and coding over wireless link: cross-layer analysis and design," *IEEE Transactions on Wireless Communications*, vol. 4, no. 3, pp. 1142–1153, May. 2005.

[29] B. Collins and R. L. Cruz, "Transmission policies for time varying channels with average delay constraints," in *Proc. Allerton International Conference on Communication, Control and Computing*, Monticello, IL, Sep. 1999, pp. 709–717.

[30] D. Rajan, A. Sabharwal, and B. Aazhang, "Delay bounded packet scheduling of bursty traffic over wireless channels," *IEEE Transactions on Information Theory*, vol. 50, no. 1, pp. 125–144, Jan. 2004.

[31] D. Rajan, "Towards universal power efficient scheduling in wireless channels," in *Proc. IEEE ICC*, Paris, France, June 2004, pp. 123–127.

[32] B. Prabhakar, E. Uysal-Biyikoglu, and A. El Gamal, "Energy-efficient transmission over a wireless link via lazy packet scheduling," in *Proc. IEEE INFOCOM*,

Anchorage, AK, Apr. 2001, pp. 386–394.

[33] R. Berry and R. G. Gallager, "Communication over fading channels with delay constraints," *IEEE Transactions on Information Theory*, vol. 48, no. 5, pp. 1135–1149, May 2002.

[34] M. K. Simon and M.S. Alouini, *Digital Communication over Fading Channels: A Unified Approach to Performance Analysis, 2nd ed.*, New York: Wiley, 2005.

[35] S. M. Alamouti, "A simple transmit diversity technique for wireless communications," *IEEE Journal on Selected Areas in Communications*, vol. 16, no. 8, pp. 1451–1458, October 1998.

[36] V. Tarokh, H. Jafarkhani, and A. Calderbank, "Space-time block codes from orthogonal designs," *IEEE Transactions on Information Theory*, vol. 45, no. 5, pp. 1456–1467, July 1999.

[37] J. Tang and X. Zhang, "Transmit selection diversity with maximal-ratio combining for multicarrier DS-CDMA wireless networks over Nakagami-m fading channels," *IEEE Journal on Selected Areas in Communications*, vol. 24, no. 1, pp. 104–112, January 2006.

[38] T. Skinner and J. Cavers, "Selective diversity for Rayleigh fading channels with a feedback link," *IEEE Transactions on Communications*, vol. 21, no. 2, pp. 117–126, Feb 1973.

[39] W. C. Jake Jr., *Microwave Mobile Communications*, New York: Wiley, 1974.

[40] E. A. Neasmith and N. C. Beaulieu, "New results on selection diversity," *IEEE Transactions on Communications*, vol. 46, no. 5, pp. 695–704, May 1998.

[41] Z. Chen, J. Yuan, and B. Vucetic, "Analysis of transmit antenna selection/maximal-ratio combining in Rayleigh fading channels," *IEEE Transactions on Vehicle Technology*, vol. 54, no. 4, pp. 1312–1321, Jul. 2005.

[42] S. Thoen, L. V. Perre, B. Gyselinckx, and M. Engels, "Performance analysis of combined transmit-SC/receive-MRC," *IEEE Transactions on Communications*, vol. 49, no. 1, pp. 5–8, Jan. 2001.

[43] M. S. Alouini and M. K. Simon, "Performance of coherent receivers with hybrid SC/MRC over Nakagami-m fading channels," *IEEE Transactions on Vehicle Technology*, vol. 48, no. 4, pp. 1155–1164, Jul. 1999.

[44] M. Z. Win and J. H. Winters, "Virtual branch analysis of symbol error probability for hybrid selection/maximal-ratio combining in Rayleigh fading," *IEEE Transactions on Communications*, vol. 49, no. 11, pp. 1926 – 1934, Nov 2001.

[45] D. A. Gore and A. J. Paulraj, "MIMO antenna subset selection with space-time coding," *IEEE Transactions on Signal Processing*, vol. 50, no. 10, pp. 2580–2588, Oct. 2002.

[46] W. H. Wong and E. G. Larsson, "Orthogonal space-time block coding with antenna selection and power allocation," *IEE Electronics Letters*, vol. 39, no. 4, pp. 379–381, Feb. 2003.

[47] A. Annamalai and C. Tellambura, "Analysis of hybrid selection/maximal-ratio diversity combiners with Gaussian errors," *IEEE Transactions on Wireless Communications*, vol. 1, no. 3, pp. 498–512, Jul. 2002.

[48] J. Cheng and T. Berger, "Capacity and performance analysis for hybrid selection/maximal-ratio combining in nakagami fading with unequal fading pa-

rameters and branch powers," in *Proc. IEEE ICC*, Seattle, WA, May 2003, vol. 5, pp. 3031–3035.

[49] Y. Chen and C. Tellambura, "Distribution functions of selection combiner output in equally correlated Rayleigh, Rician, and Nakagami-m fading channels," *IEEE Transactions on Communications*, vol. 52, no. 11, pp. 1948–1956, Nov. 2004.

[50] H. A. David, *Order Statistics, 2nd ed.*, New York: Wiley, 1981.

[51] X. Zhang, J. Tang, H.-H. Chen, S. Ci, and M. Guizani, "Cross-layer-based modeling for quality of service guarantees in mobile wireless networks," *IEEE Communications Magazine*, vol. 44, no. 1, pp. 100–106, January 2006.

[52] J. Tang and Xi Zhang, "Cross-layer-based modeling for quality of service guarantees over wireless links," *IEEE Transactions on Wireless Communications*, submitted.

[53] J. Tang and X. Zhang, "Quality-of-service driven power and rate adaptation over wireless links," *IEEE Transactions on Wireless Communications*, accepted.

[54] J. Tang and X. Zhang, "Quality-of-service driven power and rate control in mobile wireless networks," in *Proc. IEEE ICC*, Istanbul, Turkey, June 2006.

[55] J. Tang and X. Zhang, "Quality-of-service driven power and rate adaptation for multichannel communications over wireless links," *IEEE Transactions on Wireless Communications*, accepted.

[56] J. Tang and X. Zhang, "QoS-driven adaptive power and rate allocation for multichannel communications in mobile wireless networks," in *IEEE International*

Symposium on Information Theory (ISIT'06), Seattle, WA, July 2006.

[57] J. Tang and X. Zhang, "QoS-driven power and rate adaptation for multicarrier communications over mobile wireless networks," in *Conference on Information Sciences and Systems (CISS'06)*, Princeton, NJ, March 2006.

[58] J. Tang and X. Zhang, "QoS-driven power allocation over parallel fading channels with imperfect channel estimations in wireless networks," *IEEE Transactions on Wireless Communications*, submitted.

[59] J. Tang and X. Zhang, "QoS-driven power allocation over parallel fading channels with imperfect channel estimations in wireless networks," in *Proc. IEEE INFOCOM*, Anchorage, AL, May 2007.

[60] J. Tang and X. Zhang, "Cross-layer-model based adaptive resource allocation for statistical QoS guarantees in mobile wireless networks," *IEEE Transactions on Wireless Communications*, accepted.

[61] J. Tang and X. Zhang, "Cross-layer resource allocation over wireless relay networks for quality of service provisioning," *IEEE Journal on Selected Areas in Communications*, submitted.

[62] J. Tang and X. Zhang, "Error probability analysis of TAS/MRC-based scheme for wireless networks," in *IEEE Wireless Communications and Networking Conference (WCNC'05)*, March, pp. 877–882.

[63] J. Tang and X. Zhang, "Alamouti scheme with joint antenna selection and power allocation over rayleigh fading channels in wireless networks," in *Proc. IEEE GLOBECOM*, St. Louis, MO, Nov. 2005, pp. 3319–3323.

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

[65] H. S. Wang and N. Moayeri, "Finite-state Markov channel — a useful model for radio communication channels," *IEEE Transactions on Vehicle Technology*, vol. 44, no. 1, pp. 163–171, Feb. 1995.

[66] Q. Zhang and S. A. Kassam, "Finite-state Markov model for Rayleigh fading channels," *IEEE Transactions on Communications*, vol. 47, no. 11, pp. 1688–1692, Nov. 1999.

[67] C.-D. Iskander and P. T. Mathiopoulos, "Analytical level crossing rates and average fading durations for diversity techniques in Nakagami fading channels," *IEEE Transactions on Communications*, vol. 50, no. 8, pp. 1301–1309, Aug. 2002.

[68] C.-D. Iskander and P. T. Mathiopoulos, "Fast simulation of diversity Nakagami fading channels using finite-state Markov models," *IEEE Transactions on Broadcasting*, vol. 49, no. 3, pp. 269–277, Sep. 2003.

[69] Q. Liu, S. Zhou, and G. B. Giannakis, "Cross-layer modeling of adaptive wireless link for QoS support in heterogeneous wired-wireless networks," *ACM/Kluwer Journal of Wireless Networks (WINET)*, vol. 12, pp. 427–437, 2006.

[70] B. Maglaris, D. Anastassiou, P. Sen, G. Karlsson, and J. D. Robbins, "Performance models of statistical multiplexing in packet video communications," *IEEE Transactions on Communications*, vol. 36, pp. 834–843, 1988.

[71] T. S. Rappaport, *Wireless Communications: Principles and Practice, 2nd ed.*, Upper Saddle River, NJ: Prentice Hall PTR, 2001.

[72] S. Verdu, "Spectral efficiency in the wideband regime," *IEEE Transactions on Information Theory*, vol. 48, no. 6, pp. 1319–1343, June 2002.

[73] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*, Cambridge: Cambridge University Press, U.K., 2005.

[74] I.S. Gradshteyn and I.M. Ryzhik, *Table of Integral, Series, and Products*, New York: Academic Press, 1992.

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

[76] A. J. Viterbi, *CDMA : Principles of Spread Spectrum Communication*, Upper Saddle River, NJ: Prentice Hall PTR, 1995.

[77] S. Verdu, *Multiuser Detection*, Cambridge: Cambridge University Press, U.K., 1998.

[78] L. Hanzo, M. Münster, B. J. Choi, and T. Keller, *OFDM and MC-CDMA for Broadband Multi-User Communications, WLANs and Broadcasting*, New York: Wiley, 2003.

[79] G. J. Foschini, "Layered space-time architecture for wireless communications in a fading environment when using multielement antennas," *Bell Labs Technical Journal*, vol. 1, no. 6, pp. 41–59, Autumn 1996.

[80] J. Jiang, R. M. Buehrer, and W. H. Tranter, "Antenna diversity in multiuser data networks," *IEEE Transactions on Communications*, vol. 52, no. 3, pp. 490–497, March 2004.

[81] A. J. Goldsmith, *Wireless Communications*, Cambridge: Cambridge University Press, U.K., 2005.

[82] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge: Cambridge University Press, U.K., 2004.

[83] L. Ozarow, S. Shamai, and A. D. Wyner, "Infomration theoretic consideration for cellular mobile radio," *IEEE Transactions on Vehicle Technology*, vol. 43, pp. 359–378, May 1994.

[84] T. E. Klein and R. G. Gallager, "Power control for the additive white Gaussian noise channel under channel estimation errors," in *Proc. IEEE ISIT*, Washington, DC, 2001, p. 304.

[85] S. Ohno and G. B. Giannakis, "Capacity maximizing MMSE-optimal pilots for wireless OFDM over frequency-selective block Rayleigh-fading channels," *IEEE Transactions on Information Theory*, vol. 50, no. 9, pp. 2138–2145, Sep. 2004.

[86] C. Y. Wong, R. S. Cheng, K. B. Letaief, and R. D. Murch, "Multicarrier OFDM with adaptive subcarrier, bit, and power allocation," *IEEE Journal on Selected Areas in Communications*, vol. 17, no. 10, pp. 1747–1758, Oct. 1999.

[87] A. Lapidoth and S. Shamai, "Fading channels: How perfect need 'perfect side information' be," *IEEE Transactions on Information Theory*, vol. 48, pp. 1118–1134, May 2002.

[88] S. Zhou and G. B. Giannakis, "Adaptive modulation for multi-antenna transmissions with channel mean feedback," *IEEE Transactions on Wireless Communications*, vol. 3, no. 5, pp. 1626–1636, Sep. 2004.

[89] 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, no. 5, pp. 616–628, Oct. 1999.

[90] X. Zhang and J. Tang, "QoS-driven asynchronous uplink subchannel allocation algorithms for space-time OFDM-CDMA systems in wireless networks," *ACM/Kluwer Journal of Wireless Networks (WINET)*, vol. 8, pp. 411–425, Aug. 2006.

[91] S. Shakkottai, T. S. Rappaport, and P. C. Karlsson, "Cross-layer design for wireless networks," *IEEE Communications Magazine*, vol. 41, no. 10, pp. 74–80, Oct. 2003.

[92] S. Falahati, A. Svensson, T. Ekman, and M. Sternad, "Adaptive modulation systems for predicted wireless channels," *IEEE Transactions on Communications*, vol. 52, no. 2, pp. 307–316, Feb. 2004.

[93] J. Razavilar, K. J. R. Liu, and S. I. Marcus, "Jointly optimized bit-rate/delay control policy for wireless packet networks with fading channels," *IEEE Transactions on Communications*, vol. 50, no. 3, pp. 484–494, Mar. 2002.

[94] R. Knopp and P. A. Humblet, "Information capacity and power control in single-cell multiuser communications," in *Proc. IEEE ICC*, Seattle, WA, June 1995, pp. 331–335.

[95] P. Viswanath, D. N. C. Tse, and R. Laroia, "Opportunistic beamforming using

dumb antennas," *IEEE Transactions on Information Theory*, vol. 48, no. 6, pp. 1277–1294, 2002.

[96] 3GPP, "UE Radio Transmission and Reception (FDD, release 5)," TS 25.101, v5.3.0., June 2002.

[97] V. Kawadia and P. R. Kumar, "A cautionary perspective on cross-layer design," *IEEE Wireless Communications*, pp. 3–11, Feb. 2005.

[98] E. Setton, T. Yoo, X. Zhu, A. Goldsmith, and B. Girod, "Cross-layer design of ad hoc networks for real-time video streaming," *IEEE Wireless Communications*, vol. 12, no. 4, pp. 59–65, Aug. 2005.

[99] T. M. Cover and A. El Gamal, "Capacity theorem for the relay channel," *IEEE Transactions on Information Theory*, vol. 25, pp. 572–584, Sep. 1979.

[100] A. Sendonaris, E. Erkip, and B. Aazhang, "User cooperation diversity – Part I: System description," *IEEE Transactions on Communications*, vol. 51, no. 11, pp. 1927–1938, November 2003.

[101] A. Sendonaris, E. Erkip, and B. Aazhang, "User cooperation diversity – Part II: Implementation aspects and performance analysis," *IEEE Transactions on Communications*, vol. 51, no. 11, pp. 1939–1948, November 2003.

[102] N. Ahmed, M. A. Khojastepour, and B. Aazhang, "Outage minimization and optimal power control for the fading relay channel," in *Proc. IEEE ITW*, San Antonio, TX, Oct. 2004, pp. 458–462.

[103] N. Ahmed, M. A. Khojastepour, A. Sabharwal, and B. Aazhang, "Ourage minimization with limited feedback for the fading relay channel," *IEEE Transactions on Communications*, vol. 54, no. 3, pp. 659–669, Mar. 2006.

[104] J. N. Laneman, D. N. C. Tse, and G. W. Wornell, "Cooperative diversity in wireless networks: Efficient protocols and outage behavior," *IEEE Transactions on Information Theory*, vol. 50, no. 12, pp. 3062–3080, December 2004.

[105] A. Host-Madsen and J. Zhang, "Capacity bounds and power allocation for wireless relay channels," *IEEE Transactions on Information Theory*, vol. 51, no. 6, pp. 2020–2040, June 2005.

[106] M. O. Hasna and M.-S. Alouini, "Performance analysis of two-hop relayed transmissions over Rayleigh fading channels," in *Proc. IEEE VTC*, Vancouver, Canada, Sep. 2002, pp. 1992–1996.

[107] M. O. Hasna and M.-S. Alouini, "Optimal power allocation for relayed transmissions over Rayleigh-fading channels," *IEEE Transactions on Wireless Communications*, vol. 3, no. 6, pp. 1999–2004, November 2004.

[108] S. Cui and A. Goldsmith, "Cross-layer design in energy-constrained networks using cooperative MIMO techniques," *EURASIP Journal of Applied Signal Processing*, vol. 86, pp. 1804–1814, Aug. 2006.

[109] Wolfram Research Inc., "Property of the exponential integral function $E_v(z)$," 2001, avaible at http://functions.wolfram.com/06.34.06.0004.01.

[110] L.-L. Yang and L. Hanzo, "Multicarrier DS-CDMA: a multiple access scheme for ubiquitous broadband wireless communications," *IEEE Communications Magazine*, pp. 116–124, Oct. 2003.

[111] L.-L. Yang and L. Hanzo, "Performance of generalized Multicarrier DS-CDMA over Nakagami-m fading channels," *IEEE Transactions on Communications*, vol. 50, no. 6, pp. 956–966, Jun. 2002.

[112] X.-K. Zhao and X.-D. Zhang, "Peak-to-average power ratio analysis in Multi-carrier DS-CDMA," *IEEE Transactions on Vehicle Technology*, vol. 52, no. 3, pp. 561–568, May 2003.

[113] E. A. Sourour and M. Nakagawa, "Performance of orthogonal multicarrier CDMA in a multipath fading channel," *IEEE Transactions on Communications*, vol. 44, no. 3, pp. 356–367, Mar. 1996.

[114] X. Zhang, J. Tang, and H.-H. Chen, "Space-time diversity-enhanced QoS provisioning for real-time service over MC-DS-CDMA based wireless networks," *John-Wiley Journal of Wireless Communications and Mobile Computing (WCMC)*, accepted.

[115] M. S. Alouini and A. J. Goldsmith, "A unified approach for calculating error rate of linearly modulated signals over generalized fading channels," *IEEE Transactions on Communications*, vol. 47, no. 9, pp. 1324–1334, Sep. 1999.

[116] M. K. Simon and M.S. Alouini, "A unified approach to the performance analysis of digtal communications over generalized fading channels," *Proceedings of the IEEE*, vol. 86, no. 9, pp. 1860–1877, Sep. 1998.

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

[118] M. Abramowitz, *Handbook of Mathematical Functions*, New York: Dover, 1965.

[119] N. Dinur and D. Wulich, "Peak-to-average power ratio in high-order OFDM," *IEEE Transactions on Communications*, vol. 49, no. 6, pp. 1063–1072, Jun.

2001.

[120] Z. Chen, J. Yuan, B. Vucetic, and Z. Zhou, "Performance of Alamouti scheme with transmit antenna selection," *IEE Electronics Letters*, vol. 39, no. 23, pp. 1666–1668, Nov. 2003.

[121] E. N. Onggosanusi, A. Gatherer, A. G. Dabak, and S. Hosur, "Performance analysis of closed-loop transmit diversity in the presence of feedback delay," *IEEE Transactions on Communications*, vol. 49, no. 9, pp. 1618 – 1630, Sep. 2001.

[122] N. Jindal and A. Goldsmith, "Capacity and optimal power allocation for fading broadcast channels with minimum rates," *IEEE Transactions on Information Theory*, vol. 49, no. 11, pp. 2895–2909, Nov. 2003.

[123] S. Vishwanath, N. Jindal, and A. Goldsmith, "Duality, achievable rates, and sum-rate capacity of Gaussian MIMO broadcast channels," *IEEE Transactions on Information Theory*, vol. 49, no. 10, pp. 2658–2668, Oct. 2003.

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

[125] N. Jindal and A. Goldsmith, "Dirty-paper coding versus TDMA for MIMO Broadcast channels," *IEEE Transactions on Information Theory*, vol. 51, no. 5, pp. 1783–1794, May 2005.

[126] N. Jindal, W. Rhee, S. Vishwanath, S. A. Jafar, and A. Goldsmith, "Sum power iterative water-filling for multi-antenna Gaussian broadcast channels," *IEEE Transactions on Information Theory*, vol. 51, no. 4, pp. 1570–1580, Apr. 2005.

[127] L. Li, N. Jindal, and A. Goldsmith, "Outage capacities and optimal power allocation for fading multiple-access channels," *IEEE Transactions on Information Theory*, vol. 51, no. 4, pp. 1326–1347, Apr. 2005.

[128] L. Liu, P. Parag, J. Tang, W.-Y. Chen, and J.-F. Chamberland, "Resource allocation and quality of service evaluation for wireless communication systems using fluid models," *IEEE Transactions on Information Theory*, submitted.

[129] J. R. Norris, *Markov Chains*, Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge: Cambridge University Press, U.K., 1998.

[130] R. A. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge: Cambridge University Press, U.K., 1987.

[131] Wolfram Research Inc., "Property of the hypergeometric function ${}_1F_1(n; m; z)$," 2001, avaible at http://functions.wolfram.com/07.20.03.0025.01.

[132] Wolfram Research Inc., "Property of the hypergeometric function ${}_2F_1(a, b; b - n; z)$," 2001, avaible at http://functions.wolfram.com/07.23.03.0082.01.

APPENDIX A

DERIVATIONS OF EQ. (2.9)

Let α denote the sampled envelope of the channel gain at combiner output. From [67, eq. (3)], the LCR at α, denoted by $N_\alpha(\alpha)$, is determined by

$$N_\alpha(\alpha) = p_\alpha(\alpha) \int_0^\infty \dot{\alpha} p_{\dot{\alpha}|\alpha}(\dot{\alpha}|\alpha) d\dot{\alpha} \tag{A.1}$$

where $\dot{\alpha}$ denotes the time derivative of the envelope α, and $p_\alpha(\alpha)$ and $p_{\dot{\alpha}|\alpha}(\dot{\alpha}|\alpha)$ denote the corresponding pdf and conditional pdf, respectively. It can be shown that the following relation holds

$$N_\Gamma(\gamma) = N_\alpha \left(\sqrt{\frac{\Omega\gamma}{\overline{\gamma}}} \right) \tag{A.2}$$

where $\Omega = E\{\alpha^2\}$ denotes the average gain of the envelope process. Following the similar approach of deriving the unified pdf as in Eq. (2.1), the envelope α of the combined SNR for a variety of MIMO diversity scheme can be expressed as

$$\alpha = \max_{1 \le i \le M} \left\{ \alpha_{(i)} \right\} \tag{A.3}$$

where we define

$$\alpha_{(i)} \triangleq \left[\frac{1}{\beta} \sum_{j=1}^{L} \alpha_{ij}^2 \right]^{1/2} \tag{A.4}$$

where $\{\alpha_{ij} : \forall 1 \le i \le M, 1 \le j \le L\}$ are i.i.d. Nakagami-m random variables (r.v.'s), and the parameters M, L, and β are the same as those defined in TABLE I. Then, the pdf and cumulative density function (CDF) of $\alpha_{(i)}$, denoted by $p_{\alpha_{(i)}}(\alpha)$ and $P_{\alpha_{(i)}}(\alpha)$, respectively, are given by

$$p_{\alpha_{(i)}}(\alpha) = 2\left(\frac{\beta m}{\Omega}\right)^{mL} \frac{\alpha^{2mL-1}}{\Gamma(mL)} \exp\left(-\frac{\beta m \alpha^2}{\Omega}\right) \tag{A.5}$$

and

$$P_{\alpha_{(i)}}(\alpha) = \frac{\gamma\left(mL, \frac{\beta m}{\Omega}\alpha^2\right)}{\Gamma(mL)}. \tag{A.6}$$

Using the order statistics [50], the pdf $p_\alpha(\alpha)$ of α can be derived as

$$\begin{aligned} p_\alpha(\alpha) &= M p_{\alpha_{(i)}}(\alpha)\left[P_{\alpha_{(i)}}(\alpha)\right]^{M-1} \\ &= \frac{2M}{\Gamma(mL)} \sum_{i=0}^{M-1} (-1)^i \binom{M-1}{i} \\ &\quad \cdot \exp\left(-(i+1)\frac{\beta m \alpha^2}{\Omega}\right) \sum_{j=0}^{i(mL-1)} \xi_{ji} \left(\frac{\beta m}{\Omega}\right)^{j+mL} \alpha^{2(j+mL)-1}. \end{aligned} \tag{A.7}$$

On the other hand, taking the derivative of Eq. (A.4) we get

$$\dot{\alpha}_{(i)} = \frac{1}{\beta \alpha_{(i)}} \sum_{j=1}^{L} \alpha_{ij} \dot{\alpha}_{ij}. \tag{A.8}$$

According to [67, Section II-A], conditioned on $\alpha_{(i)}$, $\dot{\alpha}_{(i)}$ is a Gaussian r.v. with zero-mean and variance σ^2 equal to

$$\sigma^2 = \frac{\Omega \pi^2 f_d^2}{\beta m}. \tag{A.9}$$

Thus, using Eq. (A.1), the LCR of the *combining diversity* scheme can be derived as

$$N_{\alpha_{(i)}}(\alpha) = p_{\alpha_{(i)}}(\alpha) \int_0^\infty \frac{\dot{\alpha}}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\dot{\alpha}^2}{2\sigma^2}\right) d\dot{\alpha} = \frac{\sigma}{\sqrt{2\pi}} p_{\alpha_{(i)}}(\alpha). \tag{A.10}$$

Taking the *selection diversity* into account, the LCR can be simply obtained by replacing $p_{\alpha_{(i)}}(\alpha)$ in Eq. (A.10) to $p_\alpha(\alpha)$ in Eq. (A.7), i.e.,

$$N_\alpha(\alpha) = \frac{\sigma}{\sqrt{2\pi}} p_\alpha(\alpha). \tag{A.11}$$

Finally, substituting Eqs. (A.7) and (A.9) into Eq. (A.11), and then using the relation of Eq. (A.2), we obtain the LCR for unified MIMO diversity scheme as shown in Eq. (2.9).

APPENDIX B

PROOF OF PROPOSITION 1

Proof. The proof is similar in spirit to [1, Example 3.3]. Let us define $\nu_i(\theta, t) \triangleq \mathbb{E}\left\{e^{-\theta S(t)} \,\middle|\, R(1) = \mu_i\right\}$ and $\boldsymbol{\nu}(\theta, t) \triangleq \left\{\nu_1(\theta, t), \cdots, \nu_K(\theta, t)\right\}$, respectively. Then, we get

$$\begin{aligned}
\nu_i(\theta, t) &= \left\{e^{-\theta R(1)} \middle|\, R(1) = \mu_i\right\} \mathbb{E}\left\{e^{-\theta(S(t)-R(1))} \,\middle|\, R(1) = \mu_i\right\} \\
&\overset{(a)}{=} e^{-\mu_i\theta} \sum_{j=1}^{K} \mathbb{E}\left\{e^{-\theta(S(t)-S(1))} \,\middle|\, R(2) = \mu_j, R(1) = \mu_i\right\} \\
&\quad \cdot \Pr\left\{R(2) = \mu_j \,\middle|\, R(1) = \mu_i\right\} \\
&\overset{(b)}{=} e^{-\mu_i\theta} \sum_{j=1}^{K} \mathbb{E}\left\{e^{-\theta(S(t)-S(1))} \,\middle|\, R(2) = \mu_j\right\} p_{ij} \\
&\overset{(c)}{=} e^{-\mu_i\theta} \sum_{j=1}^{K} \mathbb{E}\left\{e^{-\theta S(t-1)} \,\middle|\, R(1) = \mu_j\right\} p_{ij} \\
&= e^{-\mu_i\theta} \sum_{j=1}^{K} \nu_j(\theta, t-1) p_{ij} \qquad \text{(B.1)}
\end{aligned}$$

where (a) is due to the chain rule of conditional probability and the fact that $R(1) = S(1)$ from the definition of $S(t)$, (b) is because of the Markovian property that given current state $R(2)$, the future state is independent of the past state $R(1)$, and (c) is due to the strong Markov property [129, Theorem 1.4.2]. Rewriting Eq. (B.1) as a matrix form and using the iterative relationship as expressed in Eq. (B.1), we get

$$\boldsymbol{\nu}(\theta, t)^T = \left(\boldsymbol{\Phi}(\theta)\mathbf{P}\right)^{t-1} \boldsymbol{\Phi}(\theta)\mathbf{1}^T \qquad \text{(B.2)}$$

where we use the relation $\boldsymbol{\nu}(\theta, 1)^T = \boldsymbol{\Phi}(\theta)\mathbf{1}^T$ with $\mathbf{1}$ denoting the K-dimensional row-vector of $\mathbf{1} = [1, ..., 1]$. Using Eq. (B.2), the moment generating function $\mathbb{E}\left\{e^{-\theta S(t)}\right\}$

can be expressed as

$$\mathbb{E}\left\{e^{-\theta S(t)}\right\} = \boldsymbol{\pi}\boldsymbol{\nu}(\theta,t)^T = \boldsymbol{\pi}\left(\boldsymbol{\Phi}(\theta)\mathbf{P}\right)^{t-1}\boldsymbol{\Phi}(\theta)\mathbf{1}^T = \boldsymbol{\pi}\left(\mathbf{P}\boldsymbol{\Phi}(\theta)\right)^t\mathbf{1}^T. \qquad \text{(B.3)}$$

Since $\mathbf{P}\boldsymbol{\Phi}(\theta)$ is a primitive nonnegative matrix, from Perron-Frobenious Theorem [130, Theorem 8.5.1], we have

$$\lim_{t\to\infty}\bigg(\boldsymbol{\pi}\left(\mathbf{P}\boldsymbol{\Phi}(\theta)\right)^t\mathbf{1}^T\bigg) = \bigg(\rho\{\mathbf{P}\boldsymbol{\Phi}(\theta)\}\bigg)^t\boldsymbol{\pi}\mathbf{y}(\theta)\mathbf{x}(\theta)\mathbf{1}^T \qquad \text{(B.4)}$$

where $\mathbf{y}(\theta)$ and $\mathbf{x}(\theta)$ are, respectively, the column and row eigenvectors of the matrix $\mathbf{P}\boldsymbol{\Phi}(\theta)$, corresponding to the maximum real-valued eigenvalue $\rho\{\mathbf{P}\boldsymbol{\Phi}(\theta)\}$ and satisfying $\mathbf{x}(\theta)\mathbf{y}(\theta) = 1$. Thus, we obtain the effective bandwidth function $\mathrm{E_C}(\theta)$ as follows:

$$\mathrm{E_C}(\theta) = -\lim_{t\to\infty}\frac{1}{\theta t}\log\left(\mathbb{E}\left\{e^{-\theta S(t)}\right\}\right) = -\frac{1}{\theta}\log\bigg(\rho\Big\{\mathbf{P}\,\boldsymbol{\Phi}(\theta)\Big\}\bigg). \qquad \text{(B.5)}$$

Thus, the proof follows. □

APPENDIX C

PROOF OF PROPOSITION 2

Proof. Since the Gärtner-Ellis limit of the service process $\Lambda_{\mathrm{C}}(\theta)$ is a convex function [1], $\Lambda_{\mathrm{C}}(-\theta)$ is also a convex function. Then, $-\Lambda_{\mathrm{C}}(-\theta)$ is a concave function. Let us redefine $f(\theta) \triangleq -\Lambda_{\mathrm{C}}(-\theta)$ for convenience. Due to the concavity for $f(\theta)$, we have $f''(\theta) \leq 0$. Then,

$$\mathrm{E_C}'(\theta) = \left(\frac{f(\theta)}{\theta}\right)' = \frac{\theta f'(\theta) - f(\theta)}{\theta^2}. \tag{C.1}$$

It is easy to see that $f(0) = 0$. Thus,

$$(\theta f'(\theta) - f(\theta))\big|_{\theta=0} = 0. \tag{C.2}$$

Moreover, for all $\theta > 0$,

$$(\theta f'(\theta) - f(\theta))' = f'(\theta) + \theta f''(\theta) - f'(\theta) = \theta f''(\theta) \leq 0. \tag{C.3}$$

From Eqs. (C.2) and (C.3), we know for all $\theta > 0$, the derivative $\mathrm{E_C}'(\theta) \leq 0$. Therefore, the effective capacity function $\mathrm{E_C}(\theta)$ is a monotonically decreasing function of θ. Thus, the proof of Claim 1 follows.

Let $\lambda(\theta) \triangleq \rho\{\mathbf{P\Phi}(\theta)\}$ and $\mathbf{x}(\theta)$ denote the maximum real-valued eigenvalue and the corresponding eigenvector of the matrix $\mathbf{P\Phi}(\theta)$. Thus, the following equation holds

$$\lambda(\theta)\mathbf{x}(\theta) = \mathbf{x}(\theta)\mathbf{P\Phi}(\theta). \tag{C.4}$$

Since the Gärtner-Ellis limit $\Lambda_{\mathrm{C}}(\theta)$ is differentiable, $\mathbf{P\Phi}(\theta)$ is also differentiable. Thus, the eigenvalue $\lambda(\theta)$ and eigenvector $\mathbf{x}(\theta)$ can be expanded as Taylor series in the first order at $\theta = 0$, i.e.,

$$\begin{cases} \lambda(\theta) = \lambda_0 + \lambda_1\theta + o(\theta) \\ \mathbf{x}(\theta) = \mathbf{x}_0 + \mathbf{x}_1\theta + o(\theta). \end{cases} \tag{C.5}$$

Since 1 and $\boldsymbol{\pi}$ are, respectively, the maximum real-valued eigenvalue and corresponding positive row-eigenvector of transition matrix $\mathbf{P}$, we obtain $\lambda_0 = 1$ and $\mathbf{x}_0 = \boldsymbol{\pi}$ at $\theta = 0$. Then, the left-handside of Eq. (C.4) can be expressed as

$$\lambda(\theta)\mathbf{x}(\theta) = \lambda_0\mathbf{x}_0 + (\lambda_0\mathbf{x}_1 + \lambda_1\mathbf{x}_0)\theta + o(\theta) = \boldsymbol{\pi} + (\mathbf{x}_1 + \lambda_1\boldsymbol{\pi})\theta + o(\theta). \tag{C.6}$$

Let $\mathbf{U} \triangleq \text{diag}\{\mu_1, \mu_2, ..., \mu_K\}$. Similarly, the right-handside of Eq. (C.4) can be written as follows:

$$\begin{aligned} \mathbf{x}(\theta)\mathbf{P}\boldsymbol{\Phi}(\theta) &= (\boldsymbol{\pi} + \mathbf{x}_1\theta)\mathbf{P}(\mathbf{I} - \theta\mathbf{U}) + o(\theta) = \boldsymbol{\pi}\mathbf{P} + (\mathbf{x}_1\mathbf{P} - \boldsymbol{\pi}\mathbf{P}\mathbf{U})\theta + o(\theta) \\ &= \boldsymbol{\pi} + (\mathbf{x}_1\mathbf{P} - \boldsymbol{\pi}\mathbf{U})\theta + o(\theta). \end{aligned} \tag{C.7}$$

Comparing Eq. (C.6) and Eq. (C.7), we have

$$\mathbf{x}_1 + \lambda_1\boldsymbol{\pi} = \mathbf{x}_1\mathbf{P} - \boldsymbol{\pi}\mathbf{U}. \tag{C.8}$$

Solving Eq. (C.8) for λ_1, we obtain

$$\lambda_1 = -\sum_{k=1}^{K} \pi_k\mu_k = -\overline{\mu}. \tag{C.9}$$

Thus, by using the definition $\lambda(\theta) \triangleq \rho\{\mathbf{P}\boldsymbol{\Phi}(\theta)\}$ and Eqs. (2.19), (C.5), and (C.9), we get

$$\lim_{\theta\to 0} \mathrm{E_C}(\theta) = -\lim_{\theta\to 0} \frac{1}{\theta} \log \lambda(\theta) = -\lambda_1 \lim_{\theta\to 0} \frac{1}{\lambda_1\theta} \log\Big(1 + \lambda_1\theta + o(\theta)\Big) = -\lambda_1 = \overline{\mu}. \tag{C.10}$$

Noting that $\mathrm{E_C}(\theta)$ is a monotonically decreasing function, the proof of <u>Claim 2</u> follows.

Letting $j = \arg\min_{1\leq i\leq K} \mu_i$, it is clear that

$$\lim_{\theta\to\infty} \lambda(\theta) = p_{jj} \exp\left(-\theta\mu_{\min}\right). \tag{C.11}$$

Thus, we get:

$$\lim_{\theta\to\infty} \mathrm{E_C}(\theta) = -\lim_{\theta\to\infty} \frac{1}{\theta} \log \lambda(\theta) = \mu_{\min}. \tag{C.12}$$

Also noting that $\mathrm{E_C}(\theta)$ is a monotonically decreasing function, the proof of Claim 3 follows. □

APPENDIX D

PROOF OF THEOREM 1

Proof. Since $\log(\cdot)$ is a monotonically increasing function, for each given $\theta > 0$, the maximization problem of (3.4) can be converted into a minimization problem as follows:

$$\min_{\mu(\theta,\gamma):\int_0^\infty \mu(\theta,\gamma)p_\Gamma(\gamma)d\gamma=1} \left\{ \int_0^\infty e^{-\theta T_f B \log_2\left(1+\mu(\theta,\gamma)\gamma\right)} p_\Gamma(\gamma)d\gamma \right\}. \tag{D.1}$$

It is clear from [82, Sec. 3.2] that in (D.1), the objective function is strictly convex and the constraint is linear with respect to $\mu(\theta,\gamma)$. Thus, the minimization problem has a unique optimal solution. Then, we can form the Lagrangian function, denoted by $\mathcal{J}$, as follows:

$$\mathcal{J} = \int_0^\infty e^{-\beta \log\left(1+\mu(\theta,\gamma)\gamma\right)} p_\Gamma(\gamma)d\gamma + \lambda \left(\int_0^\infty \mu(\theta,\gamma)p_\Gamma(\gamma)d\gamma - 1 \right). \tag{D.2}$$

where $\beta \triangleq \theta T_f B/\log 2$ is defined as the normalized QoS exponent. Differentiating the Lagrangian function given by (D.2) and setting the derivative equal to zero [81, Sec. 4.2.4], we get

$$\frac{\partial \mathcal{J}}{\partial \mu(\theta,\gamma)} = \left\{ \lambda - \beta\gamma \left[1+\mu(\theta,\gamma)\gamma\right]^{-\beta-1} \right\} p_\Gamma(\gamma) = 0. \tag{D.3}$$

Defining $\gamma_0 \triangleq \lambda/\beta$ and solving (D.3), we can obtain the optimal power and rate adaptation policy as shown by (3.5), where γ_0 is determined by the mean power constraint of (3.6). The proof follows. □

APPENDIX E

PROOF OF PROPOSITION 4

Proof. In order to prove Proposition 4, we first introduce the following lemma.

Lemma 4. *Let a channel service process be modeled as a continuous-time FSMC with N states, the service rate of the nth state be denoted by $\widetilde{\nu}_n$, ($n \in \{0, 1, ..., N-1\}$), and the corresponding generating matrix of the continuous-time FSMC be represented by $\mathbf{Q}$, respectively. If we define $\mathbf{R} \triangleq \mathrm{diag}\{\widetilde{\nu}_0, \widetilde{\nu}_1, ..., \widetilde{\nu}_{N-1}\}$, then the effective capacity of this process, denoted by $\widetilde{\mathrm{E}}_{\mathrm{C}}(\theta)$, is determined by*[1]

$$\widetilde{\mathrm{E}}_{\mathrm{C}}(\theta) = -\frac{1}{\theta}\delta\left\{\mathbf{Q} - \theta\mathbf{R}\right\} \tag{E.1}$$

where $\delta\{\cdot\}$ denotes the maximum real eigenvalue of the matrix.

Proof. The proof is similar to [12, Appendix], which is omitted for lack of space. □

There exists the close relationship between continuous-time FSMC and discrete-time FSMC. Under appropriate conditions, the discrete-time FSMC can be considered as the "samples" of the embedded continuous-time FSMC. Based on our system model, the sample interval is T_f and the service rate $\nu_n = \widetilde{\nu}_n T_f$.

The relationship between transition probability matrix $\mathbf{P}$ of a discrete-time FSMC and generating matrix $\mathbf{Q}$ of a continuous-time FSMC can be expressed as

$$\mathbf{P}(T_f) = e^{\mathbf{Q}T_f} = \mathbf{I} + \mathbf{Q}T_f + o(T_f^2) \tag{E.2}$$

[1] *Note that for continuous-time FSMC, the unit for the service rate $\widetilde{\nu}_n$ and the effective capacity $\widetilde{\mathrm{E}}_{\mathrm{C}}(\theta)$ is "bits per second".*

where we rewrite $\mathbf{P}$ by $\mathbf{P}(T_f)$ in (E.2) to emphasize that the sample interval is T_f. Given the transition probability matrix $\mathbf{P}$, the first-order approximation of the generator matrix $\mathbf{Q}$ is determined by

$$\mathbf{Q} \approx \frac{\mathbf{P}(T_f) - \mathbf{I}}{T_f}. \tag{E.3}$$

It is clear that the generating matrix $\mathbf{Q}$ can be expressed as $\mathbf{Q} = f_d \mathbf{A}$, where $\mathbf{A}$ only depends on the marginal statistics of the channel. Thus, we can approximate the effective capacity of a discrete-time FSMC by the effective capacity of a continuous-time FSMC as follows:

$$\mathrm{E_C}(\theta) \approx \widetilde{\mathrm{E}}_\mathrm{C}(\theta) T_f. \tag{E.4}$$

Based on the continuous-time FSMC approximation given by (E.4), we prove Proposition 4 as follows. From Lemma 4, we have

$$\begin{aligned}
\widetilde{\mathrm{E}}_{\mathrm{C}_1}(\theta) &= -\frac{1}{\theta}\delta\left\{\mathbf{Q}_1 - \theta\mathbf{R}\right\} \\
&= -\frac{1}{\theta}\delta\left\{f_{d_1}\mathbf{A} - \theta\mathbf{R}\right\} \\
&= -\frac{1}{\theta}\delta\left\{\frac{f_{d_1}}{f_{d_2}}\left(f_{d_2}\mathbf{A}\right) - \frac{f_{d_1}}{f_{d_2}}\left(\frac{f_{d_2}}{f_{d_1}}\theta\mathbf{R}\right)\right\} \\
&= -\left(\frac{f_{d_1}}{f_{d_2}}\right)\frac{1}{\theta}\delta\left\{f_{d_2}\mathbf{A} - \frac{f_{d_2}}{f_{d_1}}\theta\mathbf{R}\right\} \\
&= -\frac{1}{\left(\frac{f_{d_2}}{f_{d_1}}\theta\right)}\delta\left\{\mathbf{Q}_2 - \left(\frac{f_{d_2}}{f_{d_1}}\theta\right)\mathbf{R}\right\} \\
&= \widetilde{\mathrm{E}}_{\mathrm{C}_2}\left(\frac{f_{d_2}}{f_{d_1}}\theta\right).
\end{aligned} \tag{E.5}$$

Plugging the approximate relationship given by (E.4) into (E.5), the proof for Proposition 4 follows. □

APPENDIX F

PROOF OF THE STRICT CONVEXITY OF THE OBJECTIVE FUNCTION IN EQ. (4.18)

Proof. To show the strict convexity of the objective function in (4.18), we introduce the following proposition:

Proposition 12. ***If*** $\mathbf{x} = (x_1, x_2, ..., x_n)$ *and* $f(\mathbf{x}) = \prod_{i=1}^{n} x_i^{-\alpha}$*, where* $x_i > 0$ *for all* $i = 1, 2, ..., n$ *and* $\alpha > 0$*,* ***then*** $f(\mathbf{x})$ *is strictly convex on the domain where* $\mathbf{x} = (x_1, x_2, ..., x_n)$ *is defined.*

Proof. It is easy to show that

$$\frac{\partial^2 f(\mathbf{x})}{\partial x_k^2} = \frac{\alpha(\alpha+1)}{x_k^2} \prod_{i=1}^{n} x_i^{-\alpha} \tag{F.1}$$

and

$$\frac{\partial^2 f(\mathbf{x})}{\partial x_k \partial x_l} = \frac{\alpha^2}{x_k x_l} \prod_{i=1}^{n} x_i^{-\alpha}, \text{ for } k \neq l. \tag{F.2}$$

Thus, the Hessian of $f(\mathbf{x})$ can be expressed as

$$\nabla^2 f(\mathbf{x}) = \alpha \prod_{i=1}^{n} x_i^{-\alpha} \left[\alpha \mathbf{y}^T \mathbf{y} + \text{diag}\left(\frac{1}{x_1^2}, ..., \frac{1}{x_n^2} \right) \right] \tag{F.3}$$

where $\mathbf{y} = (1/x_1, 1/x_2, ..., 1/x_n)$. For any nonzero $\mathbf{v} = (v_1, v_2, ..., v_n)$, we have

$$\begin{aligned} &\mathbf{v}\left(\nabla^2 f(\mathbf{x})\right)\mathbf{v}^T \\ &\quad = \alpha \prod_{i=1}^{n} x_i^{-\alpha} \left[\alpha \left(\mathbf{v}\mathbf{y}^T \right)^2 + \sum_{i=1}^{n} \left(\frac{v_i}{x_i} \right)^2 \right] > 0. \end{aligned} \tag{F.4}$$

The Hessian of $f(\mathbf{x})$ is positive definite and therefore $f(\mathbf{x})$ is strictly convex on the domain where $\mathbf{x}$ is defined. □

By Proposition 12, since the item $[1+\mu_n(\theta,\boldsymbol{\gamma})\gamma_n] > 0$ always holds, we have $\prod_{n=1}^{N}[1+\mu_n(\theta,\boldsymbol{\gamma})\gamma_n]^{-\frac{\beta}{N}}$ is strictly convex on the space spanned by the vector expressed as $\big([1+\mu_1(\theta,\boldsymbol{\gamma})\gamma_1], ..., [1+\mu_N(\theta,\boldsymbol{\gamma})\gamma_N]\big)$. Also, since $\mu_n(\theta,\boldsymbol{\gamma})$ is just a *linear variety* of $[1+\mu_n(\theta,\boldsymbol{\gamma})\gamma_n]$, $\big(\mu_1(\theta,\boldsymbol{\gamma}), ..., \mu_N(\theta,\boldsymbol{\gamma})\big)$ preserves the strict convexity of $\prod_{n=1}^{N}[1+\mu_n(\theta,\boldsymbol{\gamma})\gamma_n]^{-\frac{\beta}{N}}$ [82, Sec. 3.2.2]. Thus, $\prod_{n=1}^{N}[1+\mu_n(\theta,\boldsymbol{\gamma})\gamma_n]^{-\frac{\beta}{N}}$ is strictly convex on the space spanned by $\big(\mu_1(\theta,\boldsymbol{\gamma}), ..., \mu_N(\theta,\boldsymbol{\gamma})\big)$. Furthermore, the integral in (4.18) is a linear operation, which also preserves the strict convexity. Thus, the objective function in (4.18) is strictly convex on the space spanned by $\big(\mu_1(\theta,\boldsymbol{\gamma}), ..., \mu_N(\theta,\boldsymbol{\gamma})\big)$. □

APPENDIX G

PROOF OF LEMMA 1

Proof. Without loss of generality, we assume $\mathcal{N}_1 = \{\gamma_1, \gamma_2, ..., \gamma_{N_1}\}$, where $N_1 < N$. Let us denote the complementary set of $\mathcal{N}_1$ by $\overline{\mathcal{N}}_1$, i.e., $\overline{\mathcal{N}}_1 = \{\gamma_{N_1+1}, \gamma_{N_1+2}, ..., \gamma_N\}$. To prove Lemma 1, we need to show that for any nonempty subset $\mathcal{C} \subseteq \overline{\mathcal{N}}_1$, there is no policy $\mu_n(\theta, \boldsymbol{\gamma})$ such that $\mu_n(\theta, \boldsymbol{\gamma}) > 0$ for all $n \in \mathcal{C} \cup \mathcal{N}_1$.

If $\mathcal{C} = \overline{\mathcal{N}}_1$, we already know there is no such a policy, due to the condition of Scenario-2.

Otherwise, if $\mathcal{C} \subset \overline{\mathcal{N}}_1$, without loss of generality, we assume that the set $\mathcal{C} = \{\gamma_{N_1+1}, \gamma_{N_1+2}, ..., \gamma_{N_1+G}\}$, where $1 \leq G \leq N - N_1 - 1$. Suppose there exists such a policy, from Section 2 we know that the policy can be expressed as

$$\mu_n(\theta, \boldsymbol{\gamma}) = \begin{cases} \dfrac{1}{\gamma_0^{\frac{N}{\omega}} \prod_{i=1}^{N_1+G} \gamma_i^{\frac{\beta}{\omega}}} - \dfrac{1}{\gamma_n}, & n \in \mathcal{C} \cup \mathcal{N}_1 \\ 0, & \text{otherwise} \end{cases} \tag{G.1}$$

where $\omega = (N_1 + G)\beta + N$. In particular, we have $\mu_{N_1+G}(\theta, \boldsymbol{\gamma}) > 0$ in (G.1), which is equivalent to the following:

$$\gamma_{N_1+G} > \left(\gamma_0^N \prod_{i=1}^{N_1+G-1} \gamma_i^{\beta} \right)^{\frac{1}{(N_1+G-1)\beta+N}}. \tag{G.2}$$

On the other hand, from the definition of $\mathcal{N}_1$, we know

$$\frac{1}{\gamma_0^{\frac{1}{\beta+1}} \prod_{i=1}^{N} \gamma_i^{\frac{\beta}{N(\beta+1)}}} \leq \frac{1}{\gamma_n} \tag{G.3}$$

where $n \in \{N_1 + 1, N_1 + 2, ..., N\}$. Plugging $n = N_1 + G$ into (G.3), we get

$$\gamma_{N_1+G} \leq \left(\gamma_0^N \prod_{i=1}^{N_1+G-1} \gamma_i^{\beta} \prod_{j=N_1+G+1}^{N} \gamma_j^{\beta} \right)^{\frac{1}{(N-1)\beta+N}}. \tag{G.4}$$

Furthermore, letting $n = (N_1 + G + 1), (N_1 + G + 2), ..., N$ in (G.3), respectively, we obtain a set of $(N - N_1 - G)$ inequalities. Multiplying the left-hand sides and right-hand sides of these $(N - N_1 - G)$ inequalities, respectively, we generate a new inequality as follows:

$$\prod_{j=N_1+G+1}^{N} \gamma_j^{\beta} \leq \left(\gamma_0^N \prod_{i=1}^{N_1+G} \gamma_i^{\beta} \right)^{\frac{\beta(N-N_1-G)}{\omega}} . \tag{G.5}$$

Finally, substituting (G.5) into the right-hand side of (G.4) and re-arranging the expression, we get

$$\gamma_{N_1+G} \leq \left(\gamma_0^N \prod_{i=1}^{N_1+G-1} \gamma_i^{\beta} \right)^{\frac{1}{(N_1+G-1)\beta+N}} \tag{G.6}$$

which contradicts (G.2). Therefore, such a policy does not exist. The proof follows.

□

APPENDIX H

PROOF OF THEOREM 2

Proof. Assume that there are exactly k channels out of M channels being assigned with nonzero power, where $1 \leq k \leq M$. It can be easily shown by contradiction that these k channels are $\widehat{\gamma}_{\pi(1)}, \widehat{\gamma}_{\pi(2)}, ..., \widehat{\gamma}_{\pi(k)}$. Thus, the Lagrangian $\mathcal{J}_1$ can be simplified to a new Lagrangian function, denoted by $\mathcal{J}_1'$, as follows:

$$\mathcal{J}_1' = \mathbb{E}_{\widehat{\boldsymbol{\gamma}}}\left[\prod_{m=1}^{k}\Big(1+\widetilde{\gamma}_{\pi(m)}P_{\pi(m)}(\boldsymbol{\nu})\Big)^{-\beta}\right] + \lambda_1 \sum_{m=1}^{k} P_{\pi(m)}(\boldsymbol{\nu})$$

where $\widetilde{\gamma}_{\pi(m)}$ is defined in Eq. (5.9). Differentiating the simplified Lagrangian $\mathcal{J}_1'$ with respect to $P_{\pi(m)}(\boldsymbol{\nu})$ and setting the derivative equal to zero, we can get a set of k equations:

$$\begin{aligned}\big[1+\widetilde{\gamma}_{\pi(m)}P_{\pi(m)}(\boldsymbol{\nu})\big]^{-(\beta+1)} \prod_{i=1,\, i\neq m}^{k} &\big[1+\widetilde{\gamma}_{\pi(i)}P_{\pi(i)}(\boldsymbol{\nu})\big]^{-\beta} \\ &= \frac{\lambda_1}{\beta\widetilde{\gamma}_{\pi(m)}}, \text{ for all } 1 \leq m \leq k. \end{aligned} \tag{H.1}$$

Solving Eq. (H.1) and considering the boundary conditions, we obtain Eq. (5.10), where

$$\omega(\boldsymbol{\nu}, k) = \left(\frac{\beta}{\lambda_1}\right)^{\frac{1}{k\beta+1}} \prod_{i=1}^{k} \widetilde{\gamma}_{\pi(i)}^{-\frac{\beta}{k\beta+1}}. \tag{H.2}$$

By choosing a proper λ_1 in Eq. (H.2) to meet the total power constraint, $\omega(\boldsymbol{\nu}, k)$ can be simplified to Eq. (5.11). The proof follows. □

APPENDIX I

PROOF OF LEMMA 3

Proof. Due to the linearity of the expectation, it is sufficient to show that the objective function inside the expectation is convex on $P_{\text{total}}(\boldsymbol{\nu}) \in \mathbb{R}_+$. Following the notation of Eq. (5.12), we differentiate $\mathcal{F}\left(\boldsymbol{\mu}^*_{\text{total}}(\boldsymbol{\nu}), \widehat{\boldsymbol{\gamma}}\right)$ with respect to $P_{\text{total}}(\boldsymbol{\nu})$ and get the following:

$$\frac{\partial \mathcal{F}\left(\boldsymbol{\mu}^*_{\text{total}}(\boldsymbol{\nu}), \widehat{\boldsymbol{\gamma}}\right)}{\partial P_{\text{total}}(\boldsymbol{\nu})} = -\frac{\beta (k\Sigma_k)^{k\beta+1}\left[1+\sigma_e^2 P_{\text{total}}(\boldsymbol{\nu})\right]^{k\beta-1}}{\Pi_k^{-k\beta}\left[1+(\sigma_e^2+\Sigma_k)P_{\text{total}}(\boldsymbol{\nu})\right]^{k\beta+1}}. \tag{I.1}$$

In particular, by Theorem 2 we can show, but omit details that, at the critical point where the number of active channels k increases from ℓ to $\ell+1$ with $1 \leq \ell < M$, the total power is equal to

$$P_{\text{total}}(\boldsymbol{\nu}) = \frac{\ell\Sigma_\ell - \widehat{\gamma}_{\pi(\ell+1)}}{\Sigma_\ell \widehat{\gamma}_{\pi(\ell+1)} - \sigma_e^2(\ell\Sigma_\ell - \widehat{\gamma}_{\pi(\ell+1)})} \tag{I.2}$$

Substituting Eq. (I.2) into Eq. (I.1) and letting *either* $k = \ell$ *or* $k = \ell + 1$, the derivative in Eq. (I.1) yields the *same* solution:

$$-\beta \Pi_\ell^{\ell\beta} \widehat{\gamma}_{\pi(\ell+1)}^{\ell\beta-1} \Sigma_\ell^{-2}\left[\Sigma_\ell \widehat{\gamma}_{\pi(\ell+1)} - \sigma_e^2(\ell\Sigma_\ell - \widehat{\gamma}_{\pi(\ell+1)})\right]^2$$

which implies that the derivative of the objective function is *continuous* on $P_{\text{total}}(\boldsymbol{\nu}) \in \mathbb{R}_+$, even though the number of active channels *discretely* increases. Once verified the continuity, the twice differentiation for each given k can be easily obtained as[2]

$$\frac{\partial^2 \mathcal{F}\left(\boldsymbol{\mu}^*_{\text{total}}(\boldsymbol{\nu}), \widehat{\gamma}\right)}{\partial P^2_{\text{total}}(\boldsymbol{\nu})} = \frac{\beta (k\Sigma_k)^{k\beta+1} \left[1+\sigma_e^2 P_{\text{total}}(\boldsymbol{\nu})\right]^{k\beta-2}}{\Pi_k^{-k\beta}\left[1+(\sigma_e^2+\Sigma_k)P_{\text{total}}(\boldsymbol{\nu})\right]^{k\beta+2}} \cdot \Big\{\Sigma_k(k\beta+1)+2\sigma_e^2\big[1+(\sigma_e^2+\Sigma_k)P_{\text{total}}(\boldsymbol{\nu})\big]\Big\} > 0$$

which demonstrates that Eq. (I.1) is a continuous and monotonically increasing function of $P_{\text{total}}(\boldsymbol{\nu})$. Thus, the objective function is strictly convex, and then the proof follows. □

[2] The twice differentiation is not continuous at the critical point when k changes. Moreover, since $\widehat{\boldsymbol{\gamma}} \in \mathbb{R}_+^M$ follows a certain continuous distribution, we have $\Sigma_k > 0$ and $\Pi_k > 0$ with probability 1.

APPENDIX J

PROOF OF THEOREM 3

Proof. Differentiating the Lagrangian function $\mathscr{J}_2$ given by Eq. (5.16) and setting the derivative equal to zero, we get

$$\frac{\partial \mathcal{F}\left(\boldsymbol{\mu}^*_{\text{total}}(\boldsymbol{\nu}), \widehat{\gamma}\right)}{\partial P_{\text{total}}(\boldsymbol{\nu})} + \lambda_2 = 0. \tag{J.1}$$

Plugging Eq. (I.1) into Eq. (J.1) with simple algebraic manipulations, we obtain Eq. (5.17), where

$$\omega^* \triangleq \left(\frac{\beta}{\lambda_2}\right)^{\frac{1}{M\beta+1}}.$$

Similar to the proof of Lemma 3, we can show that the left-hand side of Eq. (5.17) is continuous and monotonically increasing function at $P_{\text{total}}(\boldsymbol{\nu}) \in \mathbb{R}_+$, even though k changes discretely. Therefore, the positive solution $P^*_{\text{total}}(\boldsymbol{\nu}) \in \mathbb{R}_{++}$ and the corresponding number of active channels k in $\{1, ..., M\}$, if exist, are unique, respectively.

Otherwise, if we cannot find such k and $P^*_{\text{total}}(\boldsymbol{\nu})$ that satisfy the two-step power allocation, from the KKT conditions and the constraint $P_{\text{total}}(\boldsymbol{\nu}) \in \mathbb{R}_+$, we know $P^*_{\text{total}}(\boldsymbol{\nu}) = 0$. Finally, the parameter $\omega^* \in \mathbb{R}_+$ should be chosen such that the average power constraint is satisfied. The proof follows. □

APPENDIX K

PROOF OF THEOREM 4

Proof. The proof is based on the following lemma.

Lemma 5. *For all* k *in* $\{1, 2, ..., M\}$*, the following inequality always holds:*

$$k\Sigma_k\Pi_k \leq 1.$$

Proof. It can be shown by definition that $k\Sigma_k$ is the *harmonic mean* of $\{\widehat{\gamma}_{\pi(i)}\}_{i=1}^{k}$, while Π_k^{-1} is the *geometric mean* of $\{\widehat{\gamma}_{\pi(i)}\}_{i=1}^{k}$. Using the well known result that the harmonic mean is always less than or equal to the geometric mean, we know the ratio $k\Sigma_k\Pi_k \leq 1$ always holds, with equality if and only if $\widehat{\gamma}_{\pi(1)} = \widehat{\gamma}_{\pi(2)} = \cdots = \widehat{\gamma}_{\pi(k)}$.[3] The proof of Lemma 5 follows. □

Now, we prove Theorem 4. If $\omega^* \leq 1$, from Lemma 5 we know $\eta_k = (\omega^*)^{M/k} k\Sigma_k\Pi_k \leq 1$. Then, $P_{\text{total}}(\boldsymbol{\nu}) = 0$ always holds, and $\mathbb{E}_{\widehat{\boldsymbol{\gamma}}}[P_{\text{total}}(\boldsymbol{\nu})] = 0$.

Otherwise, if $\omega^* > 1$, the convergence for the average power is equivalent to the convergence of the average water-level $\omega(\boldsymbol{\nu}, k)$ given in Eq. (5.11), since the power assigned to each channel cannot exceed the water-level. Substituting Eq. (5.18) into Eq. (5.11) and removing the irrelevant terms, we get the following condition:

$$\mathbb{E}_{\widehat{\boldsymbol{\gamma}}}\left[\frac{\Sigma_k\Pi_k}{\Sigma_k - \sigma_e^2(\eta_k - 1)}\right]^+ < \infty. \tag{K.1}$$

[3] Since $\Pr\{\widehat{\gamma}_{\pi(1)} = \widehat{\gamma}_{\pi(2)} = \cdots = \widehat{\gamma}_{\pi(k)}\} = 0$ for $k \geq 2$. Therefore, the maximum $k\Sigma_k\Pi_k = 1$ is achieved by $k = 1$ with probability 1.

To prove the necessary condition, we need to find a lower-bound of Eq. (K.1), which is given by

$$\begin{aligned}\mathbb{E}_{\widehat{\gamma}}\left[\frac{\Sigma_k \Pi_k}{\Sigma_k - \sigma_e^2\left(\eta_k - 1\right)}\right]^+ &= \mathbb{E}_{\widehat{\gamma}}\left[\frac{\Pi_k}{1 - \sigma_e^2(\omega^*)^{M/k} k \Pi_k + \sigma_e^2/\Sigma_k}\right]^+ \\ &\overset{(a)}{\geq} \mathbb{E}_{\widehat{\gamma}}\left[\frac{\Pi_1}{1 - \sigma_e^2 \omega^* \Pi_1 + \sigma_e^2/\Sigma_1}\right]^+ \qquad \text{(K.2)}\end{aligned}$$

where (a) holds since Π_k is a monotonically increasing function of k, while $(\omega^*)^{M/k}$ and Σ_k are monotonically decreasing functions of k, respectively. Plugging $\Sigma_1 = \widehat{\gamma}_{\pi(1)}$ and $\Pi_1 = 1/\widehat{\gamma}_{\pi(1)}$ into Eq. (K.2), and upper-bounding it away from infinity, we get the necessary condition given in Eq. (5.23).

Similarly, to prove the sufficient condition, we need to find an upper-bound of Eq. (K.1), which is given by

$$\begin{aligned}\mathbb{E}_{\widehat{\gamma}}\left[\frac{\Sigma_k \Pi_k}{\Sigma_k - \sigma_e^2\left(\eta_k - 1\right)}\right]^+ &\leq \mathbb{E}_{\widehat{\gamma}}\left[\frac{\Pi_k}{1 - \sigma_e^2(\omega^*)^{M/k} k \Pi_k}\right]^+ \\ &\leq \mathbb{E}_{\widehat{\gamma}}\left[\frac{\Pi_M}{1 - \sigma_e^2(\omega^*)^{M} M \Pi_M}\right]^+ . \qquad \text{(K.3)}\end{aligned}$$

Upper-bounding Eq. (K.3) away from infinity, we get the sufficient condition given in Eq. (5.24). The proof of Theorem 4 follows. □

APPENDIX L

PROOF OF PROPOSITION 6

Proof. If $\omega^* \leq 1$, from Theorem 4 we know that $P^*_{\text{total}}(\boldsymbol{\nu}) = 0$ always holds. Thus, the spectral efficiency $\mathcal{C}$ given in Eq. (5.26) is always equal to zero. The outage probability is equal to one.

Otherwise, if $\omega^* > 1$, based on Eq. (5.18), it is *sufficient* to construct a nonempty region such that

$$0 \leq \max_k \{\Sigma_k\} \leq \min_k \left\{\sigma_e^2(\eta_k - 1)\right\}. \tag{L.1}$$

Inside this region, there is no positive solution for Eq. (5.18) for all $k \in \{1, 2, ..., M\}$, and thus $P^*_{\text{total}}(\boldsymbol{\nu}) = 0$ always holds. It is clear that in Eq. (L.1), $\max_k\{\Sigma_k\} = \Sigma_1 = \widehat{\gamma}_{\pi(1)}$. On the other hand,

$$\min_k \left\{\sigma_e^2\left(\eta_k - 1\right)\right\} \geq \sigma_e^2\left(\omega^* \widehat{\gamma}_{\pi(M)}/\widehat{\gamma}_{\pi(1)} - 1\right) \tag{L.2}$$

where the inequality holds since $\min_k\{(\omega^*)^{M/k}\} = \omega^*$ due to $\omega^* > 1$, $\min_k\{k\Sigma_k\} \geq \widehat{\gamma}_{\pi(M)}$ from the definition of the harmonic mean, and $\min_k\{\Pi_k\} = \Pi_1 = 1/\widehat{\gamma}_{\pi(1)}$. Combining Eqs. (L.1) and (L.2), we get the following inequalities:

$$\widehat{\gamma}_{\pi(1)} \leq \sigma_e^2\left(\omega^* \widehat{\gamma}_{\pi(M)}/\widehat{\gamma}_{\pi(1)} - 1\right). \tag{L.3}$$

Solving the inequalities given in Eq. (L.3) and noting that $\widehat{\gamma}_{\pi(1)} \geq \widehat{\gamma}_{\pi(M)} \geq 0$, we get the boundary conditions for this region as follows:

$$\begin{cases} 0 \leq \widehat{\gamma}_{\pi(M)} \leq \sigma_e^2(\omega^* - 1) \\ \widehat{\gamma}_{\pi(M)} \leq \widehat{\gamma}_{\pi(1)} \leq \min\left\{\omega^* \widehat{\gamma}_{\pi(M)}, \sigma_e^2(\omega^* - 1)\right\} \end{cases} \tag{L.4}$$

As long as $\omega^* > 1$ and $\sigma_e^2 > 0$, the probability measure of the region indicated by Eq. (L.4) is nonzero, which is a *lower-bound* for the outage probability. Therefore, the outage probability is nonzero. The proof follows. $\square$

APPENDIX M

PROOF OF THEOREM 6

Proof. Based on the concavity of $\widetilde{R}_{AF}(\boldsymbol{\nu})$, it is easy to show that $(P1'')$ is a strictly convex optimization problem and therefore has the unique optimal solution. Construct the Lagrange as follows:

$$\begin{aligned}\mathcal{J}_1 \;=\; & \mathbb{E}_{\boldsymbol{\gamma}}\left[\left(1+2\gamma_1 P_s(\boldsymbol{\nu})+\frac{2\gamma_2 P_s(\boldsymbol{\nu})\gamma_3 P_r(\boldsymbol{\nu})}{\gamma_2 P_s(\boldsymbol{\nu})+\gamma_3 P_r(\boldsymbol{\nu})}\right)^{-\frac{\beta}{2}}\right] \\ & + \lambda\Bigg(\mathbb{E}_{\boldsymbol{\gamma}}\left[P_s(\boldsymbol{\nu})+P_r(\boldsymbol{\nu})\right]-\overline{P}\Bigg). \qquad \text{(M.1)}\end{aligned}$$

If there exists the solution $\mathbf{P}(\boldsymbol{\nu})$ such that both $P_s(\boldsymbol{\nu}) > 0$ and $P_r(\boldsymbol{\nu}) > 0$, then according to Karush-Kuhn-Tucker (KKT) condition we get

$$\frac{\partial \mathcal{J}_1}{\partial P_i(\boldsymbol{\nu})} = 0, \text{ for } i \in \{s, r\} \qquad \text{(M.2)}$$

which yields

$$\left\{\begin{array}{l} \left(\gamma_1+\frac{\gamma_2\gamma_3^2 P_r^2(\boldsymbol{\nu})}{[\gamma_2 P_s(\boldsymbol{\nu})+\gamma_3 P_r(\boldsymbol{\nu})]^2}\right)\left(1+2\gamma_1 P_s(\boldsymbol{\nu})+\frac{2\gamma_2 P_s(\boldsymbol{\nu})\gamma_3 P_r(\boldsymbol{\nu})}{\gamma_2 P_s(\boldsymbol{\nu})+\gamma_3 P_r(\boldsymbol{\nu})}\right)^{-1-\frac{\beta}{2}} = \gamma_0 \\ \frac{\gamma_2^2\gamma_3 P_s^2(\boldsymbol{\nu})}{[\gamma_2 P_s(\boldsymbol{\nu})+\gamma_3 P_r(\boldsymbol{\nu})]^2}\left(1+2\gamma_1 P_s(\boldsymbol{\nu})+\frac{2\gamma_2 P_s(\boldsymbol{\nu})\gamma_3 P_r(\boldsymbol{\nu})}{\gamma_2 P_s(\boldsymbol{\nu})+\gamma_3 P_r(\boldsymbol{\nu})}\right)^{-1-\frac{\beta}{2}} = \gamma_0 \end{array}\right. \qquad \text{(M.3)}$$

where $\gamma_0 \triangleq \lambda/\beta$. Solving Eq. (M.3), we can obtain Eq. (7.10). Note that Eq. (7.10) is a feasible solution when $u > 0$ and $P_r(\boldsymbol{\nu}) > 0$. Otherwise, the AF protocol reduces to direct transmission, and thus the problem can be solved by the similar approach used in [53], which leads to Eq. (7.12). Finally, the parameter γ_0 is determined by the mean total network power constraint. The proof follows. □

APPENDIX N

POWER LIMIT FOR DIFFERENT DF RELAY PROTOCOLS

Protocol ($R0$) (Proof of Proposition 9)

Proof. As $\theta \to \infty$, the optimal resource allocation policy for the original DF relay protocol ($R0$) becomes Eq. (7.27). To prove Proposition 9, it is equivalent to show that σ in Eq. (7.27) is always equal to zero for any finite power constraint $\overline{P}$.

Using Eq. (7.27), it is easy to show that the total transmit power can be expressed as

$$P_s(\boldsymbol{\nu}) + P_r(\boldsymbol{\nu}) = \frac{(\gamma_2+\gamma_3-\gamma_1)\sigma}{2\gamma_2\gamma_3}\mathcal{I}\{\gamma_2 \geq \gamma_1\} + \frac{\sigma}{2\gamma_2}\mathcal{I}\{\gamma_2 < \gamma_1\}$$

Therefore, the constant σ is determined by the following equation:

$$\begin{aligned} \frac{2\overline{P}}{\sigma} &= \underbrace{\mathbb{E}_{\boldsymbol{\gamma}}\left[\frac{\gamma_2+\gamma_3-\gamma_1}{\gamma_2\gamma_3}\middle|\gamma_2 \geq \gamma_1\right]}_{A}\Pr\{\gamma_2 \geq \gamma_1\} \\ &\qquad + \underbrace{\mathbb{E}_{\boldsymbol{\gamma}}\left[\frac{1}{\gamma_2}\middle|\gamma_1 > \gamma_2\right]}_{B}\Pr\{\gamma_1 > \gamma_2\} \qquad \text{(N.1)} \end{aligned}$$

where we can show, but omit the details, that

$$\begin{cases} A = \lambda_3\Gamma(0^+)\left[\frac{\lambda_2}{\lambda_1}\log\left(1+\frac{\lambda_1}{\lambda_2}\right) + \frac{\lambda_1}{\lambda_1+\lambda_2}\right] \\ B = \lambda_2\left[\Gamma(0^+) - \log\left(1+\frac{\lambda_1}{\lambda_2}\right)\right] \end{cases} \qquad \text{(N.2)}$$

where $\Gamma(0^+) = \lim_{x\to 0^+}\Gamma(x) = +\infty$. Therefore, we have $A = B = +\infty$, which results in $\sigma = 0$ for any finite power constraint $\overline{P}$. The proof follows. □

Protocol ($R1$)

Proof. As $\theta \to \infty$, the optimal resource allocation policy for protocol $(R1)$ becomes

$$\begin{cases} P_s(\boldsymbol{\nu}) = \dfrac{\sigma}{2\gamma_2}\mathcal{I}\{\gamma_2 \geq \gamma_1\} + \dfrac{\sigma}{2\gamma_1}\mathcal{I}\{\gamma_2 < \gamma_1\} \\ P_r(\boldsymbol{\nu}) = \tilde{u}P_s(\boldsymbol{\nu})\mathcal{I}\{\gamma_2 \geq \gamma_1\}. \end{cases} \tag{N.3}$$

Similarly to Appendix N, we prove that for protocol $(R1)$, σ in Eq. (N.3) is always equal to zero for any finite power constraint $\overline{P}$. Omitting the details, we can show that the constant σ is determined by the following equation:

$$\begin{aligned} \frac{2\overline{P}}{\sigma} &= \underbrace{\mathbb{E}_{\boldsymbol{\gamma}}\left[\frac{\gamma_2+\gamma_3-\gamma_1}{\gamma_2\gamma_3}\middle|\gamma_2 \geq \gamma_1\right]}_{A}\Pr\{\gamma_2 \geq \gamma_1\} \\ &\qquad + \underbrace{\mathbb{E}_{\boldsymbol{\gamma}}\left[\frac{1}{\gamma_1}\middle|\gamma_1 > \gamma_2\right]}_{B'}\Pr\{\gamma_1 > \gamma_2\} \end{aligned} \tag{N.4}$$

where $A = +\infty$ from Eq. (N.2) and

$$\begin{aligned} B' &= \mathbb{E}_{\boldsymbol{\gamma}}\left[\frac{1}{\gamma_1}\middle|\gamma_1 \geq \gamma_2\right] = \int_0^\infty \left(\int_{\gamma_2}^\infty \frac{1}{\gamma_1}\lambda_1 e^{-\lambda_1\gamma_1}d\gamma_1\right)\lambda_2 e^{-\lambda_2\gamma_2}d\gamma_2 \\ &= \int_0^\infty \left[-\lambda_1 E_i(-\lambda_1\gamma_2)\right]\lambda_2 e^{-\lambda_2\gamma_2}d\gamma_2 \overset{(a)}{=} \lambda_1 \log\left(1+\frac{\lambda_2}{\lambda_1}\right) \end{aligned} \tag{N.5}$$

where $E_i(x)$ denotes the exponential integral function, and the equation of (a) holds due to the results given in [74, Sec. 6.224.1]. Again, we have $\sigma = 0$ for any finite power constraint $\overline{P}$. The proof follows. □

Protocol $(R2)$ (Proof of Proposition 10)

Proof. As $\theta \to \infty$, the optimal resource allocation policy for protocol $(R2)$ becomes

$$\begin{cases} P_s(\boldsymbol{\nu}) = \dfrac{\sigma}{2\gamma_2}\mathcal{I}\{\gamma_2 \geq \gamma_1 \text{ and } \gamma_3 \geq \gamma_1\} + \dfrac{\sigma}{2\gamma_1}\mathcal{I}\{\gamma_2 < \gamma_1 \text{ or } \gamma_3 < \gamma_1\} \\ P_r(\boldsymbol{\nu}) = \tilde{u}P_s(\boldsymbol{\nu})\mathcal{I}\{\gamma_2 \geq \gamma_1 \text{ and } \gamma_3 \geq \gamma_1\}. \end{cases} \tag{N.6}$$

To prove Proposition 10, it is equivalent to show that for protocol $(R2)$, σ in Eq. (N.6) is bounded away from zero for any finite power constraint $\overline{P}$. Likewise, the following equations hold for the constant σ:

$$\begin{aligned}
\frac{2\overline{P}}{\sigma} &= \mathbb{E}_{\boldsymbol{\gamma}}\left[\frac{\gamma_2+\gamma_3-\gamma_1}{\gamma_2\gamma_3}\middle|\gamma_2 \geq \gamma_1 \text{ and } \gamma_3 \geq \gamma_1\right]\Pr\{\gamma_2 \geq \gamma_1 \text{ and } \gamma_3 \geq \gamma_1\} \\
&\qquad + \mathbb{E}_{\boldsymbol{\gamma}}\left[\frac{1}{\gamma_1}\middle|\gamma_1 > \gamma_2 \text{ or } \gamma_1 > \gamma_3\right]\Pr\{\gamma_1 > \gamma_2 \text{ or } \gamma_1 > \gamma_3\} \\
&< \mathbb{E}_{\boldsymbol{\gamma}}\left[\frac{\gamma_2+\gamma_3-\gamma_1}{\gamma_2\gamma_3}\middle|\gamma_2 \geq \gamma_1 \text{ and } \gamma_3 \geq \gamma_1\right] + \mathbb{E}_{\boldsymbol{\gamma}}\left[\frac{1}{\gamma_1}\middle|\gamma_1 > \gamma_2 \text{ or } \gamma_1 > \gamma_3\right] \\
&< \mathbb{E}_{\boldsymbol{\gamma}}\left[\frac{1}{\gamma_2}+\frac{1}{\gamma_3}\middle|\gamma_2 \geq \gamma_1 \text{ and } \gamma_3 \geq \gamma_1\right] + \mathbb{E}_{\boldsymbol{\gamma}}\left[\frac{1}{\gamma_1}\middle|\gamma_1 > \gamma_2 \text{ or } \gamma_1 > \gamma_3\right] \\
&< \mathbb{E}_{\boldsymbol{\gamma}}\left[\frac{1}{\gamma_2}\middle|\gamma_2 \geq \gamma_1\right] + \mathbb{E}_{\boldsymbol{\gamma}}\left[\frac{1}{\gamma_3}\middle|\gamma_3 \geq \gamma_1\right] + \mathbb{E}_{\boldsymbol{\gamma}}\left[\frac{1}{\gamma_1}\middle|\gamma_1 > \gamma_2\right] + \mathbb{E}_{\boldsymbol{\gamma}}\left[\frac{1}{\gamma_1}\middle|\gamma_1 > \gamma_3\right] \\
&= \lambda_2 \log\left(1+\frac{\lambda_1}{\lambda_2}\right) + \lambda_3 \log\left(1+\frac{\lambda_1}{\lambda_3}\right) + \lambda_1 \log\left(\left(1+\frac{\lambda_2}{\lambda_1}\right)\left(1+\frac{\lambda_3}{\lambda_1}\right)\right). \quad \text{(N.7)}
\end{aligned}$$

Therefore, σ is bounded away from zero for any finite power constraint $\overline{P}$. The proof follows. □

APPENDIX O

DERIVATION OF EQ. (8.14)

Expanding the Bessel function in Eq. (8.11) by an infinite series [118], and then solving the integral of Eq. (8.10), we can derive the closed-form expression of the pdf $f_{\widetilde{\Gamma}}(\widetilde{\gamma})$ as follows:

$$\begin{aligned} f_{\widetilde{\Gamma}}(\widetilde{\gamma}) \;=\; & \frac{L_t\, \widetilde{\gamma}^{L_r-1} \exp\left(-\frac{m\widetilde{\gamma}}{(1-\rho)\overline{\gamma}}\right)}{[(L_r-1)!]^2} \left[\frac{m}{\overline{\gamma}(1-\rho)}\right]^{L_r} \\ & \cdot \sum_{i=0}^{L_t-1} (-1)^i \binom{L_t-1}{i} \sum_{j=0}^{i(L_r-1)} \Bigg\{ \xi_{ji}(j+L_r-1)! \\ & \cdot \left[\frac{1-\rho}{i(1-\rho)+1}\right]^{j+L_r} {}_1F_1\left(j+L_r\,; L_r; \frac{\rho m \widetilde{\gamma}}{[i(1-\rho)+1](1-\rho)\overline{\gamma}}\right)\Bigg\} \end{aligned} \tag{O.1}$$

where ${}_1F_1(\,\cdot\,;\,\cdot\,;\,\cdot\,)$ denotes the confluent (Kummer) hypergeometric function [118], which can be expressed as a finite series expansion by [131]. Thus, we obtain a more explicit closed-form expression for $f_{\widetilde{\Gamma}}(\widetilde{\gamma})$ as follows:

$$\begin{aligned} f_{\widetilde{\Gamma}}(\widetilde{\gamma}) \;=\; & \frac{L_t}{(L_r-1)!} \sum_{i=0}^{L_t-1} (-1)^i \binom{L_t-1}{i} \exp\left(-\frac{m(i+1)\widetilde{\gamma}}{[i(1-\rho)+1]\overline{\gamma}}\right) \\ & \cdot \sum_{j=0}^{i(L_r-1)} \Bigg\{ \xi_{ji}(j+L_r-1)! \sum_{k=0}^{j} \binom{j}{k} \left(\frac{m}{\overline{\gamma}}\right)^{k+L_r} \\ & \cdot \frac{\rho^k (1-\rho)^{j-k}}{[i(1-\rho)+1]^{j+k+L_r}} \frac{\widetilde{\gamma}^{k+L_r-1}}{(k+L_r-1)!} \Bigg\}. \end{aligned} \tag{O.2}$$

Substituting Eq. (O.2) into Eq. (8.13) and using the approach proposed in [115], we obtain the closed-form expression of $\widetilde{I}(\overline{\gamma}, \beta, \theta)$ as shown in Eq. (8.14).

APPENDIX P

DERIVATION OF EQ. (9.8)

Combining all the terms which contain γ_2 in Eq. (9.7), we can define an integral function $I(\gamma_1)$ as follows:

$$I(\gamma_1) \triangleq \int_0^{\gamma_1} \exp\left(-\left[(i+1)\lambda + \frac{g\beta}{\sin^2\theta}\right]\gamma_2\right)\gamma_2^{j+L-1} d\gamma_2. \tag{P.1}$$

Defining $\phi \triangleq (i+1)\lambda + \big(g\beta/\sin^2\theta\big)$, we can solve the integral function $I(\gamma_1)$ analytically as follows [118]:

$$I(\gamma_1) = \left(\frac{1}{\phi}\right)^{(j+L)} \gamma\Big(j+L, \phi\gamma_1\Big) \tag{P.2}$$

where $\gamma(\cdot,\cdot)$ denotes the incomplete Gamma function [118]. After solving the inner integral of Eq. (9.7), the SER P_M then becomes:

$$\begin{aligned} P_M &= \frac{\kappa N(N-1)}{\pi[(L-1)!]^2} \int_0^{\Theta}\int_0^{\infty} \gamma_1^{L-1} \exp\Big(-\Big[\lambda + \frac{g\alpha}{\sin^2\theta}\Big]\gamma_1\Big) \sum_{i=0}^{N-2} (-1)^i \binom{N-2}{i} \\ &\quad \cdot \sum_{j=0}^{i(L-1)} \xi_{ji}\lambda^{j+2L} I(\gamma_1) d\gamma_1\, d\theta \end{aligned} \tag{P.3}$$

where $I(\gamma_1)$ is specified by Eq. (P.2). Then, combining all the terms which contain γ_1 in Eq. (P.3), we can solve the inner integral of Eq. (P.3) as follows [74]:

$$\begin{aligned} &\int_0^{\infty} \exp\left(-\left[\lambda + \frac{g\alpha}{\sin^2\theta}\right]\gamma_1\right)\gamma_1^{L-1} I(\gamma_1) d\gamma_1 \\ &\quad = \frac{(j+2L-1)!}{(j+L)} \left[\frac{\sin^2\theta}{(i+2)\lambda\sin^2\theta + g(\alpha+\beta)}\right]^{j+2L} \\ &\quad\quad \cdot\ {}_2F_1\left(1, j+2L; j+L+1; \frac{\lambda\sin^2\theta + g\alpha}{(i+2)\lambda\sin^2\theta + g(\alpha+\beta)}\right). \end{aligned} \tag{P.4}$$

Then, the SER P_M can be expressed as follows:

$$\begin{aligned} P_M \;=\; & \frac{\kappa N(N-1)}{\pi[(L-1)!]^2} \int_0^{\Theta} \sum_{i=0}^{N-2} (-1)^i \binom{N-2}{i} \sum_{j=0}^{i(L-1)} \xi_{ji} \\ & \cdot \frac{(j+2L-1)!}{(j+L)} \left[\frac{\lambda \sin^2\theta}{(i+2)\lambda \sin^2\theta + g(\alpha+\beta)} \right]^{j+2L} \\ & \cdot {}_2F_1\!\left(1, j+2L; j+L+1; \frac{\lambda \sin^2\theta + g\alpha}{(i+2)\lambda \sin^2\theta + g(\alpha+\beta)} \right) d\theta. \end{aligned} \tag{P.5}$$

Furthermore, by using [132], we obtain the SER P_M as shown in Eq. (9.8).

APPENDIX Q

DERIVATION OF EQ. (9.13)

Expanding the Bessel function by representing it as a infinite series [118], we obtain an alternative expression for the conditional pdf $p_{\widetilde{\Gamma}|\Gamma}(\widetilde{\gamma}|\gamma)$ as

$$p_{\widetilde{\Gamma}|\Gamma}(\widetilde{\gamma}|\gamma) = \left(\frac{\lambda}{1-\rho}\right)^{L} \widetilde{\gamma}^{L-1} \exp\left(-\frac{\lambda(\rho\gamma+\widetilde{\gamma})}{1-\rho}\right) \sum_{k=0}^{\infty} \frac{1}{k!(L+k-1)!} \left[\frac{\lambda^2\rho\widetilde{\gamma}\gamma}{(1-\rho)^2}\right]^{k}. \tag{Q.1}$$

By using the similar approach as used in Appendix P, we first combine all the terms containing γ_2 in Eq. (9.10) to solve the inner integral. Then, we combine all the terms containing γ_1 in Eq. (9.10) to solve the outer integral. Finally, we obtain the closed-form joint pdf $f_{\widetilde{\Gamma}_1,\widetilde{\Gamma}_2}(\widetilde{\gamma}_1,\widetilde{\gamma}_2)$ expressed as follows:

$$\begin{aligned} f_{\widetilde{\Gamma}_1,\widetilde{\Gamma}_2}(\widetilde{\gamma}_1,\widetilde{\gamma}_2) = {} & \frac{N(N-1)}{[(L-1)!]^2} \exp\left(-\frac{\lambda(\widetilde{\gamma}_1+\widetilde{\gamma}_2)}{1-\rho}\right) \sum_{i=0}^{N-2} (-1)^i \binom{N-2}{i} \sum_{j=0}^{i(L-1)} \xi_{ji} \\ & \cdot \sum_{n=0}^{\infty}\sum_{k=0}^{\infty} \frac{\lambda^{2L+k+n}\rho^{k+n}\widetilde{\gamma}_1^{n+L-1}\widetilde{\gamma}_2^{k+L-1}}{(1-\rho)^{k+n-j}[i(1-\rho)+2]^{j+k+n+2L}} \\ & \cdot \frac{(j+k+n+2L-1)!}{k!n!(k+L-1)!(n+L-1)!(j+k+L)} \\ & \cdot {}_2F_1\left(1, j+k+n+2L; j+k+L+1; \frac{i(1-\rho)+1}{i(1-\rho)+2}\right). \end{aligned} \tag{Q.2}$$

Plugging Eq. (Q.2) into Eq. (9.12) and solving the inner two-fold integrals with respect to $\widetilde{\gamma}_1$ and $\widetilde{\gamma}_2$ [74], we obtain the SER $P_M^{(d)}$ as shown in Eq. (9.13).

VITA

Jia Tang received his B.S. degree in electrical engineering from Xi'an Jiaotong University, Xi'an, China, in 2001, and received his Ph.D. degree in computer engineering from the Department of Electrical and Computer Engineering, Texas A&M University, College Station, Texas, USA, in 2006. His research interests include mobile wireless communications and networks, with emphasis on cross-layer design and optimizations, wireless quality-of-service (QoS) provisioning for mobile multimedia networks, wireless diversity techniques, and wireless resource allocation.

Dr. Tang received Fouraker Graduate Research Fellowship Award from the Department of Electrical and Computer Engineering, Texas A&M University in 2005, for excellence in research performance. He is a member of Phi Kappa Phi honor society. He may be reached at the Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX 77843-3128, his e-mail address is jtang@ece.tamu.edu.