ADAPTIVE RESOURCE ALLOCATION FOR STATISTICAL QoS PROVISIONING IN MOBILE WIRELESS COMMUNICATIONS AND NETWORKS

A Dissertation

by

QINGHE DU

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 2010

Major Subject: Computer Engineering

ADAPTIVE RESOURCE ALLOCATION FOR STATISTICAL QoS PROVISIONING IN MOBILE WIRELESS COMMUNICATIONS AND NETWORKS

A Dissertation

by

QINGHE DU

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
	Jennifer L. Welch
	Dmitri Loguinov
Head of Department,	Costas N. Georghiades

December 2010

Major Subject: Computer Engineering

ABSTRACT

Adaptive Resource Allocation for Statistical QoS Provisioning in Mobile Wireless Communications and Networks. (December 2010)

Qinghe Du, B.S., Xi'an Jiaotong University, Xi'an, China;

M.S., Xi'an Jiaotong University, Xi'an, China

Chair of Advisory Committee: Dr. Xi Zhang

Due to the highly-varying wireless channels over time, frequency, and space domains, statistical QoS provisioning, instead of deterministic QoS guarantees, has become a recognized feature in the next-generation wireless networks. In this dissertation, we study the adaptive wireless resource allocation problems for statistical QoS provisioning, such as guaranteeing the specified delay-bound violation probability, upper-bounding the average loss-rate, optimizing the average goodput/throughput, etc., in several typical types of mobile wireless networks.

In the first part of this dissertation, we study the statistical QoS provisioning for mobile multicast through the adaptive resource allocations, where different multicast receivers attempt to receive the common messages from a single base-station sender over broadcast fading channels. Because of the heterogeneous fading across different multicast receivers, both instantaneously and statistically, how to design the efficient adaptive rate control and resource allocation for wireless multicast is a widely cited open problem. We first study the time-sharing based goodput-optimization problem for non-realtime multicast services. Then, to more comprehensively characterize the QoS provisioning problems for mobile multicast with diverse QoS requirements, we further integrate the statistical delay-QoS control techniques — effective capacity theory, statistical loss-rate control, and information theory to propose a QoS-driven optimization framework. Applying this framework and solving for the corresponding

optimization problem, we identify the optimal tradeoff among statistical delay-QoS requirements, sustainable traffic load, and the average loss rate through the adaptive resource allocations and queue management. Furthermore, we study the adaptive resource allocation problems for multi-layer video multicast to satisfy diverse statistical delay and loss QoS requirements over different video layers. In addition, we derive the efficient adaptive erasure-correction coding scheme for the packet-level multicast, where the erasure-correction code is dynamically constructed based on multicast receivers' packet-loss statuses, to achieve high error-control efficiency in mobile multicast networks.

In the second part of this dissertation, we design the adaptive resource allocation schemes for QoS provisioning in unicast based wireless networks, with emphasis on statistical delay-QoS guarantees. First, we develop the QoS-driven time-slot and power allocation schemes for multi-user downlink transmissions (with independent messages) in cellular networks to maximize the delay-QoS-constrained sum system throughput. Second, we propose the delay-QoS-aware base-station selection schemes in distributed multiple-input-multiple-output systems. Third, we study the queue-aware spectrum sensing in cognitive radio networks for statistical delay-QoS provisioning. Analyses and simulations are presented to show the advantages of our proposed schemes and the impact of delay-QoS requirements on adaptive resource allocations in various environments.

To My Parents

ACKNOWLEDGMENTS

First and foremost, I would like to express my sincere and deep appreciation to my Ph.D. advisor, Professor Xi Zhang, for all his contributions of time, ideas, and funding support to this research. Without his guidance and help throughout my studies at Texas A&M University, it would have been impossible for me to complete this dissertation. He offered me a lot of valuable guidance as well as sufficient freedom of surfing for research on wireless networking techniques. He is always willing to share his experiences helping and encouraging me for my research. He gives me strict training on cultivating ideas, conducting rigorous research, and writing technical papers, from which I benefit significantly to improve my ability of independent research. More importantly, his passion for exploring the new research areas and his persistent commitment to high-quality research through hard work will always inspire and motivate me in my future career.

I would like to express my gratitude to Professor Costas N. Georghiades, Professor Jennifer L. Welch and Professor Dmitri Loguinov for serving on my dissertation committee. They offered me a lot of valuable guidance and insightful advice on my research. I would like to give my special thanks to Professor Costas N. Georghiades for his guidance on modulation theory and his generous help for my Ph.D. studies. I am also very grateful to Professor Krishna R. Narayanan and Professor Jean-Franǫois Chamberland. I learned a lot from them on wireless communications and signal processing techniques.

I also thank my colleagues and friends for their friendship and support during my Ph.D. study at Texas A&M University. I especially thank Jia Tang and Hang Su for their intensive discussions and insightful suggestions on my research. I would like to sincerely thank my friends at Texas A&M University, Jie Ke, Lingjia Liu, Yang Yi,

Zhuo Feng, Yifang Liu, Jing Jiang, Yuqiang Bai, Chen Zhao, Xi Chen, and Yanjia Hong, among others. I would like to thank my M.S. advisor, Professor Shihua Zhu, at Xi'an Jiaotong University, China. I would not even have had the chance to pursue my Ph.D. at Texas A&M University without his persistent guidance, encouragement, and support on my graduate study in China. I would also like to thank Professor Pinyi Ren for his generous help during my study in Xi'an Jiaotong University. Finally, but most importantly, this dissertation is dedicated to my father, Baiqing Du, and my mother, Shuangwei Wu, for their infinite love and support. The funding for this research was supported in part by the U.S. National Science Foundation CAREER Award under Grant ECS-0348694, and in part by a Teaching Assistantship from the Department of Electrical and Computer Engineering, Texas A&M University.

TABLE OF CONTENTS

CHAPTER Page

CHAPTER Page

CHAPTER Page

CHAPTER Page

CHAPTER Page

CHAPTER Page

LIST OF TABLES

LIST OF FIGURES

FIGURE Page

FIGURE Page

FIGURE Page

FIGURE Page

FIGURE Page

CHAPTER I

INTRODUCTION

A. Background and Motivations

1. Challenges of Statistical QoS Provisioning in Mobile Wireless Networks

With the rapid evolution of communication techniques in mobile wireless networks, more and more mobile wireless services have been implemented in our daily lives, such as the traffic-avoidance navigation, web browsing, e-mail access, online gaming, text chatting, teleconferencing, on-demand video streaming, etc. As various mobile wireless services have different requirements on quality-of-service (QoS), QoS provisioning over wireless networks has become one of the well-recognized features in next-generation wireless networks.

Unlike wireline networks, the qualities of wireless fading channels are highly-varying over time, frequency, and space domains. As a result, deterministic QoS provisioning, such as hard delay bound, steady transmission/service rate, etc., is unrealistic to implement in practical wireless communications systems. Alternatively, statistical QoS guarantees have become one of the major objectives and guidelines for the design of wireless networks.

Adaptive resource allocation is one of the effective approaches for statistical QoS provisioning of mobile wireless services, which can dynamically allocate wireless resources, such as power, time slots, and frequency bandwidths, to control the service-rate process based on the channel quality and QoS requirements. However, the statistical QoS provisioning over wireless networks still faces many challenges. First,

---

The journal model is *IEEE Journal on Selected Areas in Communications.*

various mobile wireless services are interested in different statistical QoS metrics. For example, non-real-time services expect high average system throughput/goodput and low average loss rate, while not addressing the delay bound of each data packet. In contrast, real-time services not only require high system throughput, but also specify the delay bound constraint for the transmitted data stream with a small violation probability. Second, for delay-sensitive services, the delay-bound constraints vary significantly with different applications. The delay bound for video teleconferencing is typically hundreds of milliseconds [1]; but the delay bound live for text chatting may be much longer. How to accurately characterize and guarantee different levels of delay constraints over wireless channels through a systematic and unified approach is still an open problem. Third, a mobile wireless service often has requirements on multiple QoS metrics. For instance, real-time services can sacrifice certain loss to meet the delay constraint, but the loss rate needs to be controlled in a low level. As a result, the joint design over multiple QoS metrics significantly increases difficulty and complexity for networks design and implementation. Finally, with the emergence of new wireless technologies such as cognitive radio networks (CRN), there is the urgent need to develop new QoS-aware adaptive resource allocation schemes to efficiently use the scarce wireless spectrum resources. Motivated by the above challenges, in this dissertation we tackle the statistical QoS provisionings through adaptive resource allocation in several types of wireless networks, with emphasis on: 1) throughput/goodput optimization for wireless multicast services with statistical delay-QoS and/or loss-QoS provisioning; 2) statistical delay-QoS provisioning in downlink transmissions in cellular networks, distributed multi-input-multi-output (MIMO) systems, and Cognitive Radio networks, respectively.

2. Statistical Delay-QoS Guarantees through Adaptive Resource Allocations

Delay bound is one of the most important QoS metric for delay-sensitive mobile wireless services. However, the highly time-varying properties of wireless channels makes it impractical to guarantee a deterministic delay bound, as mentioned in Section A-1. Consequently, statistical delay-QoS metrics, featured with queue-length-bound violation probability, buffer-overflow probability, and delay-bound violation probability have been widely used to characterize the delay-QoS guarantees.

The pioneer research works [2–7] of statistical delay-QoS guarantees mainly focused on *effective bandwidth* theory in asynchronous transfer mode (ATM) networks, where the traffic arrival-rate processes are typically time-varying and network service rates are constant. The effective bandwidth theory shows that the queue-length-bound/delay-bound violation probabilities can be approximated as a exponentially decaying function of the specified bound, which provides us a powerful yet convenient approach for analyzing statistical delay-QoS guarantees.

In [8], the authors proposed the dual concept of effective bandwidth, namely, *effective capacity*, to analyze the queue-length-bound/delay-bound violation probabilities in wireless networks. The effective capacity theory concentrates on the queueing systems with the constant arrival-rate and time-varying service-rate process, addressing the capability of wireless channels, which are often time-varying and bottlenecks in network transmissions, in supporting delay-QoS constrained mobile wireless services. More specifically, the effective capacity defines the maximum traffic load that can be supported by a time-varying service process subject to the statistical delay-QoS constraints. One of the major advantages of the dual concepts of effective bandwidth and effective capacity is to establish the general model which can conveniently characterize delay-QoS with fine-grained delay bounds and violation probabilities. Moreover,

for queueing system with both time-varying arrival-rate and service-rate processes, the analyses on statistical delay-QoS guarantees can be tackled through decomposing the queuing system to two virtual sub-queuing-systems. One sub-queueing-system consists of a time-varying arrival-rate process and a constant departure rate, falling into the framework of effective bandwidth theory; the other is composed of a time-varying service-rate process and a constant arrival rate, which can be described by using the effective capacity theory. Accordingly, we can first study the two sub-queuing-systems separately. Then, the statistical delay-QoS provisioning status for the original queuing system can then be readily obtained by comparing the effective bandwidth and effective capacity of the two sub-queuing-systems, respectively. The details of the effective bandwidth and effective capacity theories will be introduced in Chapter II.

To efficiently use the scarce wireless resources, it is expected that the adaptive resource allocation is aware of not only the variation of the channel conditions, but also the delay-QoS requirements for various mobile wireless services. However, many research works for adaptive resource allocations in wireless networks mainly focused on the following two types of applications: 1) which can tolerate infinite delay; 2) which can not tolerate any delay. The first and the second scenarios are associated with the ergodic capacity [9–15] and outage capacity [16–18], respectively, in information theory. The resource allocation to achieve the ergodic capacity aims at maximizing the average information transmission rate (average throughput). The resource allocation that achieves the outage capacity maximizes the constant information transmission rate under a certain outage probability. Clearly, these two frameworks cannot accurately characterize the diverse delay-QoS requirements, thus resulting in inefficient design for many delay-sensitive services. Moreover, the outage capacity cannot really guarantee zero delay due to the existence of outage state, expect that the outage

probability is equal to zero. Therefore, there are urgent needs to develop QoS-driven adaptive resource allocation schemes for mobile wireless services with diverse delay-QoS requirements in wireless networks.

A great deal of research has been devoted to the QoS-driven resource allocation in wireless networks. In [19] and [20], the authors developed the delay-QoS-driven power and rate adaptation schemes for the single-channel and multi-channel wireless links, respectively. In [21], the authors derived the optimal delay-QoS driven adaptive resource allocation for the wireless link with imperfect CSI. In [22], the authors proposed the delay-QoS driven schemes for cooperative relay networks. These works mainly studied the single point-to-point wireless link, which cannot be applied for the cases with coexistence of multi-links, such as downlink multi-user transmissions with independent messages and multicast transmissions (distribute common message) in cellular wireless networks. The authors of [23] proposed the resource allocation schemes for multi-user downlink transmissions with different QoS requirements. However, advanced wireless techniques such as power control and MIMO are not addressed. Moreover, the emergence of cognitive radio technologies requires new statistical delay-QoS provisioning strategies. To address the above problem, we study the statistical delay-QoS provisioning issues through adaptive resource allocation in several different types of wireless networks. In particular, we derive the statistical delay-QoS guaranteed adaptive resource allocation schemes for wireless multicast, downlink multi-user communications with independent messages, base-station selections in distributed MIMO systems, and spectrum sensing techniques of CRNs, in Chapters IV-V, VII, VIII, and IX, respectively.

3. Adaptive Resource and Rate Allocation for Mobile Multicast

Mobile multicast, which provides a highly efficient way of distributing common data from one source to multiple location-independent receivers [24], has been considered as one of the key techniques for current and future wireless communications systems and networks [25, 26]. In this dissertation, we mainly focus on the mobile multicast services through downlink transmissions in cellular networks, where a base station (BS) is responsible to distribute the same data content to all multicast receivers through broadcast fading channels.

The heterogeneous fading properties across different multicast receivers imposes many challenges in implementing efficient adaptive resource allocation/control for multicast transmissions. A commonly used strategy for wireless multicast is to regulate the resource allocation strategy and the transmission rate based only on the worst-case channel quality among all multicast receivers at any time instant. This strategy has been shown to be inefficient to achieve high system throughput when the number of multicast receivers becomes large [27]. However, increasing the transmission rate for multicast may cause data loss for receivers with poorer instantaneous channel qualities. How to optimize the system throughput while upper-bounding the loss rate over the entire multicast group is still a widely cited open problem.

Despite throughput/goodput optimization and the loss-rate control, delay-QoS provisioning are the other important issues for mobile multicast, because many multicast applications are towards delay-sensitive applications such as teleconferencing and on-line video streaming. As mentioned in Section A-1, for real-time services, we can often sacrifice certain loss to guarantee delay QoS. Consequently, there exists fundamental tradeoff among the delay-QoS provisioning, loss-rate control, and throughput/goodput optimization for mobile multicast over wireless networks, which

need to be thoroughly studied to implement highly-efficient multicast. To overcome the above problems, in this dissertation we develop a set of adaptive resource/rate allocation and erasure-correction coding schemes for mobile multicast services under various delay and loss QoS requirements.

There has been a great deal of research devoted to adaptive resource allocation and rate allocation for mobile multicast services. The authors of [28] developed a beamforming scheme for the multiple-input single-output (MISO) multicast system, which optimizes the worst-case combined signal-to-noise ratio (SNR) for multicast receivers through the sequential quadratic programming. In [29], the authors applied the semi-definite relaxation techniques for transmit beamforming in MISO multicast systems to minimize the transmit power subject to the minimum received SNR constraint. The work of [30] developed a rate-adaptive multimedia multicasting protocol over IEEE 802.11 Wireless LANs. The transmission rate of this scheme is determined by the worst-case channel quality among all receivers. These works mainly aimed at optimizing the worst-case performance across all multicast receivers at any time instant. From information theory perspective, the authors of [27] investigated the physical-layer multicast capacity with multiple transmit antennas equipped at the sender, which goes to zero as the multicast group size becomes very large. In [31], the authors discussed the throughput-delay tradeoff problem for cellular multicast. Specifically, for a multicast session, the transmission rate is determined by the SNR which is in a *fixed* position in the ordered SNR sequence among all multicast receivers. However, we can expect that the dynamic rate-control strategy will be more efficient and flexible as compared to this fixed multicast strategy.

Moreover, many works towards QoS provisionings for multicast services have been proposed [32–36]. In [32–34], the authors developed the scalable flow control protocols and proposed the comprehensive delay analyses for multicast services over

heterogeneous multicast receivers. To handle the heterogeneous channels while efficiently distributing video data, multi-layer video distributions have proven to be one of the efficient ways. The authors of [35] proposed the efficient receiver-driven layered multicast over Internet. However, these multicast strategies in Internet cannot be directly applied into wireless networks due to the highly and rapidly varying wireless channel qualities. This may result in unstable bandwidths and thus unsatisfied loss and delay QoS, if the multicast rate adaptation is solely driven by the receivers. In [37], the authors implemented multi-layer video multicast through minimizing the maximum penalty function such that the overall quality can be guaranteed for all wireless multicast receivers. The authors of [38] proposed a cross-layer approach with adaptive power allocation and channel coding strategies for multi-layer video multicast to improve the peak signal-to-noise ratio (PSNR) QoS. However, these works do not systematically reveal how to efficiently tradeoff among the delay, loss, and throughput QoS requirements. Differen from the above previous work, our major goal in this dissertation is to design efficient adaptive resource allocation schemes to guarantee the specified statistical loss and/or delay QoS requirements while optimizing the desired system throughput performance for mobile multicast services.

4. Adaptive Resource Allocation for Downlink Multiuser Transmission with Independent Messages in Cellular Wireless Networks

The cellular wireless network is one of the most popular wireless network structure. In cellular wireless networks, the base station can simultaneously communicate with multiple mobile users with independent messages over broadcast fading channels. Correspondingly, adaptive resource allocation needs to be designed towards different users' QoS requirements. Extensive research works have been dedicated to resource allocation downlink multiuser communications from information theory per-

spective [10, 15, 18, 39]. In [10, 15], the authors derived the ergodic capacity region, which corresponds to the scenario without any delay constraints, and the corresponding optimal power and/or time-slot allocation schemes for transmitting different information to multiple users. On the other hand, the authors of [18] proposed the optimal resource allocation for outage capacities. This work guarantees a constant service rate for each user with a certain outage probability. Clearly, these results cannot effectively characterize and support fine-grained delay QoS requirements in future wireless networks. Moreover, these works assume that all mobile users have the same delay constraint. Correspondingly, how to use a unified approach to characterize different delay-QoS requirements over different mobile users is still a widely cited open problem. Using the effective capacity theory, the authors of [23] proposed the downlink multiuser resource allocation schemes with satisfied delay-QoS requirements for all mobile users. However, the adaptive power control, which is one of the major approach to combat random wireless fading for efficient QoS provisioning, is not addressed in [23]. To overcome these problems, we apply the effective capacity theory to derive the QoS-driven power and time-slot allocation schemes for downlink multi-user transmissions, while also addressing the fairness problem across multiple mobile users, which will be detailed in Chapter VII.

5. Base Station Selections in Distributed MIMO Systems

Distributed multiple-input-multiple-output (MIMO) [40–44] is an advanced techniques in wireless networks. Specifically, the distributed MIMO techniques are responsible for organizing multiple location-independent BS's to build the distributed MIMO system, such that high data transmission throughput can be achieved for mobile users. Similar to the conventional centralized MIMO system [45–47], the system capability can be significantly improved by applying the distributed MIMO tech-

niques. Accordingly, the quality-of-service (QoS) performance is also considerably enhanced as compared to the single antenna system. Recently, the works of [41, 44] have demonstrated the feasibility of the efficient synchronization techniques over distributed MIMO links through both experimental tests and theoretical analyses. Therefore, distributed MIMO techniques are promising to increase the coverage of broadband wireless networks, and thus has become a major component in wireless networks design.

In distributed MIMO system, BS's are also a type of precious resources to effectively improve delay-QoS performances. However, the computational complexity for MIMO signal processing and coding also grow rapidly as more BS's are involved. Also, as the power is allocated across location-independent BS's, the interfering area caused by the distributed MIMO transmissions drastically increases. This effect may severely degrade the spatial frequency-reuse efficiency for the entire network. Thus, it is critically important to minimize the BS usage in distributed MIMO systems, i.e., reducing the number of distributed BS's used for data transmissions.

In centralized MIMO systems, the antenna selection techniques [46, 47] can effectively reduce the complexity. It is straightforward that the antenna selection techniques can be also extended to distributed MIMO systems for the BS selection. However, most previous works for BS/antenna selections mainly focused on the scenarios of selecting a subset of BS's/antennas with the fixed number of BS's/antennas [42,43]. Clearly, by applying the dynamic BS selection strategy, we can further reduce the complexity and interfering range caused by the distributed MIMO transmissions. More importantly, note that distributed MIMO techniques are often applied for high-rate data transmissions towards multimedia applications. Thus, the efficient BS selection schemes needs to be driven by not only by the channel qualities between the BS transmitters and mobile users, but also diverse delay-QoS requirements imposed by

the mobile users. In this dissertation, to address the above problems for distributed MIMO systems, we develop the delay-QoS aware BS selection schemes for single-user and multi-user scenarios, respectively, as elaborated on in Chapter VIII.

6. Spectrum Accesses in Cognitive Radio Networks

Due to the increasing demands on wireless services and considerable underutilization of the licensed spectrums [48], cognitive radio techniques have been extensively studied to improve the utilization of *wireless spectrum resources*. In cognitive radio networks, unlicensed users, also known as secondary users (SU) are allowed to use the licensed spectrums when the licensed users, also called primary users (PU), do not occupy these spectrum bandwidths [48–51]. Moreover, the access opportunities of secondary users are discovered by using spectrum sensing techniques which can detect the spectrum holes across licensed radio bands.

Among various spectrum sensing techniques, energy detection with the threshold-based decision is widely applied [50–52]. The traditional energy-detection based schemes typically compare the received energy with the fixed threshold to decide whether the spectrum is occupied. The threshold is selected such that the probability of causing interference to the PUs does not exceed a specified small value. In such a case, the PUs will not feel the existence of the SUs. However, the traditional energy-detection based spectrum-sensing strategy uses the fixed threshold, which does not adapt to the queueing status of the SUs. As a result, the spectrum resources cannot be efficiently used for SUs, to support the statistical QoS requirements [3, 8, 53] such as the queue-length-bound violation probability or buffer-overflow probability.

To effectively decrease the queue-length-bound violation or buffer-overflow probabilities, we can take into account the queueing status of SU sender for spectrum sensing. When the queue length is small, the SUs can use a relatively conservative

strategy which causes interference to PUs with a lower probability. On the other hand, when the queue length is approaching the upper limit, SUs will apply more aggressive strategies to reduce the chances of buffer overflow or queue-length-bound violation. Then, we can significantly reduce the queue-length-bound violation probability or buffer-overflow probability can be decreased while still upper-bounding the average interfering probability to the PUs. Following this strategy, we develop the efficient QoS-aware spectrum sensing techniques for CRNs in Chapter IX.

B. Contributions of the Dissertation

In the first part of this paper:

1. We propose to maximize average multicast goodput QoS over all multicast receivers through time-sharing based rate allocation. We obtain the optimal rate adaptation scheme when all multicast receivers' wireless channels are independent and identically distributed (i.i.d.). Our results show that in order to achieve high average goodput, the transmission rate cannot be always determined by the worst-case channel quality over all receivers. Also, we design a sub-grouping strategy, which applies the optimal scheme for the i.i.d. scenario into the non-i.i.d. scenario to achieve suboptimal yet efficient multicast rate control.

2. By integrating the effective capacity theory, information theory, and statistical loss-rate control, we develop the optimal time-sharing (TS) based and superposition-coding (SPC) based adaptive resource allocation schemes, respectively, to maximize the multicast effective capacity under various delay and loss QoS requirements. The TS and SPC based multicast transmissions use time-slot allocation and power allocation, respectively, to handle the different fading

statuses of different receivers for efficient yet flexible rate adaptation. Our results reveal the fundamental tradeoff among multicast throughput, loss-rate of multicast receivers, and the statistical delay-QoS provisioning for mobile multicast services. Also, we obtain the optimal multicast policies under the two limiting scenarios of delay-QoS constraints: 1) the scenario which can tolerate infinite delay; 2) the scenario which cannot tolerate any delay.

3. We propose the statistical delay-QoS provisioning framework for multi-layer video unicast/multicast, where unicast can be treated as a special case of multicast with only one multicast receiver. The framework imposes different loss rates for different video layers due to their different importance levels. Meanwhile, the synchronous transmissions are required across all video layers, implying that all video layers have the same delay-bound and the associated violation-probability constraints. We derive the optimal time-slot allocations scheme over different video layers, which minimizes the total time-slot resource consumption while guaranteeing the specified delay and loss QoS constraints.

4. We propose an adaptive hybrid automatic repeat request-forward error correction (ARQ-FEC) erasure correcting scheme for quality of service (QoS)-driven mobile multicast services over wireless networks. Our proposed scheme can dynamically construct the erasure-correction code based on the packet-loss statuses of multicast receivers, which can achieve high error-control efficiency for mobile multicast networks while imposing low error-control complexity and overhead for mobile multicast networks.

In the second part of the dissertation:

1. We formulate and solve the sum effective capacity maximization problem in

downlink multiuser transmissions with independent message subject to the proportional fairness across different mobile users, where the effective capacity of a user represents the delay-QoS-constrained sustainable traffic loads. Our derived schemes can efficiently support downlink communications where coexisting mobile users have different statistical delay QoS requirements.

2. We propose the statistical delay-QoS-aware base station selection framework for distributed MIMO systems, which aims at minimizing the BS usage to reduce the complexity and interfering range caused by distributed MIMO transmissions. We develop the corresponding adaptive BS selection schemes for the single-user and multi-user scenarios, respectively. Our derived results can effectively satisfy diverse statistical delay QoS requirements of mobile users, while significantly decreasing the BS usage and the interfering range as compared to the fixed BS selection schemes.

3. We develop the queue-aware spectrum sensing scheme with statistical delay-QoS provisionings for secondary users in CRNs. Different from traditional energy-detection based spectrum-sensing schemes, where the threshold to detect the occupancy of the spectrum is fixed, our proposed energy-detection scheme uses a dynamic threshold varying with the queue-length of the SUs. Under the same interfering constraints to the PUs and the same statistical delay-QoS requirements, our proposed scheme can support higher traffic loads as compared to the traditional energy-detection based spectrum sensing schemes.

C. Outline of the Dissertation

The reminder of this dissertation is organized as follows.

In Chapter II, we give an introduction to the theory of statistical delay-QoS

guarantees and the dual concepts of effective capacity and effective bandwidth, which serves as the foundation theory for our works in Chapters IV, V, VII, VIII, and IX.

In Chapter III, we propose the rate-adaptation schemes for non-real-time multicasting over fading channels for mobile wireless networks. Specifically, the proposed schemes aim at achieving high system *goodput* for the entire multicast group. We first derive the optimal time-sharing (OPTS) rate policy for scenarios with independent and identically distributed (i.i.d.) channel fading across different multicast receivers. We also derive the optimal policy for two-receiver case with non.-i.i.d. channel distributions. Then, taking into account the practical considerations for more realistic systems, where the statistical information on fading channels is not available and the channel conditions are not i.i.d. over mobile users, we develop the *sub-grouping* (SG) based suboptimal rate-control policy. Chapter III is in part a reprint of the material in papers [54, 55].

In Chapter IV, applying the theory of statistical-delay QoS, we propose an *effective-capacity* optimization framework to develop the multicast rate-adaptation schemes over broadcast fading channels with the guaranteed statistical-delay QoS and loss-rate QoS. In particular, we employ the time-sharing (TS) and superposition-coding (SPC) techniques, respectively, to adapt the multicast transmission rates to the heterogeneous channel qualities across multicast receivers at each time instant. We further develop a queue-management scheme, called the *pre-drop* algorithm, and incorporate it with our optimization framework to implement the more efficient QoS-driven wireless multicast. Under our proposed framework, we derive the optimal TS-based and SPC-based adaptive multicast policies, respectively. Chapter IV is in part a reprint of the material in papers [56–58].

In Chapter V, we propose an efficient framework to model the statistical delay QoS guarantees, in terms of QoS exponent, effective bandwidth/capacity, and delay-

bound violation probability, for multi-layer video transmissions over wireless fading channels. We develop a set of optimal adaptive transmission schemes to minimize the resource consumption while satisfying the diverse delay-QoS requirements under various scenarios, including video unicast/multicast with and/or without loss tolerance. Chapter V is in part a reprint of the material in paper [59].

In Chapter VI, we propose an adaptive hybrid automatic repeat request-forward error correction (ARQ-FEC) erasure correcting scheme for quality of service (QoS)-driven mobile multicast services over wireless networks. Specifically, we propose the *non-uniformed* adaptive coding structures to achieve high error-control efficiency. Furthermore, we develop a loss covering strategy to determine the balanced error-control redundancy in each adaptation step and derive the corresponding incremental error-control redundancy as a function of the packet-loss level. Using the proposed two-dimensional adaptive error-control scheme, we design an efficient hybrid ARQ-FEC protocol for mobile multicast services with diverse loss-rate QoS requirements. Chapter VI is in part a reprint of the material in paper [60].

In Chapter VII, we propose the adaptive power and time-resource allocation schemes for multiuser downlink quality-of-service (QoS) provisionings over broadcast fading channels in cellular wireless networks. Subject to the *proportional-effective-capacity* constraint and the diverse *statistical* delay-QoS requirements over different downlink users, we formulate the *sum effective capacity* maximization problem via channel-aware power and time-slot allocation. Simulation results show the performance gain as compared to the suboptimal allocation schemes and the impact of delay-QoS requirements on resource allocations. Chapter VII is in part a reprint of the material in paper [61].

In Chapter VIII, we propose the QoS-aware BS-selection schemes for the *distributed wireless MIMO links*, which aim at minimizing the BS usages, while satisfying

diverse statistical delay-QoS constraints characterized by the delay-bound violation probability and the effective capacity technique. We develop the efficient scheme for both single-user case and multi-user case, which are demonstrated to be capable of significantly decreasing the BS-usage and the interfering range. Chapter VIII is in part a reprint of the material in paper [62].

In Chapter IX, we propose the queue-aware spectrum sensing schemes for interference-constrained opportunistic transmissions of secondary users (SUs) in cognitive radio networks. It is shown that under the specified statistical QoS requirements and interference constraints, our proposed schemes can support higher data traffic loads for SUs than the traditional energy-detection based scheme. Chapter IX is in part a reprint of the material in paper [63].

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

CHAPTER II

STATISTICAL DELAY-QOS GUARANTEES: EFFECTIVE BANDWIDTH AND EFFECTIVE CAPACITY

A. Queue-Length Distribution

The theory on statistical delay-QoS guarantees [2, 3, 8], featured with queue-length-bound violation probability, delay-bound violation probability, and buffer-overflow probability, provides a powerful approach in analyzing the queueing behavior for time-varying arrival and/or service processes. Specifically, consider a stable queueing system with the stationary and ergodic arrival and service processes. Asymptotic analyses [2] based on Large Deviation Principle show that with sufficient conditions, the queue length process $Q[t]$ converges to a random variable $Q[\infty]$ in distribution such that

$$-\lim_{Q_{\rm th}\to\infty}\frac{\log\left(\Pr\left\{Q[\infty]>Q_{\rm th}\right\}\right)}{Q_{\rm th}}=\theta \tag{2.1}$$

for a certain $\theta > 0$, where $Q_{\rm th}$ is the queue-length bound. Moreover, the queue-length bound violation probability can be approximated by

$$\Pr\{Q>Q_{\rm th}\}\approx e^{-\theta Q_{\rm th}}, \tag{2.2}$$

where we remove the index $[t]$ for $Q[t]$ to simplify notations. In the above two equations, θ is called *QoS exponent*.

B. Effective Bandwidth Versus Effective Capacity

Effective bandwidth [3] and effective capacity [8] are a pair of dual concepts. Given a stationary discrete-time arrival-rate process $A[t]$, effective bandwidth of $A[t]$, denoted

by $\mathcal{A}(\theta)$ (nats/frame), is defined as the *minimum constant service rate* required to guarantee a specified QoS exponent θ, i.e., Eqs. (2.1)-(2.2) are satisfied for the given θ. In contrast, given a stationary discrete-time service-rate process $C[t]$, effective capacity of $C[t]$, denoted by $\mathcal{C}(\theta)$ (nats/frame), is defined as the maximum *constant arrival rate* which can be supported by $C[t]$ subject to the specified QoS exponent θ. Moreover, effective bandwidth [3] and effective capacity [8] can be expressed by

$$\mathcal{A}(\theta) = \lim_{k\to\infty} \frac{1}{\theta k} \log\left(\mathbb{E}\left\{e^{\theta S_A[t]}\right\}\right) \tag{2.3}$$

and

$$\mathcal{C}(\theta) = \lim_{k\to\infty} -\frac{1}{\theta k} \log\left(\mathbb{E}\left\{e^{-\theta S_C[t]}\right\}\right), \tag{2.4}$$

respectively, where $\mathbb{E}\{\cdot\}$ denotes the expectation and

$$\begin{cases} S_A[t] \triangleq \sum_{i=1}^{k} A[i], \\ S_C[t] \triangleq \sum_{i=1}^{k} C[i]. \end{cases} \tag{2.5}$$

Note that if $A[t]$ is equal to a constant $\overline{A}$ over all t, we have $\mathcal{A}(\theta) = \overline{A}$ for all θ. Similarly, if $C[t]$ is equal to a constant $\overline{C}$, we have $\mathcal{C}(\theta) = \overline{C}$ for all θ.

Now, consider a queueing system with arrival-rate process $A[t]$ and service-rate process $C[t]$, which are both time-varying. If Eq. (2.1) holds for $\theta = \theta^*$, we have [3,53]

$$\mathcal{A}(\theta^*) = \mathcal{C}(\theta^*). \tag{2.6}$$

Eq. (2.6) implies a convenient approach to design service process to meet statistical delay-QoS requirements. Assume that the certain targeted queue-length-bound Q_{th} and violation probability P_{th} are specified for a queueing system, where we know the properties of the arrival process $A[t]$ and need to design the service process to guarantee the specified Q_{th} and P_{th}. First, plugging $\Pr\{Q > Q_{\text{th}}\} = P_{\text{th}}$ and $\theta = \theta^*$

into Eq. (2.2) and solving for the targeted θ^*, we get

$$\theta^* = -\frac{1}{Q_{\text{th}}} \log(P_{\text{th}}). \tag{2.7}$$

Second, we can determine the effective bandwidth $\mathcal{A}(\theta^*)$ based on the properties of $A[t]$. Finally, the criterion for designing the service rate $C[t]$ to meet the specified queue-length-bound violation probability is to guarantee that the effective capacity $\mathcal{C}(\theta^*)$ is equal to $\mathcal{A}(\theta^*)$.

We can see that the above approach first studies the arrival process and the service process separately, and then connect them through θ^*, which leads to equal effective capacity and effective bandwidth. Since our objective in this dissertation is to design service process through adaptive resource allocation over wireless channels, we will mainly focus on the effective capacity, which characterizes the capability of wireless channel to support data traffic with guaranteed QoS satisfaction.

C. Properties of Effective Capacity

Based on previous discussions, we can see that the targeted QoS exponent reflect delay QoS requirements. The larger θ corresponds to the more stringent QoS requirement, while the smaller θ imposes the looser delay constraint. Then, we can use QoS exponent θ as the metric to characterize the statistical delay-QoS requirements. Moreover, to comprehensively understand the effective capacity, we summarize a number of main properties of effective capacity [53] as follows. 1) For a given service-rate process, the effective capacity $\mathcal{C}(\theta)$ is a monotonically decreasing function of θ, which implies that a more stringent QoS requirement results in a lower supportable services rate. 2) As θ approaches 0, the effective capacity converges to the average throughput $\mathbb{E}\{C[t]\}$. This case corresponds to the scenario where arbitrarily long delay can be

tolerated. 3) In contrast, when θ approaches ∞ the effective capacity degrades to the minimum service rate over all time. This case is associated with the scenario which cannot tolerate any delay.

Furthermore, when the service-rate process $C[t]$ is time-uncorrelated, the effective capacity expression can be converted to

$$\mathcal{C}(\theta) = -\frac{1}{\theta} \log \left(\mathbb{E} \left\{ e^{-\theta C[t]} \right\} \right). \tag{2.8}$$

Note that in wireless channels, block-fading model is widely used to characterize the variation of wireless channels, where the channel state does not change within a time frame with fixed length, but varies independently from one frame to the other frame. Thus Eq. (2.8) provides great convenience to design the delay-QoS-driven service process (equivalently, adaptive resource allocation) over wireless channels. In this dissertation, we mainly focus on the block-fading channel model. For the scenario with time-correlated channel, please refer to [19, 53].

D. Delay-Bound Violation Probability

Previous sections show that QoS exponent θ plays an important role in QoS characterization and provisioning. It can be also used to characterize the delay-bound violation probability. Specifically, for effective capacity and effective bandwidth scenarios, the delay-bound violation probability can be approximated [3, 8] as:

$$\Pr\{D > D_{\text{th}}\} \approx e^{-\theta \mathcal{C}(\theta) D_{\text{th}}}, \text{ for effective capacity;} \tag{2.9}$$

and

$$\Pr\{D > D_{\text{th}}\} \approx e^{-\theta \mathcal{A}(\theta) D_{\text{th}}}, \text{ for effective bandwidth,} \tag{2.10}$$

where D denotes the queuing delay and $D_{\rm th}$ is the delay bound. Note that the delay in wireless transmissions may result from multiple factors such as propagation, queueing, and decoding. In this dissertation, we mainly focus on queueing delay, because the wireless channel is often the bottleneck for data transmissions.

Equations (2.2), (2.9), and (2.10) are good approximations for relatively large $Q_{\rm th}$ and $D_{\rm th}$ as shown in [8,64]. When $Q_{\rm th}$ and $D_{\rm th}$ are relatively small, the more accurate approximations expressions than Eqs. (2.2), (2.9), and (2.10) are given in [8, 64] as follows:

$$\begin{cases} \Pr\{Q > Q_{\rm th}\} & \approx & \varrho e^{-\theta Q_{\rm th}}; \\ \Pr\{D > D_{\rm th}\} & \approx & \varrho e^{-\theta \mathcal{C}(\theta) D_{\rm th}}; \\ \Pr\{D > D_{\rm th}\} & \approx & \varrho e^{-\theta \mathcal{A}(\theta) D_{\rm th}}, \end{cases}$$

where ϱ denotes the probability that the queue is nonempty. These approximations are upper-bounded by the corresponding approximations given in Eqs. (2.2)-(2.10). Thus, directly using Eqs. (2.2)-(2.10) for the system design often guarantees more stringent QoS than the specified requirements. Moreover, for typical mobile services with stringent delay-QoS requirements such as wireless video transmissions, the delay bound $D_{\rm th}$ is typically hundreds of milliseconds (ms), which are thus much larger than the adaptive-transmission period scale (e.g., the physical-layer time-frame length) of the wireless transmission system, where the adaptive-transmission period typically varies from a few milliseconds (ms) to tens of milliseconds (ms). Therefore, Eqs. (2.2)-(2.10) are good approximating expressions in designing efficient wireless video-transmission schemes with statistical QoS guarantees.

CHAPTER III

ADAPTIVE RATE ALLOCATION FOR AVERAGE GOODPUT OPTIMIZATION OF NON-REAL-TIME MULTICAST SERVICES OVER FADING CHANNELS

A. Introduction

As more and more broadband wireless network services are getting widely used, they impose the great motivations and challenges in developing the new and highly-efficient wireless communications paradigms with the limited wireless resources. One of the emerging highly efficient wireless communication techniques is the mobile multicast, which disseminates the common data/infomation to multiple location-independent receivers over broadcast wireless channels. As a result, in wireless networks mobile multicast has received a great deal of research attention [25–27, 31, 65]. In addition, due to its wide spectrum of applications, including wireless data downloading, highway mobile traffic monitoring, air traffic control, and remote teleconferencing, mobile multicast has been already considered as one of the key techniques in the current/future wireless communications systems such as 3G [25] and 802.16x [26] networks.

In wireless communication environments, the time-varying fading nature of physical-layer channels has significant impact on supporting the wideband services. However, the physical-layer techniques to improve performances for mobile multicast, despite their vital importance, have been neither well understood nor thoroughly studied. For time-varying fading channels, the sender can usually adapt the data rate according to the variation of channel qualities to implement efficient and high-rate transmission by using, e.g., the *adaptive modulation and coding* (AMC) techniques [66]. However, the diverse fading properties across different multicast receivers prevent the adaptive

transmission from being efficiently implemented. Consider a multicast session in a cellular-structured network, where the sender transmits a *single* data stream to all mobile multicast receivers through broadcast wireless channels. Since all multicast receivers must receive the same data signal, a straightforward strategy is to determine the transmission rate based only on the worst-case instantaneous channel quality among all multicast receivers, which is called the instantaneous worst-case dominating strategy. As a result, the achieved throughput is typically very low due to the heterogeneous fading of different send-receiver pairs, especially when the multicast group size gets large [27]. Thus, to achieve the higher multicast transmission rate, the multicast sender needs to apply the better transmission-rate control policies than the instantaneous worst-case dominating strategy. While this causes data losses for receivers with poorer channel qualities, we can apply sophisticated erasure-correcting techniques [67,68] at the upper protocol layers to recover the lost data packets. However, if the losses level is too high, the effective/useful data rate achieved by multicast receivers will also be very low, which leads to poor transmission efficiency and cannot support wideband services for mobile users. Consequently, in order to provide high-speed wideband services for multicast receivers, it is critically important to optimize the effective/userful data rate achieved by multicast receivers. Therefore, in this chapter we characterize the effective/userful data rate achievable for the entire multicast group by the *average multicast goodput*, and build up the framework to optimize the physical-layer average multicast goodput via transmission rate adaptation.

In [31], the authors investigated the scaling law of throughput-delay tradeoff for cellular multicast under a cross-layer structure. For single multicast session with *independent and identically distributed* (i.i.d.) fading across different receivers, the static rate-scheduling algorithm was studied, where a *fixed* portion of receivers can decode the transmitted data in each fading state. In their schemes, the transmission

rate is determined by the signal-to-noise ratio (SNR) which is in a *fixed* quantile of the ordered sequence of the instantaneous SNR's received from all receivers. It can be shown that their static rate control scheme is less efficient than the dynamic adaptive rate control algorithms in optimizing the average multicast goodput. In addition, their schemes did not consider the non-i.i.d. receiving channel fading scenarios, which are more general but more challenging in deriving the optimal rate control for multicast receivers.

It should be noted that although many literatures on information theory have studied various schemes to efficiently communicate through broadcast fading channels, they do not offer any insightful guidance on rate adaptation for mobile multicast. For instance, while a great deal of researches considered scenarios of distributing independent information to different receivers, e.g., [13, 15], mobile multicast is a special service to transmit the *common* information to multiple different mobile receivers. Authors of [11, 12] studied resource allocation schemes for scenarios where both common and different information are simultaneously transmitted to multiple receivers over Gaussian broadcast channels. However, the rate for the common data part is still controlled by the instantaneous worst-case channel quality among all receivers, resulting in very low average throughput/goodput over fading channels.

To overcome the above problems, we propose the adaptive rate control schemes for mobile multicast. Specifically, we focus on non-real-time multicast services and our proposed schemes aim at achieving high system *goodput*. First, we formulate the time-sharing based goodput-optimization problem over block-fading channels to derive the optimal rate adaptation policies. Second, under the formulated optimization problem, we solve for the optimal time-sharing (OPTS) policy under the i.i.d. fading environments across multicast receivers. Third, by developing the *SNR-plane partition* technique, we derive the OPTS policy for two-receiver cases with the non-i.i.d.

fading channels. Finally, taking into account the practical considerations for more realistic systems, where the statistical information of fading channels is not available, we develop a *sub-grouping* (SG) based suboptimal rate-control policy.

The rest of this chapter is organized as follows. Section B describes the system model. Section C formulates the time-sharing based multicast goodput-optimization problem. Section D derives the optimal rate-control policies for the scenarios with i.i.d. fading conditions over different receiving channels. Section E develops the OPTS policy for two-receiver cases with non-i.i.d. fading channels towards the multicast receivers. Section F proposes the SG based suboptimal rate-control policy without the knowledge of channel distributions. Section G evaluates the performance through numerical and simulation analyses. The chapter concludes with Section H.

B. System Model

1. System Description

We consider a discrete-time mobile multicast scenario in wireless networks as shown in Fig. 1. The multicast sender transmits a *single* data stream to all N (multicast group size) mobile multicast receivers through wireless broadcast fading channels. The sender uses single antenna to transmit the multicast signals, and each receiver also uses single antenna to receive these signals. We use the flat-fading channel models, and thus the signal transmitted can be expressed by $y_n = h_n x + \nu_n$, where h_n is the channel gain between the sender and the nth receiver, x is the complex multicast signal with spectral bandwidth B and *constant* power P, y_n denotes the signal received by the nth receiver, and ν_n's are independent complex additive white Gaussian noise (AWGN) with power spectral density $N_0/2$ per dimension. Then, the channel-state information (CSI) received at the multicast receivers can be characterized by the

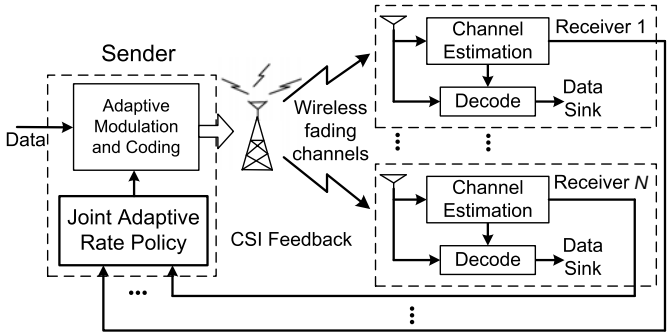


Fig. 1. System model for mobile multicast over fading channels.

received instantaneous SNR, which is defined by $\gamma_n \triangleq P|h_n|^2/(N_0 B)$, $n = 1, 2, \ldots, N$, where B denotes the system spectral-bandwidth. We also model the discrete-time fading channel by an ergodic and stationary block-fading process, which is widely used for wireless channels. In particular, $\{\gamma_n\}_{n=1}^N$ are invariant within a time frame of length T, but vary independently from time frame to time frame. The time-frame length T is sufficiently long such that the information-theoretic assumption of infinitely-long code-block length is meaningful [69]. Moreover, the CSI can be perfectly estimated at each receiver and then be reliably fed back to the sender without delay through the dedicated control channels.

We use $f_{\Gamma_n}(\gamma)$ and $F_{\Gamma_n}(\gamma)$ to represent the probability density function (pdf) and cumulative distribution function (CDF) of γ_n, respectively. The joint pdf and CDF of $\{\gamma_n\}_{n=1}^N$ are denoted by $f_{\mathbf{\Gamma}}(\boldsymbol{\gamma})$ and $F_{\mathbf{\Gamma}}(\boldsymbol{\gamma})$, respectively, where $\boldsymbol{\gamma} \triangleq (\gamma_1, \gamma_2, \ldots, \gamma_N)^\tau$ represents a particular fading state and $(\cdot)^\tau$ denotes transpose. In addition, we sort the SNR vector $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \ldots, \gamma_N)^\tau$ and then denote the ordered instantaneous SNR vector by $\widehat{\boldsymbol{\gamma}} \triangleq (\widehat{\gamma}_1, \widehat{\gamma}_2, \ldots, \widehat{\gamma}_N)^\tau$, where $\widehat{\gamma}_1 \geq \widehat{\gamma}_2 \geq \ldots \geq \widehat{\gamma}_N$. We assume that $F_{\mathbf{\Gamma}}(\boldsymbol{\gamma})$ is continuous over $(\mathbb{R}^+)^N$, where $\mathbb{R}^+$ denotes the domain of nonnegative real numbers. Note that the most widely used fading models such as Rayleigh, Rician, and Nakagami distributions [70] satisfy this assumption. In this chapter, we use Rayleigh distribu-

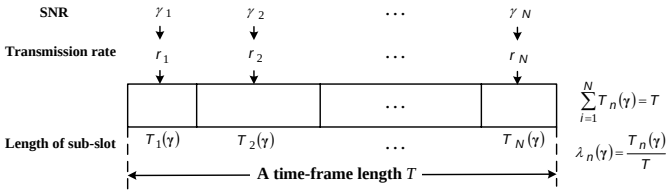


Fig. 2. Time-sharing based rate allocation.

tion [70] as the typical example for numerical and simulation analyses. For Rayleigh fading channels, we have $f_{\Gamma_n}(\gamma) = \exp(-\gamma/\overline{\gamma}_n)/\overline{\gamma}_n$ and $F_{\Gamma_n}(\gamma) = 1 - \exp(-\gamma/\overline{\gamma}_n)$, where $\overline{\gamma}_n$ is the mean of γ_n.

2. The Time-Sharing Based Rate-Adaptation Policy

Given the CSI vector $\boldsymbol{\gamma}$, the sender selects the transmission rate, denoted by r, which is selected from a discrete set $\mathcal{R}(\boldsymbol{\gamma}) \triangleq \{r_1, r_2, \ldots, r_N\}$, where $r_n = B\log(1+\gamma_n)$ is the Shannon capacity for the given SNR γ_n. In order to build up a more general framework, we apply the *time-sharing* strategy [71] among these N possible transmission rates of $\mathcal{R}(\boldsymbol{\gamma}) = \{r_1, r_2, \ldots, r_N\}$. In particular, as shown in Fig. 2, the sender divides each time-frame into N sub-slots with lengths of $T_1(\boldsymbol{\gamma}), T_2(\boldsymbol{\gamma}), \ldots, T_N(\boldsymbol{\gamma})$, corresponding to N multicast receivers, and $\sum_{i=1}^{N} T_i(\boldsymbol{\gamma}) = T$. The transmission rate in the nth sub-slot is set equal to $r = r_n$. Accordingly, we denote the time proportion of the nth sub-slot by $\lambda_n(\boldsymbol{\gamma})$, $n = 1, 2, \ldots, N$, where $\lambda_n(\boldsymbol{\gamma}) \triangleq T_n(\boldsymbol{\gamma})/T$. Since $\sum_{n=1}^{N} T_n(\boldsymbol{\gamma}) = T$, we have $\sum_{n=1}^{N} \lambda_n(\boldsymbol{\gamma}) = 1$. Then, we can control time proportions of the N sub-slots to implement the time-sharing strategy, and characterize the time-sharing policy by a vector function

$$\boldsymbol{\lambda}(\boldsymbol{\gamma}) \triangleq \big(\lambda_1(\boldsymbol{\gamma}), \lambda_2(\boldsymbol{\gamma}), \ldots, \lambda_N(\boldsymbol{\gamma})\big)^{\tau}. \tag{3.1}$$

It is worth noting that through the time-sharing strategy, we can *continuously* control the average multicast transmission rate within a time frame, although the instantaneous transmission rate can be selected only from the discrete set $\mathcal{R}(\boldsymbol{\gamma})$. For convenience of presentation, we define $\widehat{r}_i \triangleq B\log(1+\widehat{\gamma}_i)$ with $i = 1, 2, \ldots, N$, and denote the time proportion of the sub-slot having multicast transmission rate $r = \widehat{r}_i$ by $\widehat{\lambda}_i(\boldsymbol{\gamma})$. Thus, we can also characterize the time-sharing rate-control policy by a new vector defined by $\widehat{\boldsymbol{\lambda}}(\boldsymbol{\gamma}) \triangleq \big(\widehat{\lambda}_1(\boldsymbol{\gamma}), \widehat{\lambda}_2(\boldsymbol{\gamma}), \ldots, \widehat{\lambda}_N(\boldsymbol{\gamma})\big)^\tau$, which corresponds to the ordered SNR vector $\widehat{\boldsymbol{\gamma}} = (\widehat{\gamma}_1, \widehat{\gamma}_2, \ldots, \widehat{\gamma}_N)^\tau$.

C. Formulating Optimization Problem for Maximizing Average Goodput

Under the time-sharing policies, we can formulate an optimization problem to derive the optimal rate-control policies which can maximize the average multicast goodput at the *physical* layer with the CSI feedback $\boldsymbol{\gamma}$. We assume that the capacity-achieving codes are used. Thus, if the Shannon capacity for γ_n is higher than or equal to the transmission rate r, the nth receiver can correctly decode the received signal in a sub-slot; otherwise, the nth receiver cannot correctly decode the received signals. Given that the selected multicast transmission rate is r and the nth receiver's SNR is γ_n, the instantaneous multicast rate achieved by the nth receiver, denoted by $c(\gamma_n, r)$, is expressed as

$$c(\gamma_n, r) = \begin{cases} r, & \text{if } B\log(1+\gamma_n) \geq r; \\ 0, & \text{if } B\log(1+\gamma_n) < r. \end{cases} \tag{3.2}$$

We then define the average multicast goodput as follows.

Definition 1. *The nth receiver's achieved average rate, denoted by* $\overline{R}_n\big(\boldsymbol{\lambda}(\boldsymbol{\gamma})\big)$*, is the expectation of the instantaneous rate achieved by the nth receiver over all fading*

states, which is determined by

$$\overline{R}_n\big(\boldsymbol{\lambda}(\boldsymbol{\gamma})\big) \triangleq \mathbb{E}\{\mathbf{c}_n(\boldsymbol{\gamma})^\tau \boldsymbol{\lambda}(\boldsymbol{\gamma})\} = \int_0^\infty \int_0^\infty \cdots \int_0^\infty \mathbf{c}_n(\boldsymbol{\gamma})^\tau \boldsymbol{\lambda}(\boldsymbol{\gamma}) f_{\boldsymbol{\Gamma}}(\boldsymbol{\gamma}) d\gamma_1 d\gamma_2 \cdots d\gamma_N, \quad (3.3)$$

where $\mathbf{c}_n(\boldsymbol{\gamma}) \triangleq \big(c(\gamma_n, r_1), c(\gamma_n, r_2), \ldots, c(\gamma_n, r_N)\big)^\tau$, $r_n = B\log(1+\gamma_n)$, *and* $\mathbb{E}\{\cdot\}$ *denotes expectation.*

Definition 2. *The* average multicast goodput *of a multicast group under a time-sharing policy* $\boldsymbol{\lambda}(\boldsymbol{\gamma})$, *denoted by* $\overline{R}^{\text{gp}}\big(\boldsymbol{\lambda}(\boldsymbol{\gamma})\big)$, *or* $\overline{R}^{\text{gp}}$ *in short, is defined as the* minimum $\overline{R}_n\big(\boldsymbol{\lambda}(\boldsymbol{\gamma})\big)$ *among all multicast receivers as follows:*

$$\overline{R}^{\text{gp}} = \min_{1\le n\le N} \Big\{\overline{R}_n\big(\boldsymbol{\lambda}(\boldsymbol{\gamma})\big)\Big\}, \quad (3.4)$$

where $\overline{R}_n\big(\boldsymbol{\lambda}(\boldsymbol{\gamma})\big)$ *is the nth receiver's achieved average rate given by Eq. (3.3). Correspondingly,* $\overline{R}^{\text{gp}}/B$ *is called the normalized average multicast goodput.*

We can then derive the optimal rate-control policies by solving the following goodput-optimization problem:

$$\begin{aligned} &\boldsymbol{\lambda}^*(\boldsymbol{\gamma}) = \arg\max_{\boldsymbol{\lambda}(\boldsymbol{\gamma})} \Big\{\overline{R}^{\text{gp}}\Big\} = \arg\max_{\boldsymbol{\lambda}(\boldsymbol{\gamma})} \left\{\min_{1\le n\le N} \Big\{\overline{R}_n\big(\boldsymbol{\lambda}(\boldsymbol{\gamma})\big)\Big\}\right\}, \\ &\text{s.t.:} \quad \lambda_i(\boldsymbol{\gamma}) \ge 0, 1 \le i \le N; \\ &\qquad \sum_{i=1}^{N} \lambda_i(\boldsymbol{\gamma}) = 1. \end{aligned} \quad (3.5)$$

Since our proposed framework aims at optimizing the average multicast goodput other than guaranteeing the transient performance, this framework is particularly suitable for data services such as wireless data downloading. While the average multicast goodput defines the data rate achievable for all multicast receivers at the physical layer, different multicast receivers usually lose different part of the transmitted signals due to the highly-varying and heterogenous wireless channels. To provide reliability for all multicast receivers, we can apply the sophisticated erasure-correcting codes

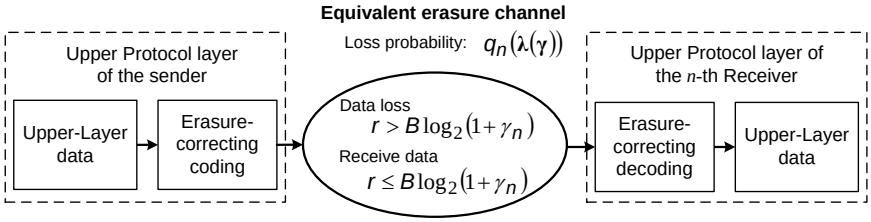


Fig. 3. The equivalent erasure channel at the upper protocol layer.

at the upper protocol layer. Specifically, consider an example of multicasting a data file with very large size. The original data file is first encoded by using the erasure-correcting codes. Then, the sender passes the encoded data to the physical layer and multicasts them to all receivers, where the amount of the encoded data transmitted in each frame is determined by the rate-adaptation policy $\boldsymbol{\lambda}(\boldsymbol{\gamma})$. The physical-layer signal transmission model given by Eq. (3.2) can be viewed as an equivalent erasure channel at the upper protocol layer, as illustrated in Fig. 3. The average loss probability of the erasure channel of the nth receiver, denoted by $\rho_n(\boldsymbol{\lambda}(\boldsymbol{\gamma}))$, is determined by the average multicast goodput and average transmission rate as follows:

$$\rho_n(\boldsymbol{\lambda}(\boldsymbol{\gamma})) = \frac{\mathbb{E}\left\{\sum_{n=1}^{N} r_n \lambda_n(\boldsymbol{\gamma})\right\} - \overline{R}^{\mathrm{gp}}(\boldsymbol{\lambda}(\boldsymbol{\gamma}))}{\mathbb{E}\left\{\sum_{n=1}^{N} r_n \lambda_n(\boldsymbol{\gamma})\right\}}, \tag{3.6}$$

where $\mathbb{E}\left\{\sum_{n=1}^{N} r_n \lambda_n(\boldsymbol{\gamma})\right\}$ is the average transmission rate. Accordingly, the capacity for the above equivalent erasure channel, denoted by C_n, $n = 1, 2, \ldots, N$, is determined by

$$C_n = \mathbb{E}\left\{\sum_{n=1}^{N} r_n \lambda_n(\boldsymbol{\gamma})\right\}\Big(1 - \rho_n(\boldsymbol{\lambda}(\boldsymbol{\gamma})\Big) = \overline{R}_n(\boldsymbol{\lambda}(\boldsymbol{\gamma})), \quad \text{nats/s}, \tag{3.7}$$

We assume that the capacity-achieving erasure-correcting code is used and the code rate is set equal to $\min_{1\le n\le N}\{1 - \rho_n(\boldsymbol{\lambda}(\boldsymbol{\gamma})\}$. Then, the information rate included in

the encoded data is equal to

$$\min_{1\leq n\leq N}\{1-\rho_n(\boldsymbol{\lambda}(\boldsymbol{\gamma})\}\times\mathbb{E}\left\{\sum_{n=1}^{N}r_n\lambda_n(\boldsymbol{\gamma})\right\}=\overline{R}^{\mathrm{gp}}(\boldsymbol{\lambda}(\boldsymbol{\gamma})).$$

As a result, all receivers can successfully recover the original data file because their equivalent erasure channels' capacities C_n, $n=1,2,\ldots,N$, are all beyond the information rate included in the encoded data, which is also equal to the average multicast goodput $\overline{R}^{\mathrm{gp}}(\boldsymbol{\lambda}(\boldsymbol{\gamma}))$. Based on the above analyses, we can see that the encoding at high protocol layer is across many time frames, and thus this model is for the non-realtime data services. Correspondingly, we need to optimize the average multicast goodput, which can also minimize the average delay to download large-size data.

The above erasure-correcting model is also meaningful and can provide insightful guidance for realistic systems, because there are many near capacity-achieving codes with low implementation complexity. Fountain codes [68] are well-known as a type of powerful codes for reliable multicast transmissions with the following properties. (i) The encoding at upper protocol layer can be performed online in each time frame with low complexity. (ii) The maximum supportable error-control redundancy is virtually not upper-bounded. The sender can keep on generating encoded data packets as long as there are requests from receivers. (iii) How to encode is independent of the loss status and channel status. (iv) A multicast receiver can recover the original uncoded file with a certain number of correctly-received encoded data packets, regardless of which parts of the transmitted packets are received. (v) The amount of the encoded data required for successful file-recovering is only slightly higher than the size of the original file. With the above properties, we can see that the physical-layer rate adaptation can be effectively applied with the help from fountain codes. Properties (iv) and (v) show that fountain codes are a type of near capacity-achieving erasure-correcting codes. Thus, the average multicast goodput presents a tight upper-bound

for the data rate achievable for the entire multicast group in realistic systems. More importantly, property (ii) implies that how to construct fountain codes does not need the information of the loss probability $\rho_n(\boldsymbol{\lambda}(\boldsymbol{\gamma}))$ or channel status. As a result, the physical-layer rate adaptation and upper protocol layer encodings can be designed separately, and the physical layer only needs to tell the upper protocol layer how many encoded data are required in each frame. Therefore, we can focus on physical-layer rate adaptation while leaving the error control to upper protocol layer.

D. Optimal Rate-Control Policy for I.I.D. Fading across Multicast Receivers

1. The Optimal Time-Sharing Rate-Adaptation Policy in I.I.D. Fading Channels

We first focus on scenarios where the instantaneous SNR's $\{\gamma_n\}_{n=1}^N$ are i.i.d. in this section. The more practical but more complex non-i.i.d. fading scenarios will be studied in Sections E and F. Proposition 1 given below derives the OPTS policy which maximizes the average multicast goodput in the i.i.d. fading environments.

Proposition 1. *Assume that $\{\gamma_n\}_{n=1}^N$ are i.i.d. The optimal time-sharing policy $\boldsymbol{\lambda}^*(\boldsymbol{\gamma})$ to the goodput-optimization problem formulated by Eq. (3.5) is given in its equivalent form $\widehat{\boldsymbol{\lambda}}^*(\boldsymbol{\gamma})$ as follows:*

$$\widehat{\lambda}_i^*(\boldsymbol{\gamma}) = \begin{cases} 1, & \text{if } i = i^*; \\ 0, & \text{if } i \neq i^*, 1 \leq i \leq N, \end{cases} \tag{3.8}$$

where

$$i^* = \arg\max_{1\leq i\leq N} \left\{ B \sum_{\forall\, k, \gamma_k \geq \widehat{\gamma}_i} \log(1+\widehat{\gamma}_i) \right\} = \arg\max_{1\leq i\leq N} \left\{ iB\log(1+\widehat{\gamma}_i) \right\}. \tag{3.9}$$

If there exist multiple indices $i_1, i_2, \ldots, i_S$ (S is an integer, $1 \leq S \leq N$) such that $i_s B\log(1+\widehat{\gamma}_{i_s}) = \max_{1\leq i\leq N}\{iB\log(1+\widehat{\gamma}_i)\}$ holds for all $s = 1, 2, \ldots, S$, i^* is set to

be $i^* = \max\{i_1, i_2, \ldots, i_S\}$. *Also, the OPTS policy* $\boldsymbol{\lambda}^*(\boldsymbol{\gamma})$ *yields*

$$\overline{R}^{\text{gp}} = \overline{R}_1\big(\boldsymbol{\lambda}^*(\boldsymbol{\gamma})\big) = \overline{R}_2\big(\boldsymbol{\lambda}^*(\boldsymbol{\gamma})\big) = \cdots = \overline{R}_N\big(\boldsymbol{\lambda}^*(\boldsymbol{\gamma})\big). \tag{3.10}$$

Proof. The proof is provided in Appendix A. □

Based on Proposition 1, for a given $\boldsymbol{\gamma}$ the sender only uses the transmission rate $\widehat{r}_{i^*} = B\log(1+\widehat{\gamma}_{i^*})$ within the entire time frame. Also, different $\boldsymbol{\gamma}$'s with the same ordered version $\widehat{\boldsymbol{\gamma}}$ yield the same transmission rate. Since $\boldsymbol{\lambda}^*(\boldsymbol{\gamma})$ determines transmission rate only based on the ordered CSI vector $\widehat{\boldsymbol{\gamma}}$, $\boldsymbol{\lambda}^*(\boldsymbol{\gamma})$ benefits all multicast receivers *evenly* in the i.i.d. fading environments. Consequently, all multicast receivers achieve the same average rate as shown in Eq. (3.10). Furthermore, the objective function in Eq. (3.9) can be rewritten as

$$B \sum_{\forall\, k, \gamma_k \geq \widehat{\gamma}_i} \log(1+\widehat{\gamma}_i) = \sum_{k=1}^{N} c(\gamma_k, \widehat{r}_i), \tag{3.11}$$

which calculates the *sum of achieved rates over all multicast receivers*, or the sum of achieved rates in short, with $r = \widehat{r}_i$. Thus, Proposition 1 implies that maximizing the goodput $\overline{R}^{\text{gp}}$ is equivalent to maximizing the sum of achieved rate in each fading state *independently*.

For a given $\boldsymbol{\gamma}$, the optimal time-sharing rate-adaptation policy selects the possible transmission rate r only from the finite set $\mathcal{R}(\boldsymbol{\gamma}) \triangleq \{\widehat{r}_1, \widehat{r}_2, \ldots, \widehat{r}_N\}$. We show below that any transmission rate $r' \notin \mathcal{R}(\boldsymbol{\gamma})$ cannot maximize the sum of achieved rate. Define $\widehat{\gamma}_0 = \infty$ and $\widehat{\gamma}_{N+1} = 0$. Without loss of generality, given any transmission rate $r' \notin \mathcal{R}(\boldsymbol{\gamma})$ we can write r' as $r = B\log(1+\gamma')$ with a certain γ' satisfying $\widehat{\gamma}_j > \gamma' > \widehat{\gamma}_{j+1}$ for some j, $0 \leq j < N$. Then, we derive

$$\sum_{k=1}^{N} c\big(\gamma_k, B\log(1+\gamma')\big) = jB\log(1+\gamma') < jB\log(1+\widehat{\gamma}_j) = \sum_{k=1}^{N} c\big(\gamma_k, \widehat{r}_j\big) \tag{3.12}$$

which verifies the above claim. Since $r' \notin \mathcal{R}(\boldsymbol{\gamma})$ cannot maximize the sum of achieved rates in each fading state, it cannot maximize the average multicast goodput, either.

2. Performance Analyses and Case Investigations for I.I.D. Fading Channels

Under the OPTS policy given by Proposition 1, we first derive the closed-form expression of the average multicast goodput for the two-receiver case ($N = 2$) over i.i.d. Rayleigh fading channels. Plugging Eqs. (3.8) and (3.9) into Eqs. (3.3) and (3.4), we get

$$\overline{R}^{\mathrm{gp}} = Be^{\frac{2}{\overline{\gamma}}}\left[\mathcal{G}\left(0,\frac{2}{\overline{\gamma}}\right) - \frac{1}{\overline{\gamma}}H\left(1,\frac{1}{2},\frac{1}{\overline{\gamma}},\frac{1}{\overline{\gamma}}\right) - \frac{2}{\overline{\gamma}}H\left(1,2,\frac{1}{\overline{\gamma}},\frac{1}{\overline{\gamma}}\right)\right] + Be^{\frac{1}{\overline{\gamma}}}\mathcal{G}\left(0,\frac{1}{\overline{\gamma}}\right), \quad (3.13)$$

where $\overline{\gamma} = \overline{\gamma}_1 = \cdots = \overline{\gamma}_N$, $\mathcal{G}(w,z) \triangleq \int_z^\infty t^{w-1}e^{-t}dt$, $w, z \geq 0$, is the incomplete Gamma function [72], and $H(w,z,u,v) \triangleq \int_w^\infty e^{-ut-vt^z}\log t dt$ with $w, z, u, v \geq 0$ and $u + v > 0$. The numerical results of these two functions can be obtained by using many math softwares, e.g., Mathematica. Usually, there are no simple closed-form expressions for the cases with $N \geq 3$. In contrast, we focus on studying the monotonic property which is summarized by Proposition 2 as follows.

Proposition 2. *Assume that $\{\gamma_n\}_{n=1}^N$ are i.i.d. Under $\boldsymbol{\lambda}^*(\boldsymbol{\gamma})$ given in Proposition 1, the average multicast goodput $\overline{R}^{\mathrm{gp}}$ is a monotonic decreasing function of the multicast group size N.*

Proof. The proof is provided in Appendix B. □

Next, we will study the asymptotic performance of the average multicast goodput to exam the scalability of the OPTS rate-control policy.

Proposition 3. *Assume that $\{\gamma_n\}_{n=1}^N$ are i.i.d. and denote the CDF of each γ_n by $F_\Gamma(\gamma)$. As the multicast group size $N \to \infty$, the OPTS rate-control policy $\boldsymbol{\lambda}^*(\boldsymbol{\gamma})$ given*

in Proposition 1 converges to a constant rate policy $r = B\log(1+\gamma_\infty^*)$ *almost surely (a.s.), where*

$$\gamma_\infty^* \triangleq \arg\max_{0<\gamma<\infty}\{(1-F_\Gamma(\gamma))\log(1+\gamma)\} \tag{3.14}$$

Correspondingly, the average multicast goodput $\overline{R}^{\rm gp}$ *is determined by*

$$\lim_{N\to\infty}\overline{R}^{\rm gp} = B(1-F_\Gamma(\gamma_\infty^*))\log(1+\gamma_\infty^*), \qquad a.s. \tag{3.15}$$

Proof. Under the OPTS policy $\boldsymbol{\lambda}^*(\boldsymbol{\gamma})$, we have $r = \widehat{r}_{i^*} = B\log(1+\widehat{\gamma}_{i^*})$, where $\widehat{\gamma}_{i^*}$ is obtained through Eq. (3.9). Define $\mathcal{N}(\gamma)$ as the number of receivers with SNR higher than or equal to a given γ. Following Eqs. (3.9) and (3.12), we get

$$\begin{aligned}\widehat{\gamma}_{i^*} &= \arg\max_{\widehat{\gamma}_i, 1\le i\le N}\left\{\sum_{\forall\, k,\gamma_k\ge\widehat{\gamma}_i}\log(1+\widehat{\gamma}_i)\right\}\\ &= \arg\max_{0<\gamma<\infty}\left\{\frac{1}{N}\sum_{\forall\, k,\gamma_k\ge\gamma}\log(1+\gamma)\right\} = \arg\max_{0<\gamma<\infty}\left\{\frac{\mathcal{N}(\gamma)}{N}\log(1+\gamma)\right\}. \end{aligned}\tag{3.16}$$

Letting $N\to\infty$, we get

$$\lim_{N\to\infty}\frac{\mathcal{N}(\gamma)}{N} \overset{(a)}{=} \Pr\{\Gamma\ge\gamma\} = 1-F_\Gamma(\gamma) \qquad \text{a.s.} \tag{3.17}$$

where (a) holds according to [73, Proposititon 4.24].

Defining $\gamma_\infty^* \triangleq \arg\max_{0<\gamma<\infty}\{(1-F_\Gamma(\gamma))\log(1+\gamma)\}$ and plugging Eq. (3.17) into Eq. (3.16), we get $\lim_{N\to\infty}\widehat{\gamma}_{i^*} = \gamma_\infty^*$ a.s., also implying $\lim_{N\to\infty}\widehat{r}_{i^*} = B\log(1+\gamma_\infty^*)$ a.s. Moreover,

$$\lim_{N\to\infty}\overline{R}_n\big(\boldsymbol{\lambda}^*(\boldsymbol{\gamma})\big) = \int_{\gamma_\infty^*}^{\infty} B\log(1+\gamma_\infty^*)f_\Gamma(\gamma)d\gamma = B(1-F_\Gamma(\gamma_\infty^*))\log(1+\gamma_\infty^*), \quad \text{a.s.} \tag{3.18}$$

for all $1\le n\le N$. By using Eqs. (3.4) and (3.18), we obtain Eq. (3.15), which completes the proof of Proposition 3. □

The above propositions show that although the average multicast goodput is a monotonic decreasing function of multicast group size N, the degradation speed is limited. As $N \to \infty$, the goodput converges to a non-zero constant. In contrast, if the transmission rate is determined by the worst-case SNR among all receivers, the transmission rate and the goodput approaches to 0 with $N \to \infty$. This indicates that our derived OPTS policy has good scalability in avoiding goodput degradation. Moreover, we derive the closed-form asymptotic average multicast goodput in Rayleigh fading channels. Applying Eq. (3.14), we obtain $\gamma_{\infty}^{*} = \max_{0<\gamma<\infty}\left\{\exp\left(-\frac{\gamma}{\overline{\gamma}}\right)\log(1+\gamma)\right\} = \overline{\gamma}/W(\overline{\gamma}) - 1$, where for $z \neq 0$ $W(z)$ is the Lambert W-function [74].[1] Correspondingly, the transmission rate r converges to a constant rate $B\log(\overline{\gamma}/W(\overline{\gamma}))$, and then through Eq. (3.15), the average multicast goodput is determined by $\overline{R}^{\text{gp}} = B\exp\left(-1/W(\overline{\gamma}) + 1/\overline{\gamma}\right)(\log\overline{\gamma} - \log W(\overline{\gamma}))$.

E. Two-Receiver Cases with Non-I.I.D. Fading across Multicast Receivers

The rate adaptation policy given in Proposition 1 is not optimal policy when $\{\gamma_n\}_{n=1}^{N}$ are non-i.i.d. This is because the policy only maximizes sum of achieved rates over the entire multicast group. However, it cannot guarantee to maximize the average multicast goodput in non-i.i.d. fading environments and thus may cause severe goodput degradation. In the following, we focus on two-receiver scenarios with non-i.i.d. $\{\gamma_n\}_{n=1}^{N}$ and derive the corresponding OPTS policy. Note that in this section, we assume that the channel distribution information is available at the sender.

[1] Lambert W-function is the inverse function of $Z(w) = we^{w}$.

1. Problem Reformulation for Goodput Optimization with Non-I.I.D. Fading

For two-receiver cases, we can reformulate the goodput optimization problem given by Eq. (3.5) as follows:

$$\begin{aligned}
&\boldsymbol{\lambda}^*(\boldsymbol{\gamma}) = \arg\max_{\boldsymbol{\lambda}(\boldsymbol{\gamma})} \left\{\overline{R}_1\big(\boldsymbol{\lambda}(\boldsymbol{\gamma})\big)\right\},\\
&\text{s.t.: } \lambda_1(\boldsymbol{\gamma}), \lambda_2(\boldsymbol{\gamma}) \geq 0, \lambda_1(\boldsymbol{\gamma}) + \lambda_2(\boldsymbol{\gamma}) = 1,\\
&\qquad \overline{R}_1\big(\boldsymbol{\lambda}(\boldsymbol{\gamma})\big) = \overline{R}_2\big(\boldsymbol{\lambda}(\boldsymbol{\gamma})\big).
\end{aligned} \tag{3.19}$$

The above reformulation is obtained based on the following arguments. Consider any policy $\boldsymbol{\lambda}(\boldsymbol{\gamma})$ with unequal $\overline{R}_n\big(\boldsymbol{\lambda}(\boldsymbol{\gamma})\big)$'s. Without loss of generality, we assume $\overline{R}_1\big(\boldsymbol{\lambda}(\boldsymbol{\gamma})\big) < \overline{R}_2\big(\boldsymbol{\lambda}(\boldsymbol{\gamma})\big)$. Then from Eq. (3.4), the goodput $\overline{R}^{\text{gp}}$ of the multicast group is equal to $\overline{R}_1\big(\boldsymbol{\lambda}(\boldsymbol{\gamma})\big)$. Next, we reallocate time proportions by increasing $\lambda_1(\boldsymbol{\gamma})$ to improve $\overline{R}_1\big(\boldsymbol{\lambda}(\boldsymbol{\gamma})\big)$ such that $\overline{R}^{\text{gp}}$ can be also improved. Correspondingly, we define $\mu_n(\boldsymbol{\gamma})$, $n = 1, 2$, as

$$\mu_n(\boldsymbol{\gamma}) \triangleq c(\gamma_n, r_1) - c(\gamma_n, r_2) = c(\gamma_n, B\log(1+\gamma_1)) - c(\gamma_n, B\log(1+\gamma_2)), \tag{3.20}$$

which evaluates the increment of the nth receiver's instantaneously achieved rate if changing the transmission rate from r_2 to r_1 (see Eq. (3.2) for the definition of $c(\cdot,\cdot)$). In particular, we get

$$0 \leq \mu_1(\boldsymbol{\gamma}) = \begin{cases} B\log(1+\gamma_1) - B\log(1+\gamma_2), & \text{if } \gamma_1 \geq \gamma_2;\\ B\log(1+\gamma_1), & \text{if } \gamma_1 < \gamma_2, \end{cases} \tag{3.21}$$

$$0 \geq \mu_2(\boldsymbol{\gamma}) = \begin{cases} -B\log(1+\gamma_2), & \text{if } \gamma_1 > \gamma_2;\\ B\log(1+\gamma_1) - B\log(1+\gamma_2), & \text{if } \gamma_1 \leq \gamma_2. \end{cases} \tag{3.22}$$

Then in fading state $\boldsymbol{\gamma}$, if the increment of $\lambda_1(\boldsymbol{\gamma})$ is given by $\delta(\boldsymbol{\gamma}) > 0$, the improvement of $\overline{R}_1\big(\boldsymbol{\lambda}(\boldsymbol{\gamma})\big)$ and $\overline{R}_2\big(\boldsymbol{\lambda}(\boldsymbol{\gamma})\big)$, denoted by $I_1(\boldsymbol{\gamma})$ and $I_2(\boldsymbol{\gamma})$, respectively, can be expressed

as

$$\begin{cases} I_1(\boldsymbol{\gamma}) &= \delta(\boldsymbol{\gamma})\mu_1(\boldsymbol{\gamma})f_{\mathbf{\Gamma}}(\boldsymbol{\gamma})d\gamma_1 d\gamma_2 \ \geq \ 0; \\ I_2(\boldsymbol{\gamma}) &= \delta(\boldsymbol{\gamma})\mu_2(\boldsymbol{\gamma})f_{\mathbf{\Gamma}}(\boldsymbol{\gamma})d\gamma_1 d\gamma_2 \ \leq \ 0. \end{cases} \tag{3.23}$$

Equation (3.23) shows that under the time-proportion reallocation procedure, $\overline{R}_1\big(\boldsymbol{\lambda}(\boldsymbol{\gamma})\big)$ always increases while $\overline{R}_2\big(\boldsymbol{\lambda}(\boldsymbol{\gamma})\big)$ always decreases. Consequently, we can keep on enlarging $\delta(\boldsymbol{\gamma})$ to increase $\overline{R}_1\big(\boldsymbol{\lambda}(\boldsymbol{\gamma})\big)$ (or, equivalently, $\overline{R}^{\text{gp}}$) until $\overline{R}_1\big(\boldsymbol{\lambda}(\boldsymbol{\gamma})\big) = \overline{R}_2\big(\boldsymbol{\lambda}(\boldsymbol{\gamma})\big)$ is satisfied. Then, a new time-sharing policy with $\overline{R}_1\big(\boldsymbol{\lambda}(\boldsymbol{\gamma})\big) = \overline{R}_2\big(\boldsymbol{\lambda}(\boldsymbol{\gamma})\big)$ and a larger $\overline{R}^{\text{gp}}$ is obtained.

The existence of such a new policy satisfying $\overline{R}_1\big(\boldsymbol{\lambda}(\boldsymbol{\gamma})\big) = \overline{R}_2\big(\boldsymbol{\lambda}(\boldsymbol{\gamma})\big)$ can be explained as follows. If we eventually reallocate all time proportions to r_1, we get $\lambda_1(\boldsymbol{\gamma}) = 1$ for all $\boldsymbol{\gamma}$'s, which results in $\overline{R}_1\big(\boldsymbol{\lambda}(\boldsymbol{\gamma})\big) - \overline{R}_2\big(\boldsymbol{\lambda}(\boldsymbol{\gamma})\big) \geq 0$. On the other hand, we originally have $\overline{R}_1\big(\boldsymbol{\lambda}(\boldsymbol{\gamma})\big) - \overline{R}_2\big(\boldsymbol{\lambda}(\boldsymbol{\gamma})\big) \leq 0$ as assumed previously. Then, since $\delta(\boldsymbol{\gamma})$ can be tuned up continuously, such a new policy must exist.

2. Compensation Efficiency

The strategy to obtain the time-sharing rate-adaptation policy discussed in Section E-1 actually increases the goodput $\overline{R}^{\text{gp}}$ or, equivalently, the first receiver's achieved average rate $\overline{R}_1\big(\boldsymbol{\lambda}(\boldsymbol{\gamma})\big)$ at the cost of degrading $\overline{R}_2\big(\boldsymbol{\lambda}(\boldsymbol{\gamma})\big)$. Hence, in order to achieve a high $\overline{R}^{\text{gp}}$, we need to use the *minimum* degradation of $\overline{R}_2\big(\boldsymbol{\lambda}(\boldsymbol{\gamma})\big)$ to obtain the *maximum* improvement of $\overline{R}_1\big(\boldsymbol{\lambda}(\boldsymbol{\gamma})\big)$. To evaluate the efficiency of improving $\overline{R}_1\big(\boldsymbol{\lambda}(\boldsymbol{\gamma})\big)$ in each fading state and determine how to reallocate corresponding time proportions, we introduce a metric which is called *compensation efficiency* and denoted by $\eta(\boldsymbol{\gamma})$ (or η for simplicity). Specifically, we define $\eta(\boldsymbol{\gamma})$ as the ratio of the increment $I_1(\boldsymbol{\lambda})$ of $\overline{R}_1\big(\boldsymbol{\lambda}(\boldsymbol{\gamma})\big)$ to the decrement $(-I_2(\boldsymbol{\lambda}))$ of $\overline{R}_2\big(\boldsymbol{\lambda}(\boldsymbol{\gamma})\big)$ when the sender reallocates a

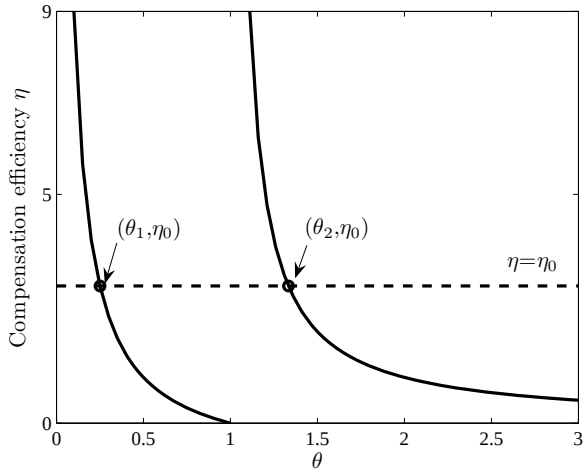


Fig. 4. Compensation efficiency η as a function of θ.

higher time proportion to transmission rate r_1. Using Eq. (3.23), we get

$$\eta(\boldsymbol{\gamma}) \triangleq \frac{I_1(\boldsymbol{\gamma})}{-I_2(\boldsymbol{\gamma})} = \frac{\mu_1(\boldsymbol{\gamma})}{-\mu_2(\boldsymbol{\gamma})} \geq 0, \quad \text{if } \gamma_1 \neq \gamma_2. \tag{3.24}$$

Note that for $\gamma_1 = \gamma_2$, we set $\eta(\boldsymbol{\gamma}) = 0$ without loss of generality.

To better understand the compensation efficiency, we introduce an auxiliary variable $\theta(\boldsymbol{\gamma})$ (or θ in short), which is defined as the ratio of instantaneous Shannon capacity between the two receivers and expressed by $\theta(\boldsymbol{\gamma}) \triangleq \log(1+\gamma_2)/\log(1+\gamma_1)$. Then using the definition of θ and Eqs. (3.21)-(3.22), we can express $\eta(\boldsymbol{\gamma})$ in terms of $\theta(\boldsymbol{\gamma})$ as

$$\eta(\boldsymbol{\gamma}) = \begin{cases} \frac{1-\theta(\boldsymbol{\gamma})}{\theta(\boldsymbol{\gamma})}, & \text{if } \theta(\boldsymbol{\gamma}) \in (0,1]; \\ \frac{1}{\theta(\boldsymbol{\gamma})-1}, & \text{if } \theta(\boldsymbol{\gamma}) \in (1,\infty). \end{cases} \tag{3.25}$$

We plot the compensation efficiency η as a function of θ in Fig. 4, which shows that η is a piece-wise decreasing function of θ. As θ increases from 0 to 1, η decreases from ∞

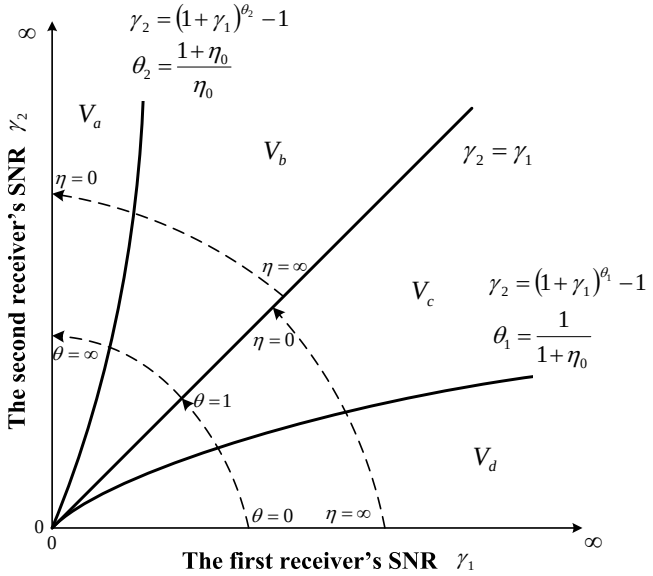


Fig. 5. SNR-plane partition for two-receiver cases.

to 0. At $\theta = 1$, there is a discontinuity for η with $\lim_{\theta\to 1^-} \eta = 0$ and $\lim_{\theta\to 1^+} \eta = \infty$.[2] When θ keeps increasing from 1 to ∞, η again decreases from ∞ to 0.

Furthermore, through the definition of θ, the SNR-plane formed by γ_1 and γ_2 (as shown in Fig. 5) can be considered as a set of curves. Each curve is characterized by the single parameter θ, $0 \le \theta < \infty$, and is expressed as

$$\gamma_2 = (1 + \gamma_1)^{\theta} - 1. \tag{3.26}$$

Note that any two curves intersect only at the origin. Also, the points on the same curve correspond to the same θ, and thus have the same compensation efficiency η. As shown in Fig. 5, all curves with $\theta \in (0, 1]$ form the region where $\gamma_1 \ge \gamma_2$. When θ varies from 0 to 1, the corresponding curve changes from $\gamma_2 = 0$ to $\gamma_2 = \gamma_1$ along the direction marked with an arrow. In the meantime, η decreases from ∞ to 0 (also see

[2]The notations $\lim_{w\to(\cdot)^-} f(w)$ and $\lim_{w\to(\cdot)^+} f(w)$ are the left limit and right limit of a given function $f(w)$, respectively, as w approaches a specified number.

Fig. 4). In contrast, all curves with $\theta \in (1, \infty)$ form the region where $\gamma_1 < \gamma_2$. As θ varies from 1 to ∞, the corresponding curve changes from $\gamma_2 = \gamma_1$ to $\gamma_2 = \infty$ and again η decreases from ∞ to 0.

3. The OPTS Policy Obtained through the Compensation-Efficiency Based SNR-Plane Partition

We use the SNR-plane partition to derive the OPTS policy for two-receiver cases. Given any compensation efficiency $\eta = \eta_0$, $\eta_0 \in (0, \infty)$, we obtain two solutions $\theta = \theta_1$ and $\theta = \theta_2$ to Eq. (3.25) as shown in Fig 4, which are determined by

$$\theta = \begin{cases} \theta_1 = \frac{1}{1+\eta_0}, & \text{if } \gamma_1 > \gamma_2; \\ \theta_2 = \frac{1+\eta_0}{\eta_0}, & \text{if } \gamma_1 < \gamma_2. \end{cases} \tag{3.27}$$

The points on the curves $\gamma_2 = (1+\gamma_1)^{\theta_1} - 1$ and $\gamma_2 = (1+\gamma_1)^{\theta_2} - 1$ correspond to the same η_0. That is, these two curves are contours for $\eta = \eta_0$. The curve $\gamma_2 = (1+\gamma_1)^{\theta_1} - 1$ lies in the region where $\gamma_1 \geq \gamma_2$, while the curve $\gamma_2 = (1+\gamma_1)^{\theta_2} - 1$ lies in the region where $\gamma_1 < \gamma_2$. Then, as shown in Fig. 5, the curves $\gamma_2 = (1+\gamma_1)^{\theta_2} - 1$, $\gamma_2 = (1+\gamma_1)^{\theta_1} - 1$, and $\gamma_2 = \gamma_1$ divide the SNR-plane into four exclusive regions, denoted by V_a, V_b, V_c, and V_d, respectively. Moreover, these regions are expressed by

$$\begin{cases} V_a \triangleq \big\{\boldsymbol{\gamma} | (1+\gamma_1)^{\theta_2} - 1 \leq \gamma_2\big\}; \\ V_b \triangleq \big\{\boldsymbol{\gamma} | \gamma_1 \leq \gamma_2 < (1+\gamma_1)^{\theta_2} - 1\big\}; \\ V_c \triangleq \big\{\boldsymbol{\gamma} | (1+\gamma_1)^{\theta_1} - 1 \leq \gamma_2 < \gamma_1\big\}; \\ V_d \triangleq \big\{\boldsymbol{\gamma} | 0 < \gamma_2 < (1+\gamma_1)^{\theta_1} - 1\big\}. \end{cases} \tag{3.28}$$

Clearly, we have $\eta \leq \eta_0$ for $\boldsymbol{\gamma} \in V_a \cup V_c$ and $\eta(\boldsymbol{\gamma}) > \eta_0$ for $\boldsymbol{\gamma} \in V_b \cup V_d$.

As discussed in Section E-2, in order to achieve high $\overline{R}^{\text{gp}}$, we need to use the *minimum* degradation of $\overline{R}_2\big(\boldsymbol{\lambda}(\boldsymbol{\gamma})\big)$ to obtain the *maximum* improvement of $\overline{R}_1\big(\boldsymbol{\lambda}(\boldsymbol{\gamma})\big)$.

That is, time proportions in the fading states that correspond to higher compensation efficiency need to be reallocated to transmission rate $r_1 = B\log(1+\gamma_1)$ with higher priority. Following this principle, we set $\lambda_1(\boldsymbol{\gamma}) = 1$ for all $\boldsymbol{\gamma}$'s in V_b and V_d, while setting $\lambda_1(\boldsymbol{\gamma}) = 0$ for all $\boldsymbol{\gamma}$'s in V_a and V_c. We call the above strategy *SNR-partition based policy*, and summarize it by $\widetilde{\boldsymbol{\lambda}}(\boldsymbol{\gamma})$ given as follows:

$$\widetilde{\boldsymbol{\lambda}}(\boldsymbol{\gamma}) = \begin{cases} (1,0)^\tau, & \text{if } \boldsymbol{\gamma} \in V_b \cup V_d \\ (0,1)^\tau, & \text{if } \boldsymbol{\gamma} \in V_a \cup V_c. \end{cases} \tag{3.29}$$

That is, the transmission rate r is equal to r_1 if $\boldsymbol{\gamma}$ falls into V_b or V_d, but is equal to r_2 if $\boldsymbol{\gamma}$ falls into V_a or V_c. Equivalently, r is determined by the worst-case SNR when $\boldsymbol{\gamma} \in V_b \cup V_c$, however, is determined by the best-case SNR when $\boldsymbol{\gamma} \in V_a \cup V_d$. Thus, we can see that receiver 1 can correctly receive the data in regions V_b, V_c, and V_d, while receiver 2 can achieve the transmission rate r in regions V_a, V_b, and V_c.

Without loss of generality, we initially set $\widetilde{\lambda}_1(\boldsymbol{\gamma}) = 0$ for all $\boldsymbol{\gamma}$'s, which is equivalent to $\eta_0 = \infty$. We then gradually decreases η_0 until the constriction $\overline{R}_1\big(\widetilde{\boldsymbol{\lambda}}(\boldsymbol{\gamma})\big) = \overline{R}_2\big(\widetilde{\boldsymbol{\lambda}}(\boldsymbol{\gamma})\big)$ in Eq. (3.19) is satisfied, where the η_0 satisfying this constriction is denoted by η_0^*. When η_0 decreases, the boundaries $\gamma_2 = (1+\gamma_1)^{\theta_1} - 1$ and $\gamma_2 = (1+\gamma_1)^{\theta_2} - 1$ both vary along the direction indicated by the dash-lined arrow in Fig. 5. Correspondingly, the areas of V_b and V_d get larger and more time proportions are reallocated to rate r_1. Thus, $\overline{R}_1\big(\widetilde{\boldsymbol{\lambda}}(\boldsymbol{\gamma})\big)$ is a decreasing function of η_0, while $\overline{R}_2\big(\widetilde{\boldsymbol{\lambda}}(\boldsymbol{\gamma})\big)$ is an increasing function of η_0. Moreover, define

$$\begin{cases} \Lambda_a & \triangleq & \int\int_{V_a} r_2 f_{\boldsymbol{\Gamma}}(\boldsymbol{\gamma}) d\gamma_1 d\gamma_2; \\ \Lambda_b & \triangleq & \int\int_{V_b} r_1 f_{\boldsymbol{\Gamma}}(\boldsymbol{\gamma}) d\gamma_1 d\gamma_2; \\ \Lambda_c & \triangleq & \int\int_{V_c} r_2 f_{\boldsymbol{\Gamma}}(\boldsymbol{\gamma}) d\gamma_1 d\gamma_2; \\ \Lambda_d & \triangleq & \int\int_{V_d} r_1 f_{\boldsymbol{\Gamma}}(\boldsymbol{\gamma}) d\gamma_1 d\gamma_2, \end{cases} \tag{3.30}$$

which evaluate the average transmission rate r over the regions V_a, V_b, V_c, and V_d, respectively. Then, according to the discussions of Eq. (3.29), $\overline{R}_1\big(\widetilde{\boldsymbol{\lambda}}(\boldsymbol{\gamma})\big) = \overline{R}_2\big(\widetilde{\boldsymbol{\lambda}}(\boldsymbol{\gamma})\big)$ implies $\Lambda_b + \Lambda_c + \Lambda_d = \Lambda_a + \Lambda_b + \Lambda_c$. Eliminating Λ_b and Λ_c and using η_0 to express the integral region (see Eqs. (3.27) and (3.28)), we get

$$\begin{aligned}\Lambda_d &= \int_0^{\infty}\int_0^{(1+\gamma_1)^{\frac{1}{1+\eta_0}}-1} \log(1+\gamma_1) f_{\boldsymbol{\Gamma}}(\boldsymbol{\gamma}) d\gamma_2 d\gamma_1 \\ &= \int_0^{\infty}\int_0^{(1+\gamma_1)^{\frac{\eta_0}{1+\eta_0}}-1} \log(1+\gamma_2) f_{\boldsymbol{\Gamma}}(\boldsymbol{\gamma}) d\gamma_1 d\gamma_2 \;=\; \Lambda_a. \end{aligned} \tag{3.31}$$

Since $F_{\boldsymbol{\Gamma}}(\boldsymbol{\gamma})$ is a continuous function over $(\mathbb{R}^+)^N$ as assumed in Section B-1, Λ_d and Λ_a are both continuous functions of η_0. Also, it is not difficult to derive $\Lambda_d < \Lambda_a$ if $\eta_0 = \infty$ and $\Lambda_d > \Lambda_a$ if $\eta_0 = 0$. Thus, the solution η_0^* to Eq. (3.31) must exist. Furthermore, the following Proposition 4 shows that the policy $\widetilde{\boldsymbol{\lambda}}(\boldsymbol{\gamma})$ is also the OPTS policy to Eq. (3.5).

Proposition 4. *For two-receiver cases, the OPTS policy to the goodput-optimization problem formulated in Eq. (3.5) is given by* $\boldsymbol{\lambda}^*(\boldsymbol{\gamma}) = \widetilde{\boldsymbol{\lambda}}(\boldsymbol{\gamma})|_{\eta_0=\eta_0^*}$, *where* η_0^* *is the solution of* η_0 *to Eq. (3.31), and* $\widetilde{\boldsymbol{\lambda}}(\boldsymbol{\gamma})$ *is determined by Eqs. (3.27)-(3.29).*

Proof. The proof is provided in Appendix C. □

Note that if $\{\gamma_n\}_{n=1}^N$ have discrete distributions, i.e., $F_{\boldsymbol{\Gamma}}(\boldsymbol{\gamma})$ is not a continuous function. There may not exist the solution $\eta_0 = \eta_0^*$ satisfying Eq. (3.31). In contrast, there must exist a certain η_{dis} satisfying (i) $\overline{R}_1\big(\widetilde{\boldsymbol{\lambda}}(\boldsymbol{\gamma})\big) \le \overline{R}_2\big(\widetilde{\boldsymbol{\lambda}}(\boldsymbol{\gamma})\big)$ with $\eta_0 \to \eta_{\text{dis}}^+$; (ii) $\overline{R}_1\big(\widetilde{\boldsymbol{\lambda}}(\boldsymbol{\gamma})\big) \ge \overline{R}_2\big(\widetilde{\boldsymbol{\lambda}}(\boldsymbol{\gamma})\big)$ with $\eta_0 \to \eta_{\text{dis}}^-$, where $\widetilde{\boldsymbol{\lambda}}(\boldsymbol{\gamma})$ follows Eqs. (3.27)-(3.29). Accordingly, the OPTS should be modified as

$$\boldsymbol{\lambda}^*(\boldsymbol{\gamma}) = \begin{cases} \widetilde{\boldsymbol{\lambda}}(\boldsymbol{\gamma})|_{\eta_0=\eta_{\text{dis}}}, & \text{for } \boldsymbol{\gamma} \text{ with } \eta \neq \eta_{\text{dis}} \\ (\lambda_{\text{dis}}, 1-\lambda_{\text{dis}})^{\tau}, & \text{for } \boldsymbol{\gamma} \text{ with } \eta = \eta_{\text{dis}}, \end{cases} \tag{3.32}$$

where we set

$$\lambda_{\rm dis}=\frac{\lim\limits_{\eta_0\to\eta_{\rm dis}^+}\overline{R}_1\left(\widetilde{\boldsymbol{\lambda}}(\boldsymbol{\gamma})\right)-\lim\limits_{\eta_0\to\eta_{\rm dis}^+}\overline{R}_2\left(\widetilde{\boldsymbol{\lambda}}(\boldsymbol{\gamma})\right)}{(1+\eta_{\rm dis})\int\int_{\boldsymbol{\gamma}:\eta(\boldsymbol{\gamma})=\eta_{\rm dis}}\mu_2(\boldsymbol{\gamma})f_{\mathbf{\Gamma}}(\boldsymbol{\gamma})d\gamma_1 d\gamma_2} \tag{3.33}$$

to achieve $\overline{R}_1\big(\boldsymbol{\lambda}^*(\boldsymbol{\gamma})\big)=\overline{R}_2\big(\boldsymbol{\lambda}^*(\boldsymbol{\gamma})\big)$. The proof of Eqs. (3.32) and (3.33) is similar to that of Proposition 4.

4. Case Investigation

Consider Rayleigh fading channels and assume that γ_1 and γ_2 are independent. Then, using Eqs. (3.27)-(3.28) and (3.30), we derive

$$\Lambda_a = Be^{\frac{1}{\overline{\gamma}_2}}\mathcal{G}\left(0,\frac{1}{\overline{\gamma}_2}\right)-\frac{B}{\overline{\gamma}_2}e^{\frac{1}{\overline{\gamma}_1}+\frac{1}{\overline{\gamma}_2}}H\left(1,\frac{\eta_0}{1+\eta_0},\frac{1}{\overline{\gamma}_2},\frac{1}{\overline{\gamma}_1}\right); \tag{3.34}$$

$$\Lambda_b = Be^{\frac{1}{\overline{\gamma}_1}+\frac{1}{\overline{\gamma}_2}}\left[\frac{\overline{\gamma}_2}{\overline{\gamma}_1+\overline{\gamma}_2}\mathcal{G}\left(0,\frac{1}{\overline{\gamma}_1}+\frac{1}{\overline{\gamma}_2}\right)-\frac{1}{\overline{\gamma}_1}H\left(1,\frac{1+\eta_0}{\eta_0},\frac{1}{\overline{\gamma}_1},\frac{1}{\overline{\gamma}_2}\right)\right]; \tag{3.35}$$

$$\Lambda_c = Be^{\frac{1}{\overline{\gamma}_1}+\frac{1}{\overline{\gamma}_2}}\left[\frac{\overline{\gamma}_1}{\overline{\gamma}_1+\overline{\gamma}_2}\mathcal{G}\left(0,\frac{1}{\overline{\gamma}_1}+\frac{1}{\overline{\gamma}_2}\right)-\frac{1}{\overline{\gamma}_2}H\left(1,1+\eta_0,\frac{1}{\overline{\gamma}_2},\frac{1}{\overline{\gamma}_1}\right)\right]; \tag{3.36}$$

$$\Lambda_d = Be^{\frac{1}{\overline{\gamma}_1}}\mathcal{G}\left(0,\frac{1}{\overline{\gamma}_1}\right)-\frac{B}{\overline{\gamma}_1}e^{\frac{1}{\overline{\gamma}_1}+\frac{1}{\overline{\gamma}_2}}H\left(1,\frac{1}{1+\eta_0},\frac{1}{\overline{\gamma}_1},\frac{1}{\overline{\gamma}_2}\right). \tag{3.37}$$

According to the discussion of Eq. (3.29), the goodput is given by

$$\overline{R}^{\rm gp}=\overline{R}_2=\overline{R}_1=(\Lambda_b+\Lambda_c+\Lambda_d)|_{\eta_0=\eta_0^*}. \tag{3.38}$$

(i) When γ_1 and γ_2 are i.i.d., we have $\overline{\gamma}_1=\overline{\gamma}_2$ and use $\overline{\gamma}$ to denote the average SNR. To obtain $\Lambda_a=\Lambda_d$, the equation $1/(1+\eta_0)=\eta_0/(1+\eta_0)$ is required according to Eqs. (3.34) and (3.37). We solve this equation and get $\eta_0^*=1$. Then, using Eqs. (3.34)-(3.38), we can get the same results as Eq. (3.13). Also, it is easy to verify that $\widetilde{\boldsymbol{\lambda}}(\boldsymbol{\gamma})$ given in Eq. (3.29) with $\eta_0^*=1$ is equal to the policy given in Proposition 1.
(ii) We fix $\overline{\gamma}_2=10$ dB and varies $\overline{\gamma}_1$ from 9 dB to 11 dB. We numerically calculate η_0^* and plot the SNR-plane partition in Fig. 6. In Fig. 6, with the decreasing of $\overline{\gamma}_1$,

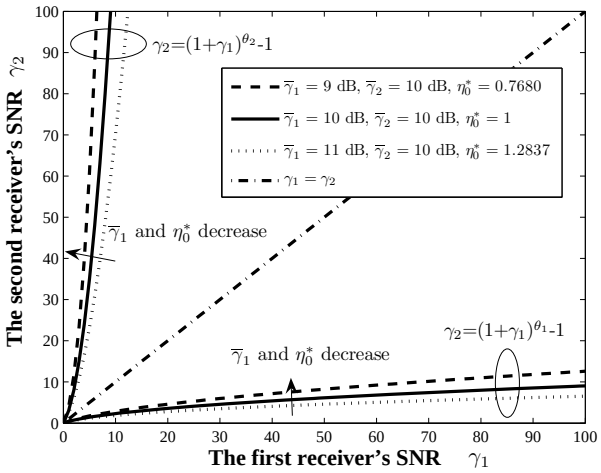


Fig. 6. SNR-plane partition for two-receiver scenarios. γ_1 and γ_2 are independent. $\overline{\gamma}_2 = 10$ dB and $\overline{\gamma}_1 = 9, 10, 11$ dB.

the boundaries $\gamma_2 = (1+\gamma_1)^{\theta_1} - 1$ and $\gamma_2 = (1+\gamma_1)^{\theta_2} - 1$ for the SNR-plane partition gradually move along the directions of arrows, respectively. Thus, more and more time proportions are allocated to the transmission rate r_1, which implies that the receiver with poorer statistical channel quality usually gets more time proportions (result in a smaller η_0^*) to determine the transmission rate.

F. A Sub-Grouping Based Suboptimal Rate-Adaptation Policy

Sections D and E derive the OPTS policies for goodput-maximization in some specific fading channels. However, the optimal time-sharing policy to maximize the average multicast goodput of a multicast group with any number of receivers and non-i.i.d. fading channels is still difficult to obtain. Also, the knowledge of channel distribution is usually unavailable for realistic systems. To solve above problems, we derive a simple *sub-grouping* based suboptimal rate-adaptation policy which can achieve good performance without the knowledge of channel distribution.

The multicast transmission begins at the 1st time interval and we use ℓ, $\ell \geq 1$,

to index the time intervals. Similarly to Propositions 1 and 4, we only use one transmission rate within each time interval. The transmission rate r and instantaneous SNR γ_n in the ℓth time interval are denoted by $r[\ell]$ and $\gamma_n[\ell]$, respectively. Note that under the max-min definition of the average multicast goodput (see Eq. (3.4)), we usually need to allocate more time proportion to the transmission rate determined by CSI's from receivers with *statistically* poorer channels. However, because the channel distribution is unknown, we cannot get the statistical channel qualities in advance. To solve this problem, we introduce a metric termed *accumulative rate*, which is defined as follows.

Definition 3. *The accumulative rate of the nth receiver until the ℓth time interval, denoted by $\zeta_n[\ell]$, is this receiver's average achieved data rate over the 1st time interval through the ℓth time interval. In particular, $\zeta_n[\ell]$ is determined by*

$$\zeta_n[\ell] \triangleq \frac{1}{\ell}\sum_{j=1}^{\ell} c\big(\gamma_n[j], r[j]\big), \qquad n = 1, 2, \ldots, N. \tag{3.39}$$

Furthermore, we define $\zeta_{\min}[\ell] \triangleq \min_{1\le n\le N}\{\zeta_n[\ell]\}$ and

$$\Omega[\ell] \triangleq \{n|\ \zeta_n[\ell] - \zeta_{\min}[\ell] \le \omega\, \zeta_{\min}[\ell]\}\,, \tag{3.40}$$

where $\frac{\zeta_n[\ell]-\zeta_{\min}[\ell]}{\zeta_{\min}[\ell]}$ evaluates the difference level between $\zeta_n[\ell]$ and $\zeta_{\min}[\ell]$ and ω is a small positive real number termed *difference-level threshold*. In the ℓth time interval, we construct a new $N^{\Omega}[\ell]\times 1$ vector by using all $\gamma_n[\ell]$, $n \in \Omega[\ell-1]$, where $N^{\Omega}[\ell]$ is the cardinality of $\Omega[\ell-1]$ and we denote this new vector by $\boldsymbol{\gamma}^{\Omega}[\ell]$. Note that eventually, the nth receiver's average achieved data rate and the average multicast goodput of the multicast group are given by $\overline{R}_n\big(\boldsymbol{\lambda}(\boldsymbol{\gamma})\big) = \lim_{\ell\to\infty}\{\zeta_n[\ell]\}$ and $\overline{R}^{\mathrm{gp}} = \lim_{\ell\to\infty}\{\zeta_{\min}[\ell]\}$, respectively.

Based on Eq. (3.40), the set $\Omega[\ell]$ includes multicast receivers with lower accumu-

lative rates, which usually implies that their statistical channel qualities are poorer. Consequently, we need to efficiently benefit these receivers such that the high average multicast goodput can be obtained. Following this principle, we determine the transmission rate $r[\ell]$ in the ℓth time interval by CSI from receivers belonging to $\Omega[\ell-1]$. In particular, using $N^{\Omega}[\ell]$ instead of N, we apply the policy $\boldsymbol{\lambda}^*(\boldsymbol{\gamma})$ given in Proposition 1 to $\boldsymbol{\gamma}^{\Omega}[\ell]$ to determine $r[\ell]$. More specifically, we set

$$r[\ell] = B\log\left(1+\widehat{\gamma}_{i^*}^{\Omega}[\ell]\right), \tag{3.41}$$

where

$$i^* = \arg\max_{1\le i\le N^{\Omega}[\ell]}\left\{iB\log\left(1+\widehat{\gamma}_{i}^{\Omega}[\ell]\right)\right\}. \tag{3.42}$$

The policy given in Eqs. (3.41) and (3.42) attempts to maximize the sum of achieved rates over receivers with lower accumulative rates. While Eq. (3.40) removes the receivers with higher accumulative rates from the set $\Omega[\ell]$, which usually have statistically better channel qualities. As a result, this strategy benefits receivers with lower accumulative rate and thus can effectively improve the average multicast goodput. Since the above policy dynamically constructs a sub-group of receivers within the entire multicast group in each time-interval, we call the above strategy *sub-grouping* (SG) based rate-control policy, which is summarized as follows.

Operation in the ℓth time interval:

Step 1) Determine $\Omega[\ell-1]$ by using Eq. (3.40).

Step 2) Determine the transmission rate $r[\ell]$ by using Eqs. (3.41) and (3.42).

Step 3) Multicast data at rate $r[\ell]$; update

$$\zeta_n[\ell+1] := \frac{\ell\zeta_n[\ell]+c(\gamma_n[\ell],r[\ell])}{\ell+1}.$$

The above policy is a special time-sharing rate-adaptation policy. This policy can ef-

fective apply the algorithm derived in i.i.d. fading environments into non-i.i.d. fading environments across multicast receivers. If the differences among all receivers' accumulative rates are small enough, this policy reduces to $\boldsymbol{\lambda}^*(\boldsymbol{\gamma})$ for i.i.d. fading channels. The value of ω will affect the goodput performance, which will be investigated through simulations In Section G.

G. Numerical and Simulation Evaluations

We evaluate the performance of our proposed time-sharing policies through numerical and simulation evaluations. We also compare the performance of our proposed policies with other existing schemes, which are described as follows.

(i) The mth largest SNR-dominating (m-LSD) policy

The m-LSD policy is a simple rate adaptation policy with a static strategy, where the transmission rate is determined by $r = B\log(1+\widehat{\gamma}_m)$. When $m = 1$, we get the best-case SNR-dominating (BSD) policy. When $m = N$, we get the worst-case SNR-dominating (WSD) policy. In i.i.d. fading channels, we also exam m^*-LSD policy, where m^* maximize goodput over all m-LSD policies, $1 \leq m \leq N$. In [31], The m-LSD scheme was studied for the scaling law of the throughput-delay tradeoff. In this chapter, we simulate the average multicast goodput of these schemes.

(ii) The *constant*-rate (CR) policy

The CR policy is a nonadaptive policy. Also, it is usually not a time-sharing policy. The constant transmission rate r is equal to $B\log(1+\gamma_{\text{th}})$, where γ_{th} is selected to maximize $\overline{R}^{\text{gp}}$ over all constant rates and is determined by

$$\gamma_{\text{th}} = \arg\max_{\gamma}\left\{\min_{1\leq n\leq N}\left\{(1-F_{\Gamma_n}(\gamma_{\text{th}}))B\log(1+\gamma_{\text{th}})\right\}\right\}.$$

It is clear that r and $\overline{R}^{\text{gp}}$ only depend on the marginal distributions of $\{\gamma_n\}_{n=1}^N$.

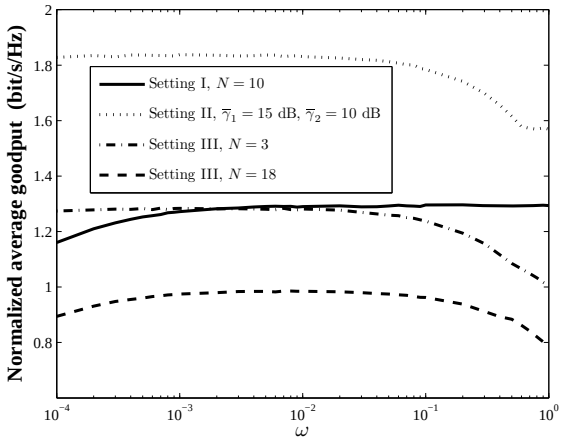


Fig. 7. The impact of ω on the normalized average multicast goodput $\overline{R}^{\rm gp}/B$ achieved by using the SG rate-adaptation policy.

Through Proposition 3, in i.i.d. fading channels, the OPTS policy converges to the CR policy as $N \to \infty$.

We use Rayleigh fading model as the typical example for numerical and simulation evaluation. In particular, we test various policies under the following three environment settings.

Setting I: The instantaneous SNR $\{\gamma_n\}_{n=1}^N$ are i.i.d. with 0 dB $\leq \overline{\gamma} \leq$ dB and $N = 1, 2, \ldots, 20$, where $\overline{\gamma} = \overline{\gamma}_1 = \cdots = \overline{\gamma}_N$.

Setting II: $N = 2$. γ_1 and γ_2 are independent with $\overline{\gamma}_2 = 10$ dB. $\overline{\gamma}_1$ varies from 5 dB to 15 dB.

Setting III: $\{\gamma_n\}_{n=1}^N$ are independent and $N = 3, 6, 9, 12, 15, 18$. All receivers are divided into three small groups, each including $N/3$ receivers with the same average SNR. The average SNR's for the three groups are 7 dB, 10 dB, and 13 dB, respectively.

For the OPTS policies under Setting I with $N \leq 2$ and those under Setting II, m-LSD policies under Settings I and II, and the CR policy under Settings I and II, the average multicast goodput are numerically calculated. The results for all other scenarios are obtained by using simulations.

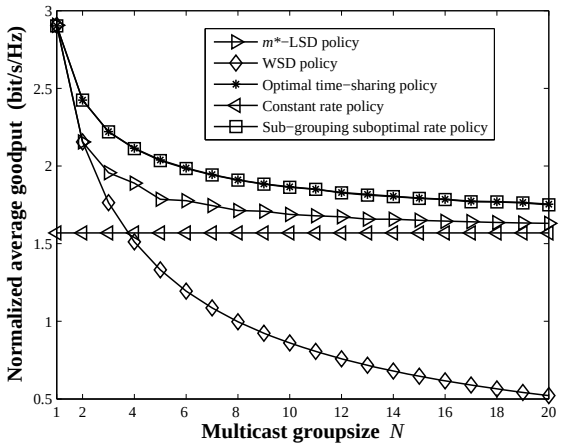


Fig. 8. The normalized average multicast goodput $\overline{R}^{\mathrm{gp}}/B$ versus multicast group size N under Setting I with $\overline{\gamma} = 10$ dB.

We first investigate the selection of the difference-level threshold ω in Eq. (3.40) for our proposed SG rate-adaptation policy. Fig. 7 plots the impact of ω on the average multicast goodput. We can see from Fig. 7 that either a too large or too small ω may lead to goodput degradation. If ω approaches 0, the receiver with the minimum accumulative data rate will determine the transmission rate r, which, however, cannot efficiently benefit the entire multicast group (the sum of achieved rates). When ω gets too large, the suboptimal policy reduces to the policy given in Proposition 1, which cannot efficiently improve the average achieved rate of receivers with statistically poor channels. We also observe that when ω falls into [0.001, 0.03], the suboptimal policy can achieve relatively better goodput and the performance is not sensitive to the variation of ω. Thus, we simply set $\omega = 0.01$ for the SG rate-adaptation policy under all three environment settings.

Fig. 8 plots the average multicast goodput versus the multicast group size N under Setting I with $\overline{\gamma} = 10$ dB. As discussed in Section D-1, in order to maximize goodput in i.i.d. fading channels, we need to maximize the sum of achieved rates in each fading independently, which can be attained by neither the CR policy nor

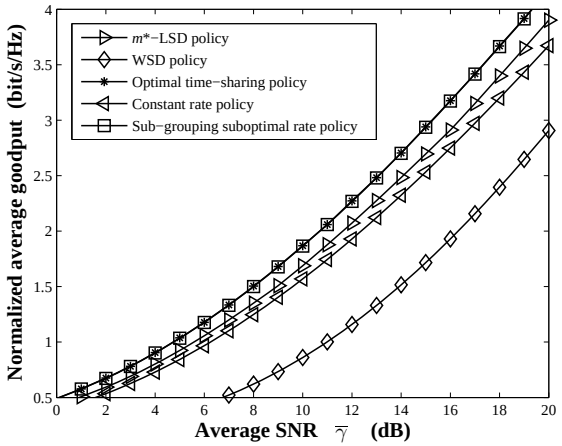


Fig. 9. The normalized average multicast goodput $\overline{R}^{\mathrm{gp}}/B$ versus the average SNR $\overline{\gamma}$ under Setting I with $N = 10$.

the m-LSD policy. Consequently, we can see from Fig. 8 that our derived OPTS policy achieves the highest goodput over all other policies. Moreover, our proposed SG rate-adaptation policy can achieve almost the same performance as the OPTS policy. In addition, we observe that all time-sharing policies' goodput are decreasing functions of N, which confirms our claim for the OPTS policy in Proposition 2. As N gets larger, the average multicast goodput of our derived OPTS policy approaches that of the CR policy, which verifies our claim in Proposition 3. Fig. 9 plots the average multicast goodput versus the average SNR under Setting I with $N = 10$ dB. From Fig. 9, we see that with the increasing of channel qualities, the average multicast goodputs under all policies also increase, while our proposed OPTS and SG rate-adaptation policies outperform all others. More importantly, the performance improvement gained by our proposed policies will also increase as $\overline{\gamma}$ becomes larger.

Under Setting II, Fig. 10 plots the average multicast goodput as the first receiver's average SNR varies. As shown in Fig. 10, our derived OPTS policy outperforms all other rate-adaptation policies. Moreover, our proposed SG rate-adaptation policy can achieve the performance very close to the OPTS policy. Also note that except for

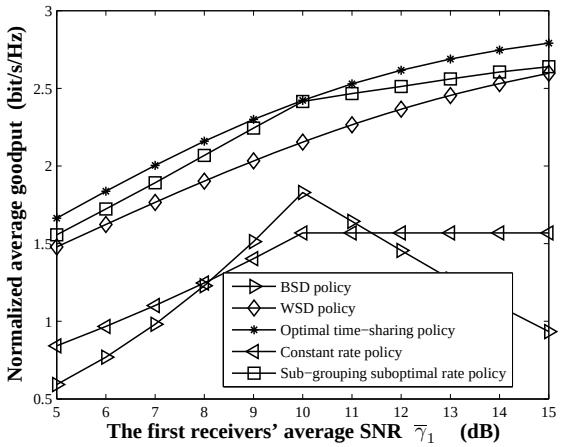


Fig. 10. The normalized average multicast goodput $\overline{R}^{\text{gp}}/B$ versus the first receiver's average SNR $\overline{\gamma}_1$ under Setting II.

the CR policy and the BSD policy, all other policies can achieve higher goodput as $\overline{\gamma}_1$ increases. In particular, the average multicast goodput of the CR policy remains unchanged when $\overline{\gamma}_1 > \overline{\gamma}_2$. This is expected since the goodput of the CR policy only depends on marginal distributions of γ_n's such that the goodput performance is determined by the receiver with lower statistical channel qualities. More surprisingly, the average multicast goodput of the BSD policy even degrades when $\overline{\gamma}_1$ becomes larger than $\overline{\gamma}_2$. This phenomenon can be explained as follows. Under the BSD policy, only the receiver with the best-case instantaneous SNR can correctly receive the data. Thus, as long as the average SNR's of different receivers are unequal, the receiver with higher average SNR will occupy more opportunities for reliable data reception while other receivers' suffer severe data losses.

Fig. 11 plots the average multicast goodput as a function of multicast group size N, respectively, for various policies under a non-i.i.d. environment – Setting III. Since all receivers are divided into three smaller groups according to the average SNR, we test $m = N/3$, $m = 2N/3$, and $m = N$ for the m-LSD policy. We can see from Fig. 11(a) that our proposed SG rate-adaptation policy achieve at least 25% perfor-

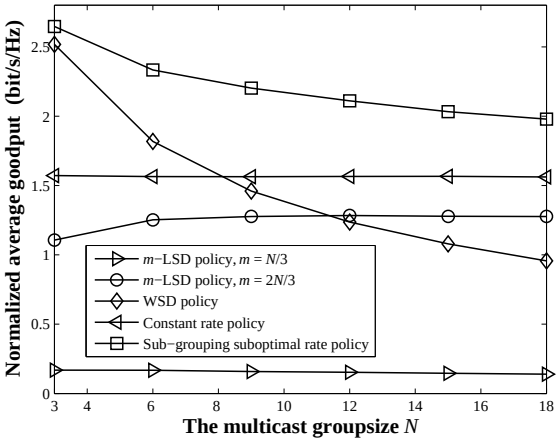


Fig. 11. The normalized average multicast goodput $\overline{R}^{\text{gp}}/B$ versus the Multicast group size N under Setting III.

mance improvement over all other policies when $6 \leq N \leq 18$, which again verifies the effectiveness of our proposed SG rate-adaptation policy. In contrast, the m-LSD policies, e.g., $m = 2N/3$ and $m = N/3$, result in significant goodput degradation, which can be explained by using the similar arguments for the performance of BSD in Fig. 10. When $m = N$, the m-LSD policy reduces to the WSD policy and the average multicast goodput degrades quickly as N increases. In addition, our proposed SG rate-adaptation policy can achieve more goodput gain over the CR policy in non-i.i.d. fading channels than in i.i.d. fading channels.

H. Summary

In this chapter, we derived the optimal time-sharing rate-adaptation policy for mobile multicast with i.i.d. fading channels. In i.i.d. fading environments, to maximize average multicast goodput is equivalent to maximizing the sum of achieved rates over receivers in each fading state independently. The derived optimal policy has good scalability in term of the multicast group size. As the multicast group size approaches infinity, the derived optimal policy converges to a constant rate policy with a non-

zero goodput. By using a SNR-plane partition based method, we also derived the optimal time-sharing policy for two-receiver cases with non-i.i.d. fading channels. To solve the problem that the statistical channel information is usually unavailable to the sender, we developed a sub-grouping based suboptimal rate-adaptation policy, which can effective apply the algorithm derived in i.i.d. fading environments into non-i.i.d. fading environments across multicast receivers. Simulation and numerical analyses show that our proposed policies significantly outperform the existing rate adaptation schemes.

CHAPTER IV

EFFECTIVE CAPACITY OF MULTICAST OVER FADING CHANNELS IN WIRELESS NETWORKS

A. Introduction

While mobile multicast has received a great deal of research attention [27–29,31,58,75, 76], the design of efficient wireless multicast schemes for diverse QoS-constrained services still faces many challenges. One the one hand, due to the heterogeneous channel-fading status across multicast receivers, it is difficult to achieve high throughput and reliability simultaneously. On the other hand, the provisionings of deterministic quality-of-service (QoS) requirements such as hard delay bounds are usually unrealistic because of the highly-varying wireless channels. Correspondingly, statistical-QoS guaranteed approaches need to be developed for wireless multicast services with diverse QoS requirements.

As discussed in Chapters I and III, in order to achieve high system throughput for wireless multicast, the multicast transmission rate cannot be always limited by the worst-case channel quality. Identifying that there is a fundamental tradeoff between multicast service rate and the reliability, in [76] we investigated the impact of the average loss-rate requirements on optimizing the average multicast throughput over broadcast fading channels, where for any given time instant the multicast transmission rate is not necessarily determined by the minimum achievable rate among all multicast receivers. Further note that optimizing the average multicast throughput cannot effectively characterize the QoS provisionings for delay-sensitive services, such as video and audio multimedia multicasting. For these services, the delay-bound guarantee is even more important than the optimization for average throughput. Thanks to the

effective capacity theory developed by the authors of [8], we can use it to develop efficient rate-adaptation techniques in wireless networks and to examine the system throughput subject to the diverse statistical delay-QoS constraint. However, it still remains as one of major challenges to integrate the effective capacity approaches into the design of efficient rate-adaptation schemes for wireless multicast with diverse QoS requirements.

To overcome the aforementioned problems, we propose an efficient framework for developing multicast rate-adaptation schemes over broadcast fading channels with statistical-delay and loss-rate QoS constraints. For rate adaptation, we employ the time-sharing (TS) and superposition-coding (SPC) techniques, respectively, to handle the heterogeneous qualities over channels across multicast receivers. We also develop a *pre-drop* scheme to implement the more efficient QoS-driven wireless multicasting. Then, given the statistical delay-QoS requirement and the average loss-rate threshold, we formulate the effective capacity maximization problem to derive the optimal channel-aware rate adaptation and pre-drop schemes. The optimal TS-based and SPC-based multicast policies are derived respectively. Extensive simulations are also conducted to evaluate the effective-capacity performance of our derived optimal schemes.

The rest of the paper is organized as follows. Section B presents the system model. Section C proposes the framework of effective capacity optimization with statistical delay and loss QoS constraints for multicast over fading channels in wireless networks. Section D develops the optimal TS-based adaptive multicast transmission policies. Section E obtains the optimal TS-based multicast policies under two limiting scenarios of delay-QoS constraints. Section F solves for the optimal SPC based adaptive multicast transmission policies. Section G compares the performances of our derived optimal adaptive multicast policies with the suboptimal policies through

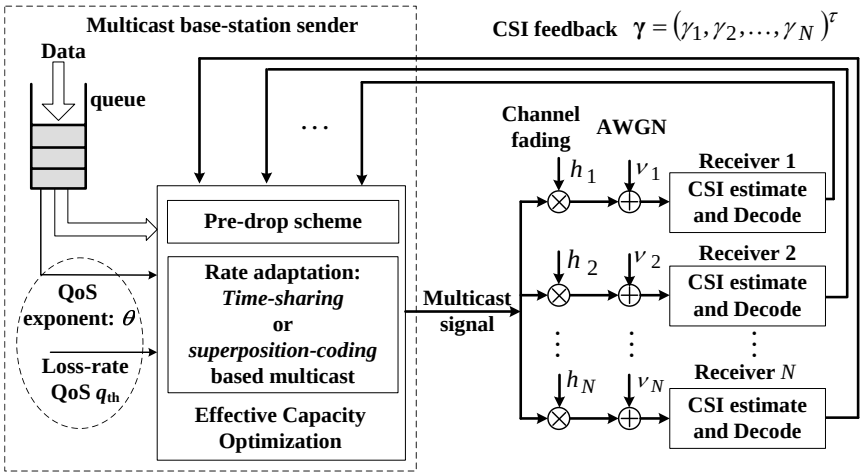


Fig. 12. The system model for mobile multicast with adaptive multicast transmissions over broadcast fading channels in wireless networks.

simulation evaluations. The chapter concludes with Section H.

B. System Model

We consider a discrete-time one-to-many multicast system as illustrated in Fig. 12, where the base-station sender transmits a single data stream to N multicast receivers over broadcast fading channels. The sender employs a single antenna for transmission and each multicast receiver uses a single antenna to receive data. We focus on the flat-fading channels. Then, the physical-layer multicasting model can be given by

$$y_n[k] = h_n[k]x[k] + v_n[k], \tag{4.1}$$

where $[k]$ is the index for consecutive time frames each with the fixed length equal to T. In Eq. (4.1), $x[k]$ is the complex multicast signal with the spectral bandwidth B, $h_n[k]$ is the complex channel gain between the sender and the nth multicast receiver, $y_n[k]$ denotes the nth receiver's received signal, and $v_n[k]$'s are circular complex additive white Gaussian noise (AWGN) with power spectral density σ_0. We model the time-

varying channel gain $h_n[k]$ as an ergodic and stationary block-fading process, where $h_n[k]$, $n = 1, 2, \ldots, N$, is invariant within a time frame, but varies independently from frame to frame. When the context is clear, we will drop the time index $[k]$ to simplify notations.

The transmit power $\mathbb{E}\{x^2\}$ within each time frame is equal to a constant denoted by $\overline{P}$, where $\mathbb{E}\{\cdot\}$ denotes the expectation. Then, we characterize the channel-state information (CSI) by the instantaneous SNR received at each multicast receiver, which is denoted by γ_n and defined as $\gamma_n \triangleq \overline{P}|h_n|^2/(\sigma_0 B)$, $n = 1, 2, \ldots, N$. The SNR vector is formed as $\boldsymbol{\gamma} \triangleq (\gamma_1, \gamma_2, \ldots, \gamma_N)^\tau$, representing a fading state, where $(\cdot)^\tau$ denotes the transpose operator. Unless otherwise mentioned, we assume that the SNR's $\{\gamma_n\}_{n=1}^N$ follow independent and identically distributed (i.i.d.) distribution, and denote the average SNR of γ_n, $n = 1, 2, \ldots, N$, by $\overline{\gamma}$. We further sort the elements of $\boldsymbol{\gamma}$ in the decreasing order and get an ordered SNR vector denoted by $\boldsymbol{\gamma}_\pi \triangleq (\gamma_{\pi(1)}, \gamma_{\pi(2)}, \ldots, \gamma_{\pi(N)})^\tau$, where $\pi(\cdot)$ is the sorting operator such that $\gamma_{\pi(1)} \geq \gamma_{\pi(2)} \geq \ldots \geq \gamma_{\pi(N)}$. We suppose that CSI information are available at both the sender and multicast receivers.

C. Framework of Effective Capacity Optimization for Mobile Multicast

1. Rate Adaptation for Multicast Transmissions

The multicast receivers' channel qualities are different at each time instant. Then, in order to efficiently regulate the multicast transmission rate based on all multicast receivers' CSI, we divide the multicast data to be transmitted within each time frame into N parts. The transmission rate of the nth data part, denoted by r_n (nats/frame), is set just within the capability of γ_n for correctly decoding. Accordingly, only the receivers with SNR higher than or equal to γ_n can correctly decode this part of data.

We define the transmission-rate vector for the N data parts as $\boldsymbol{r} \triangleq (r_1, r_2, \ldots, r_N)^\tau$. In addition, we use $r_{\pi(i)}$ to denote the transmission rate associated with the ordered SNR $\gamma_{\pi(i)}$. The above rate-adaptation strategy for wireless multicast can be implemented by either the time-sharing (TS) or superposition-coding (SPC) techniques, as illustrated in Fig. 12. The details for the TS-based and SPC-based rate-adaptation policies are elaborated on as follows.

a. Time-sharing (TS) based rate-adaptation policy

The sender divides each time-frame into N time slots with lengths equal to $T_1, T_2, \ldots, T_N$, which are associated with $\gamma_1, \gamma_2, \ldots, \gamma_N$, respectively, where $\sum_{i=1}^{N} T_i = T$. We suppose that the capacity-achieving codes are used within each frame, and then the transmission rate r_n in the nth time slot is set equal to the Shannon capacity $TB\log(1+\gamma_n)$ of γ_n. We denote the time proportion of the nth time slot by λ_n, $n = 1, 2, \ldots, N$, where $\lambda_n \triangleq T_n/T$, $0 \leq \lambda_n \leq 1$, and $\sum_{n=1}^{N} \lambda_n = 1$. Then, we can control time proportions of the N time slots to regulate the TS strategy, and we can characterize the TS rate-adaptation policy by a vector function denoted by $\boldsymbol{\lambda} \triangleq (\lambda_1, \lambda_2, \ldots, \lambda_N)^\tau$. Also, the time proportion associated with $\gamma_{\pi(i)}$ is denoted by $\lambda_{\pi(i)}$, and we define $\widetilde{\boldsymbol{\lambda}} \triangleq (\widetilde{\lambda}_1, \widetilde{\lambda}_2, \ldots, \widetilde{\lambda}_N)^\tau$. In addition, if we have $\gamma_{\pi(i)} = \gamma_{\pi(k)}$ for some $i < k$, implying $r_{\pi(i)} = r_{\pi(k)}$, we set $\lambda_{\pi(i)} = 0$ without loss of generality.

b. Superposition-coding (SPC) based rate-adaptation policy

For SPC [71] based rate-adaptation policy, signals of different data parts are superimposed to each other within the entire time frame. The power allocated to the nth layer is denoted by P_n, $n = 1, 2, \ldots, N$, where $\sum_{n=1}^{N} P_n = \overline{P}$. We further denote the power proportion of the nth data part by $\mu_n \triangleq P_n/\overline{P}$, $n = 1, 2, \ldots, N$, where $\sum_{n=1}^{N} \mu_n = 1$. Information theory results [13, 71] have shown that by using the SPC,

the achievable region of the rate vector $\boldsymbol{r}$ is a convex region given by

$$\left\{ \boldsymbol{r} \;\middle|\; r_{\pi(i)} = BT \log\left(1 + \frac{\mu_{\pi(i)}\gamma_{\pi(i)}}{1 + \gamma_{\pi(i)} \sum_{j=1}^{i-1} \mu_{\pi(j)}}\right), \right.$$
$$\left. i = 1, 2, \ldots, N,\ \mu_{\pi(i)} \in [0, 1],\ \sum_{n=1}^{N} \mu_{\pi(i)} \leq 1 \right\}, \quad (4.2)$$

where $\mu_{\pi(i)}$ is the power proportion for the data part associated with $\gamma_{\pi(i)}$. With $\sum_{i=1}^{N} \mu_{\pi(i)} = 1$, the achieved rate vector falls onto the boundary of the region [13] given by Eq. (4.2). Accordingly, for the nth multicast receiver, it can successfully decode all data parts associated with SNR lower than or equal to γ_n by using the successive interference cancellation (SIC) technique [13,71]. Then, in SPC-based multicast systems, we can control the rate adaptation through regulating the power vector $\boldsymbol{\mu}$, where $\boldsymbol{\mu} \triangleq (\mu_1, \mu_2, \ldots, \mu_N)^{\tau}$. Similarly to the TS-based multicast transmissions, if we have $\gamma_{\pi(i)} = \gamma_{\pi(k)}$ for some $i < k$, we set $\mu_{\pi(i)} = 0$ without loss of generality, where $\mu_{\pi(i)}$ denotes the power proportion corresponding to $\gamma_{\pi(i)}$. For the details of SPC and SIC techniques, please refer to [13, 71].

2. Pre-Drop Scheme

To implement the more flexible adaptive multicast transmissions with statistical QoS provisionings, we develop an adaptive queue-management scheme, called the *pre-drop* scheme, as follows. In each frame, the sender can drop some data from the head of the queue. Then, the *pre-drop rate* is defined as the amount of data dropped in a frame, which is denoted by R_d (nats/frame). We employ this strategy because of the following reasons. For delay-sensitive services, once the delay-bound has been violated for the data in the front end of the queue, these data are usually useless for receivers, and the attempt to keep transmitting these data will further delay other data backlogged in the queue. When the channel qualities becomes poorer, we

can *predict* that more data will be backlogged in the queue, and the possibilities of violating the delay/queue-length bound will increase accordingly. To suppress the rapid growth of queue length and achieve the more robust queuing behaviors, we can drop some data from the queue head in each time frame. In order to efficiently decrease the violating probability for the delay/queue-length bound, we can expect that more data need to be dropped under the poorer channel qualities and vice versa. Since this strategy uses the channel quality to *predict* future queuing behaviors, we term this scheme the *pre-drop* scheme. We treat the dropped data as the transmitted data, but count them as losses to all receivers. Clearly, we need to carefully choose R_d based on the CSI such that the loss level will not violate the loss-QoS constraint (the loss-QoS constraint is detailed in Section B-4).

3. Statistical Metrics for Multicast Rate Adaptation

Before getting into the details of the effective-capacity optimization for multicast transmissions, we need to define a set of statistical metrics in the followings. Note that some metrics defined in this chapter have the same meaning as those defined in Chapter III. However, to avoid confusions due to the differences between the frameworks in these two chapters, we redefine all corresponding metrics in this chapter.

Definition 4. *The* sending rate *for mobile multicast, denoted by R_s, is the total transmission rate averaged within a time frame, which is determined by*

$$R_s \triangleq \begin{cases} \sum_{i=1}^{N} \lambda_{\pi(i)} r_{\pi(i)}, & \textit{for TS-based policy;} \\ \sum_{i=1}^{N} r_{\pi(i)}, & \textit{for SPC-based policy,} \end{cases} \tag{4.3}$$

where

$$r_{\pi(i)} \triangleq \begin{cases} BT\log(1+\gamma_{\pi(i)}), & \textit{for TS-based policy;} \\ BT\log\left(1+\frac{\mu_{\pi(i)}\gamma_{\pi(i)}}{1+\gamma_{\pi(i)}\sum_{j=1}^{i-1}\mu_{\pi(j)}}\right), & \textit{for SPC-based policy} \end{cases} \tag{4.4}$$

as described in Section C-1. Then, the instantaneous throughput*, denoted by R, is defined as the sum of pre-drop rate R_d and sending rate R_s:*

$$R \triangleq R_s + R_d, \tag{4.5}$$

which characterizes the total service (departure) rate of multicast data within a time frame. Moreover, $\mathbb{E}_{\boldsymbol{\gamma}}\{R\}$ is called the average throughput*, where $\mathbb{E}_{\boldsymbol{\gamma}}\{\cdot\}$ denotes the expectation over $\boldsymbol{\gamma}$.*

Definition 5. *The* instantaneous goodput *of the nth multicast receiver, denoted by g_n, is the sum rate over the data which can be correctly decoded by the nth receiver's within a time frame, which is determined by*

$$g_n \triangleq \begin{cases} \sum_{i=1}^{N}\lambda_{\pi(i)}\delta(\gamma_n \geq \gamma_{\pi(i)})r_{\pi(i)}, & \textit{for TS-based policy;} \\ \sum_{i=1}^{N}\delta\left(\gamma_n \geq \gamma_{\pi(i)}\right)r_{\pi(i)}, & \textit{for SPC-based policy,} \end{cases} \tag{4.6}$$

where $\delta(\cdot)$ is the indication function (for a given statement u, $\delta(u)=1$ if u is true, and 0 otherwise). The expectation $\mathbb{E}_{\boldsymbol{\gamma}}\{g_n\}$ is the nth receiver's average goodput*.*

Definition 6. *The* instantaneous sum goodput*, denoted by g_{sum}, is defined as the sum of g_n over all multicast receivers:*

$$g_{\text{sum}} \triangleq \sum_{n=1}^{N} g_n = \begin{cases} \sum_{i=1}^{N}\lambda_{\pi(i)}v_{\pi(i)}r_{\pi(i)}, & \textit{for TS-based policy;} \\ \sum_{i=1}^{N}v_{\pi(i)}r_{\pi(i)}, & \textit{for SPC-based policy.} \end{cases} \tag{4.7}$$

where $v_{\pi(i)}$ denotes the number of receivers whose SNR is higher than or equal to $\gamma_{\pi(i)}$. The expectation $\mathbb{E}_{\boldsymbol{\gamma}}\{g_{\text{sum}}\}$ is then called the average sum goodput*.*

Definition 7. *The* average loss rate *of the nth receiver, denoted by q_n, is the fraction of the average throughput which cannot be correctly decoded by the nth receiver. Then, q_n is given by*

$$q_n \triangleq \frac{\mathbb{E}_{\boldsymbol{\gamma}}\{R\} - \mathbb{E}_{\boldsymbol{\gamma}}\{g_n\}}{\mathbb{E}_{\boldsymbol{\gamma}}\{R\}} = 1 - \frac{\mathbb{E}_{\boldsymbol{\gamma}}\{g_n\}}{\mathbb{E}_{\boldsymbol{\gamma}}\{R\}}. \tag{4.8}$$

Definition 8. *We define the* average group loss rate*, denoted by q_0, as*

$$q_0 \triangleq \frac{1}{N}\sum_{n=1}^{N} q_n = 1 - \frac{\mathbb{E}_{\boldsymbol{\gamma}}\{g_{\text{sum}}\}}{N\mathbb{E}_{\boldsymbol{\gamma}}\{R_d + R_s\}}, \tag{4.9}$$

which is the loss rate averaged over all multicast receivers.

4. The Framework for Effective Capacity Optimization

We use the effective capacity (see Chapter II for the detailed definition of effective capacity) for multicast as the main performance metric and explain the QoS exponent θ as the delay-QoS constraint. Moreover, we use a *loss-rate threshold*, denoted by q_{th}, $0 \leq q_{\text{th}} < 1$, to characterize the tolerable loss-rate threshold, where $q_n \leq q_{\text{th}}$ needs to be satisfied for all $n = 1, 2, \ldots, N$. The rationale of setting a tolerable loss level in multicast transmissions is explained as follows. For delay-sensitive services, some data loss is usually acceptable in order to meet the delay-QoS constraint. Moreover, for multicast services, a certain amount of redundancy is often injected into upper layer data (e.g., the application layer) to combat data loss caused by the heterogenous channel qualities across different multicast receivers [67]- [68]. Therefore, it is reasonable to setup a tolerable loss-rate threshold, depending on the specific requirements from various applications. Given the above delay and loss QoS constraints, we focus on developing the optimal adaptive multicast transmission schemes to maximize the

effective capacity of multicast transmissions over the wireless fading channels. It can be expected that more stringent QoS requirements, such as the larger θ and smaller q_{th}, will degrade the effective capacity, implying the existence of fundamental tradeoffs between effective capacity and these QoS metrics. Then, we identify the optimal tradeoff and derive the optimal adaptive multicast schemes by solving the following optimization problems:

$$\textbf{IV-}\boldsymbol{P1}: \quad \text{TS-based effective capacity maximization:}$$

$$\max_{(\boldsymbol{\lambda},R_d)} \left\{\mathscr{C}(\theta)\right\} = \max_{(\boldsymbol{\lambda},R_d)} \left\{-\frac{1}{\theta}\log\left(\mathbb{E}_{\boldsymbol{\gamma}}\left\{e^{-\theta R}\right\}\right)\right\}$$

$$\begin{aligned} \text{s.t.:} \quad & (i) \quad \lambda_n \geq 0,\ n=1,2,\ldots,N,\ \textstyle\sum_{n=1}^{N}\lambda_n = 1, \quad \forall\, \boldsymbol{\gamma}; \\ & (ii) \quad R_d \geq 0, \quad \forall\, \boldsymbol{\gamma}; \\ & (iii) \quad q_n \leq q_{\text{th}}, \quad \forall\ 1 \leq n \leq N. \end{aligned}$$

$$\textbf{IV-}\boldsymbol{P2}: \quad \text{SPC-based effective capacity maximization:}$$

$$\max_{(\boldsymbol{\mu},R_d)} \left\{\mathscr{C}(\theta)\right\} = \max_{(\boldsymbol{\mu},R_d)} \left\{-\frac{1}{\theta}\log\left(\mathbb{E}_{\boldsymbol{\gamma}}\left\{e^{-\theta R}\right\}\right)\right\}$$

$$\begin{aligned} \text{s.t.:} \quad & (i) \quad \mu_n \geq 0,\ n=1,2,\ldots,N,\ \textstyle\sum_{n=1}^{N}\mu_n = 1, \quad \forall\, \boldsymbol{\gamma}; \\ & (ii) \quad R_d \geq 0, \quad \forall\, \boldsymbol{\gamma}; \\ & (iii) \quad q_n \leq q_{\text{th}}, \quad \forall\ 1 \leq n \leq N, \end{aligned}$$

where R is defined by Eqs. (4.3)-(4.5), q_n is given by Eq. (4.8), and $\mathscr{C}(\theta)$ is the effective capacity of the multicast service-rate process under the specified QoS exponent θ.

D. Optimal TS-Based Adaptive Transmission Policy for Wireless Multicast

1. Effective Capacity Optimization Under the Relaxed Loss-Rate Constraint

It is difficult to solve problem **IV-*P*1** directly due to the complicated loss-rate constraint of **IV-*P*1**. In contrast, we can simplify **IV-*P*1** through the following approach.

Specifically, we derive

$$q_n \leq q_{\text{th}},\ \forall\ n \quad \overset{(a)}{\Longrightarrow} \quad q_0 \leq q_{\text{th}} \tag{4.10}$$

$$\overset{(b)}{\Longleftrightarrow} \quad \mathbb{E}_{\boldsymbol{\gamma}}\Big\{\rho(R_s + R_d) - g_{\text{sum}}\Big\} \leq 0, \tag{4.11}$$

where $\rho \triangleq N(1-q_{\text{th}})$ and (a) and (b) hold by applying Eq. (4.9). The above derivations imply that the constraint given by Eq. (4.11) is weaker than the original loss-rate constraint of **IV-*P*1**. Correspondingly, we call the inequality given by Eq. (4.11) the *relaxed* loss-rate constraint (or the group loss-rate constraint). Next, we replace the original constraint (iii) of problem **IV-*P*1** by the relaxed loss-rate constraint, and then get a new optimization problem **IV-*P*1**-a as follows:

$$\begin{aligned}
\textbf{IV-}\boldsymbol{P}\textbf{1-}a: \quad & \min_{(\boldsymbol{\lambda}, R_d)} \left\{\mathbb{E}_{\boldsymbol{\gamma}}\left\{e^{-\theta(R_s+R_d)}\right\}\right\} \\
\text{s.t.:} \quad & (i) \quad \lambda_n \geq 0,\ n = 1, 2, \ldots, N,\ \textstyle\sum_{n=1}^{N}\lambda_n = 1, \quad \forall\ \boldsymbol{\gamma}; \\
& (ii) \quad R_d \geq 0, \quad \forall\ \boldsymbol{\gamma}; \\
& (iii) \quad \mathbb{E}_{\boldsymbol{\gamma}}\Big\{\rho(R_s + R_d) - g_{\text{sum}}\Big\} \leq 0,
\end{aligned}$$

where the optimal multicast policy of **IV-*P*1**-a is denoted by $(\boldsymbol{\lambda}^\star, R_d^\star)$. Since constraint (iii) of **IV-*P*1** is stronger than the relaxed loss-rate constraint, the feasible solution set for **IV-*P*1** is a subset of that for **IV-*P*1**-a. We will show later that $(\boldsymbol{\lambda}^\star, R_d^\star)$ is also feasible for **IV-*P*1**, which suggests that $(\boldsymbol{\lambda}^\star, R_d^\star)$ is also optimal to **IV-*P*1**. Consequently, we can concentrate on how to derive $(\boldsymbol{\lambda}^\star, R_d^\star)$ rather than to solve **IV-*P*1** directly. In order to obtain $(\boldsymbol{\lambda}^\star, R_d^\star)$, we further formulate the other

problem **IV-*P*1**-b as follows:

$$\begin{aligned}\textbf{IV-}\boldsymbol{P}\textbf{1}\text{-}b: \quad & \min_{(R_s,R_d)} \left\{\mathbb{E}_{\boldsymbol{\gamma}}\left\{e^{-\theta(R_s+R_d)}\right\}\right\} \\ & \text{s.t.:} \quad (i) \quad \mathbb{E}_{\boldsymbol{\gamma}}\Big\{\rho(R_s+R_d)-\widetilde{g}_{\text{sum}}(R_s)\Big\}\leq 0; \\ & \qquad\quad (ii) \quad R_d\geq 0,\ R_s\in\left[r_{\pi(N)},r_{\pi(1)}\right],\ \ \forall\,\boldsymbol{\gamma},\end{aligned}$$

where we denote the optimal solution by (R_s^*, R_d^*) and define

$$\widetilde{g}_{\text{sum}}(R_s) \triangleq \max_{\boldsymbol{\lambda}:\,\lambda_{\pi(i)}\geq 0,\,\forall\, i}\Big\{g_{\text{sum}}\Big\}. \tag{4.12}$$

$$\text{s.t.}: \quad \sum_{i=1}^{N}\lambda_{\pi(i)}r_{\pi(i)}=R_s,\ \ \sum_{i=1}^{N}\lambda_{\pi(i)}=1. \tag{4.13}$$

In the above formulation, we combine constraints (i) and (iii) of **IV-*P*1**-a into constraint (i) of **IV-*P*1**-b through the function $\widetilde{g}_{\text{sum}}(R_s)$. Accordingly, the numbers of optimization variables and constraints are decreased, and thus **IV-*P*1**-b is easier to solve. In Section D-3, we will show how to obtain $(\boldsymbol{\lambda}^\star, R_d^\star)$ through (R_s^*, R_d^*). Before further proceeding, we need to identify the properties of $\widetilde{g}_{\text{sum}}(R_s)$, which play an important role in deriving the optimal TS-based adaptive multicast policy.

2. Properties of $\widetilde{g}_{\text{sum}}(R_s)$

Consider any fading state $\boldsymbol{\gamma}$ and a sending rate $R_s\in[r_{\pi(N)}, r_{\pi(1)}]$. There may exist many different TS policies generating this sending rate. Among all these TS policies, $\widetilde{g}_{\text{sum}}(R_s)$ represents the *maximum* achievable instantaneous sum goodput, which equivalently *minimizes* the total data loss over all multicast receivers. Next, we focus on deriving the analytical expression for the TS policy which achieves $\widetilde{g}_{\text{sum}}(R_s)$. Given any TS policy $\boldsymbol{\lambda}$ satisfying $\sum_{i=1}^{N}\lambda_{\pi(i)}=1$, we can obtain a two-dimensional (2-D) point $A\triangleq(R_s, g_{\text{sum}})$ in the "instantaneous sending rate – instantaneous sum goodput" (ISR-ISG) plane, where the horizontal and vertical axes represent the instantaneous

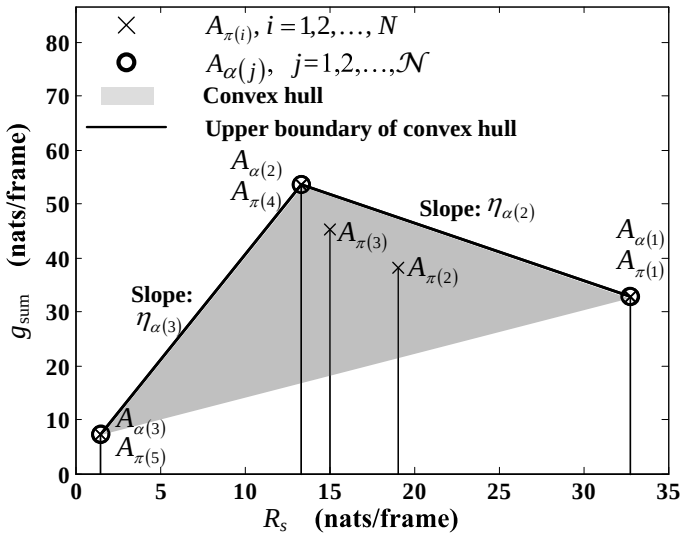


Fig. 13. An example of the convex hull and $\{A_{\alpha(i)}\}_{i=1}^{\mathcal{N}}$ in the ISR-ISG plane, where $N = 5$, $\boldsymbol{\gamma}_\pi = (14.08, 7.56, 5.45, 4.48, -8.03)$ dB, $B = 10^4$ Hz, and $T = 1$ ms.

sending rate and instantaneous sum goodput, respectively. Then, defining N points $\{A_{\pi(i)}\}_{i=1}^{N}$ by

$$A_{\pi(i)} \triangleq \big(r_{\pi(i)}, v_{\pi(i)} r_{\pi(i)}\big) = \big(BT \log(1+\gamma_{\pi(i)}), v_{\pi(i)} BT \log(1+\gamma_{\pi(i)})\big) \tag{4.14}$$

and applying Eqs. (4.3) and (4.7), we can write the point A as a convex combination [77] of $\{A_{\pi(i)}\}_{i=1}^{N}$:

$$A = \sum_{i=1}^{N} \lambda_{\pi(i)} A_{\pi(i)}, \tag{4.15}$$

where $\sum_{i=1}^{N} \lambda_{\pi(i)} = 1$. As a result, the point set yielded by all TS policies with $\sum_{i=1}^{N} \lambda_{\pi(i)} = 1$ form the *convex hull* [77] of $\{A_{\pi(i)}\}_{i=1}^{N}$. Since $A_{\pi(i)}$'s are discrete points, the convex hull must be in the form of a polygon. An example of $\{A_{\pi(i)}\}_{i=1}^{N}$ with $N = 5$ and the corresponding convex hull is illustrated in Fig. 13. For efficient algorithms to determine the vertices of a convex hull, please refer to [78].

Because $\widetilde{g}_{\text{sum}}(R_s)$ is the maximum g_{sum} under the specified R_s, $(R_s, \widetilde{g}_{\text{sum}}(R_s))$ must be on the boundary of the convex hull. Accordingly, we define the *upper boundary* of the convex hull, denoted by $\widetilde{G}$, as $\widetilde{G} \triangleq \left\{(R_s, \widetilde{g}_{\text{sum}}(R_s))|r_{\pi(N)} \leq R_s \leq r_{\pi(1)}\right\}$. The points in $\widetilde{G}$ are called *upper-boundary points*. We denote the convex hull's vertices which are on the upper boundary by $A_{\alpha(j)} = \left(r_{\alpha(j)}, i_{\alpha(j)} r_{\alpha(j)}\right)$, $j = 1, 2, \ldots, \mathcal{N}$, where $\mathcal{N}$ is cardinality of $\widetilde{G}$, the subscript $\alpha(\cdot)$ is the permuting operator such that $r_{\alpha(1)} > r_{\alpha(2)} > \cdots > r_{\alpha(\mathcal{N})}$, and $i_{\alpha(j)}$ represents the number of receivers whose current Shannon capacity is higher than or equal to the transmission rate $r_{\alpha(j)}$. In addition, it is clear that we have $A_{\alpha(1)} = A_{\pi(1)}$ and $A_{\alpha(\mathcal{N})} = A_{\pi(N)}$, and all other $\{A_{\pi(i)}\}_{i=2}^{\mathcal{N}-1}$ must be the vertices lying beyond the straight line $\overline{A_{\alpha(1)} A_{\alpha(\mathcal{N})}}$. Because the upper boundary consists of connected polygon edges of the convex hull, $\widetilde{g}_{\text{sum}}(R_s)$ is a continuous and piecewise linear function, which can be expressed by

$$\widetilde{g}_{\text{sum}}(R_s) = \begin{cases} i_{\alpha(j)} r_{\alpha(j)}, & \text{if } R_s = r_{\alpha(j)},\ 1 \leq j \leq \mathcal{N} \\ i_{\alpha(j)} r_{\alpha(j)} + \eta_{\alpha(j)} \left(R_s - r_{\alpha(j)}\right), & \text{if } r_{\alpha(j)} < R_s < r_{\alpha(j-1)},\ 2 \leq j \leq \mathcal{N}, \end{cases} \tag{4.16}$$

where $\eta_{\alpha(j)}$, $j = 2, 3, \ldots, \mathcal{N}$, is the slope of the straight line $\overline{A_{\alpha(j)} A_{\alpha(j-1)}}$. Based on Eqs. (4.12)-(4.13), it follows immediately that $\widetilde{g}_{\text{sum}}(R_s)$ is a *concave* function over R_s, implying that the slope $\eta_{\alpha(j)}$ strictly increases as j gets larger. In addition, it is easy to verify $\eta_{\alpha(j)} < N$ for all j. For presentational convenience, we define $\eta_{\alpha(1)} \triangleq -\infty$ and $\eta_{\alpha(\mathcal{N}+1)} \triangleq N$, and then get

$$-\infty = \eta_{\alpha(1)} < \eta_{\alpha(2)} < \cdots < \eta_{\alpha(\mathcal{N})} < \eta_{\alpha(\mathcal{N}+1)} = N. \tag{4.17}$$

According to the piecewise linearity of $\widetilde{g}_{\text{sum}}(R_s)$, any TS policy generating an upper-boundary point with $R_s \in [r_{\alpha(j)}, r_{\alpha(j-1)})$ is implemented by time sharing only between $A_{\alpha(j-1)}$ and $A_{\alpha(j)}$. Denoting the time proportion allocated to $A_{\alpha(j-1)}$ and

$A_{\alpha(j)}$ by $\lambda_{\alpha(j-1)}$ and $\lambda_{\alpha(j)}$, respectively. we have

$$\begin{cases} \lambda_{\alpha(j-1)} r_{\alpha(j-1)} + \lambda_{\alpha(j)} r_{\alpha(j)} = R_s; \\ \lambda_{\alpha(j-1)} + \lambda_{\alpha(j)} = 1. \end{cases} \tag{4.18}$$

Solving Eq. (4.18), we obtain

$$\begin{cases} \lambda_{\alpha(j-1)} = \frac{R_s - r_{\alpha(j)}}{r_{\alpha(j-1)} - r_{\alpha(j)}}; \\ \lambda_{\alpha(j)} = \frac{r_{\alpha(j-1)} - R_s}{r_{\alpha(j-1)} - r_{\alpha(j)}}. \end{cases} \tag{4.19}$$

Using Eq. (4.19), we can express the TS policy corresponding to the upper-boundary point as a function of R_s, which is denoted by $\boldsymbol{\lambda}(R_s)$, as follows. For $R_s \in \left[r_{\alpha(j)}, r_{\alpha(j-1)}\right)$, we have

$$\widetilde{\lambda}_{\pi(i)}(R_s) = \begin{cases} \frac{r_{\alpha(j-1)} - R_s}{r_{\alpha(j-1)} - r_{\alpha(j)}}, & \text{if } r_{\pi(i)} = r_{\alpha(j)} \wedge \left(r_{\pi(i)} > r_{\pi(i+1)} \vee i = N\right); \\ \frac{R_s - r_{\alpha(j)}}{r_{\alpha(j-1)} - r_{\alpha(j)}} & \text{if } r_{\pi(i)} = r_{\alpha(j-1)} \wedge \left(r_{\pi(i)} > r_{\pi(i+1)} \vee i = N\right); \\ 0, & \text{otherwise.} \end{cases} \tag{4.20}$$

In Eq. (4.20), $\widetilde{\boldsymbol{\lambda}}_{\pi}(R_s) \triangleq \left(\widetilde{\lambda}_{\pi(1)}(R_s), \widetilde{\lambda}_{\pi(2)}(R_s), \ldots, \widetilde{\lambda}_{\pi(N)}(R_s)\right)^{\tau}$ is the ordered version of $\boldsymbol{\lambda}(R_s)$, where $\widetilde{\lambda}_{\pi(i)}(R_s)$ is the time proportion associated with the transmission rate $BT\log(1+\gamma_{\pi(i)})$.

Obtaining the analytical expression of $\boldsymbol{\lambda}(R_s)$, we further define a useful variable $\widetilde{R}_s$ as

$$\widetilde{R}_s \triangleq \arg\max_{R_s} \left\{\widetilde{g}_{\text{sum}}(R_s)\right\}, \qquad \forall\, \boldsymbol{\gamma}, \tag{4.21}$$

which maximizes the achievable instantaneous sum goodput among all sending rates. Since $\widetilde{g}_{\text{sum}}(R_s)$ is a piecewise linear function, it is not differentiable over the entire domain. Then, we need to introduce the concepts of subgradient and subdifferential [79] in Definition 9 to derive $\widetilde{R}_s$. Further applying Definition 9, we obtain a number of

properties of $\widetilde{R}_s$ which are summarized in Lemma 1.

Definition 9. *Consider a convex function $h : \mathcal{D} \to \mathbb{R}$, where $\mathbb{R}$ denotes the set of real numbers and $\mathcal{D} \subset \mathbb{R}^n$ is a convex set. Then, an $n \times 1$ vector $\boldsymbol{\xi}$, for $\boldsymbol{\xi} \in \mathbb{R}^n$, is called a* subgradient *[79] at $\mathbf{d} \in \mathcal{D}$ if $h(\mathbf{d}') \geq h(\mathbf{d}) + \boldsymbol{\xi}^T(\mathbf{d}' - \mathbf{d})$ for all $\mathbf{d}' \in \mathcal{D}$, where $(\cdot)^T$ represents the transpose. The collection of subgradients at $\mathbf{d}$ form a set called the* subdifferential *[79] of $h(\cdot)$ at $\mathbf{d}$, denoted by $\partial h(\mathbf{d})$. If $h(\cdot)$ is differentiable at $\mathbf{d}$, the subgradient at $\mathbf{d}$ is unique and becomes the gradient. Moreover, the sufficient and necessary condition that $\mathbf{d}^*$ minimizes $h(\mathbf{d})$ is that $\mathbf{0} \in \partial h(\mathbf{d}^*)$. When $h(\cdot)$ is a* concave *function, the subgradient and subdifferential at $h(\mathbf{d})$ is defined in the similar way as the convex case, except that the required inequality becomes $h(\mathbf{d}') \leq h(\mathbf{d}) + \boldsymbol{\xi}^\tau(\mathbf{d}' - \mathbf{d})$ for all $\mathbf{d}' \in \mathcal{D}$.*

Lemma 1. *Given any fading state $\boldsymbol{\gamma}$ and the corresponding point set $\{A_{\alpha(j)}\}_{j=1}^{\mathcal{N}}$, the following Claims hold.*

Claim 1: The subdifferential of $\widetilde{g}_{\text{sum}}(R_s)$ w.r.t. R_s, denoted by $\partial\widetilde{g}_{\text{sum}}(R_s)$, is given by

$$\partial\widetilde{g}_{\text{sum}}(R_s) = \begin{cases} \left\{\eta_{\alpha(j)}\right\}, & \text{if } r_{\alpha(j)} < R_s < r_{\alpha(j-1)} \ \& \ \mathcal{N} \geq j \geq 2; \\ \left[\eta_{\alpha(j)},\ \eta_{\alpha(j+1)}\right], & \text{if } R_s = r_{\alpha(j)} \ \& \ \mathcal{N} \geq j \geq 1. \end{cases} \tag{4.22}$$

Claim 2: There exists the unique k, $1 < k \leq \mathcal{N}$, such that

$$\eta_{\alpha(k+1)} > 0 \geq \eta_{\alpha(k)}, \tag{4.23}$$

If $\eta_{\alpha(k)} < 0$, $\widetilde{R}_s$ is unique and equal to $r_{\alpha(k)}$; if $\eta_{\alpha(k)} = 0$, any R_s within $[r_{\alpha(k)}, r_{\alpha(k-1)}]$ maximizes $\widetilde{g}_{\text{sum}}(R_s)$.

Proof. The proof is provided in Appendix D. □

When $\widetilde{R}_s$ is not unique as addressed in Claim 2, we set $\widetilde{R}_s = r_{\alpha(k)}$ without loss of generality, which also causes the minimum loss among all $R_s \in [r_{\alpha(k)}, r_{\alpha(k-1)}]$. Then,

$\widetilde{R}_s$ can be written through a unified expression as:

$$\widetilde{R}_s = r_{\alpha(k)}, \quad \forall\, \boldsymbol{\gamma} \tag{4.24}$$

where the integer k is the unique solution to Eq. (4.23).

3. Derivations of the Optimal Solutions

Lemma 2 below characterizes the optimal multicast policy of **IV-*P*1**-*a* through (R_s^*, R_d^*) of **IV-*P*1**-*b*.

Lemma 2. *The optimal adaptive multicast policy* $(\boldsymbol{\lambda}^\star, R_d^\star)$ *of* **IV-*P*1**-*a can be obtained by using the optimal solution* (R_s^*, R_d^*) *to* **IV-*P*1**-*b, which is derived as*

$$\begin{cases} \boldsymbol{\lambda}^\star &= \widetilde{\boldsymbol{\lambda}}(R_s^*); \\ R_d^\star &= R_d^* \end{cases}$$

for all $\boldsymbol{\gamma}$*, where* $\widetilde{\boldsymbol{\lambda}}(R_s)$ *is given by Eq. (4.20).*

Proof. The proof is provided in Appendix E. □

In light of Lemma 2, we only need to focus on solving **IV-*P*1**-*b*. Note that: 1) the objective function of **IV-*P*1**-*b* is convex over (R_s, R_d); 2) the constraint function $\mathbb{E}_{\boldsymbol{\gamma}}\{\rho(R_s + R_d) - \widetilde{g}_{\text{sum}}(R_s)\}$ of **IV-*P*1**-*b* is convex over (R_s, R_d) due to the concavity of $\widetilde{g}_{\text{sum}}(R_s)$; 3) constraint (ii) is a linear constraint. Therefore, **IV-*P*1**-*b* is a *convex optimization problem* [77, pp. 136-137], and we can obtain the optimal solution through the Lagrangian method. In particular, the Lagrangian function for problem **IV-*P*1**-*b*, denoted by $L(R_s, R_d; \psi)$, is constructed as

$$L(R_s, R_d; \psi) = \mathbb{E}_{\boldsymbol{\gamma}}\left\{\ell(R_s, R_d; \psi)\right\} \tag{4.25}$$

where

$$\ell(R_s, R_d; \psi) \triangleq e^{-\theta(R_s+R_d)} + \psi\left(\rho(R_s + R_d) - \widetilde{g}_{\text{sum}}(R_s)\right) \tag{4.26}$$

and $\psi \geq 0$ is the Lagrangian multiplier associated with constraint (i) of **IV-*P*1**-b. We denote the subdifferentials of $\ell(R_s, R_d; \psi)$ with respect to (w.r.t.) R_s and R_d by $\partial\ell_{R_s}(R_s, R_d; \psi)$ and $\partial\ell_{R_d}(R_s, R_d; \psi)$, respectively. Then, the optimal (R_s^*, R_d^*) and the optimal Lagrangian multiplier, denoted by ψ^*, are the solution to the following equations [79]:

$$0 \in \partial\ell_{R_s}(R_s^*, R_d^*; \psi^*); \tag{4.27}$$

$$0 \in \partial\ell_{R_d}(R_s^*, R_d^*; \psi^*); \tag{4.28}$$

$$0 = \mathbb{E}_{\boldsymbol{\gamma}}\Big\{\rho(R_s^* + R_d^*) - \widetilde{g}_{\text{sum}}(R_s^*)\Big\}. \tag{4.29}$$

Applying Definition 9 and Eq. (4.16) into Eq. (4.26), we derive

$$\partial\ell_{R_s}(R_s, R_d; \psi) = \begin{cases} \Big\{\psi(\rho - \eta_{\alpha(j)}) - \theta e^{-\theta(R_s+R_d)}\Big\}, & \text{if } r_{\alpha(j)} < R_s < r_{\alpha(j-1)} \ \&\ \mathcal{N} \geq j \geq 2; \\ \Big[\psi(\rho - \eta_{\alpha(j+1)}) - \theta e^{-\theta(R_s+R_d)},\ \psi(\rho - \eta_{\alpha(j)}) - \theta e^{-\theta(R_s+R_d)}\Big], & \\ & \text{if } R_s = r_{\alpha(j)} \ \&\ \mathcal{N} \geq j \geq 1, \end{cases} \tag{4.30}$$

and

$$\partial\ell_{R_d}(R_s, R_d; \psi) = \begin{cases} \Big(-\infty, \psi\rho - \theta e^{-\theta(R_s+R_d)}\Big], & \text{if } R_d = 0; \\ \Big\{\psi\rho - \theta e^{-\theta(R_s+R_d)}\Big\}, & \text{if } R_d > 0. \end{cases} \tag{4.31}$$

Plugging Eqs. (4.30)-(4.31) into Eqs. (4.27)-(4.29) and solving these equations, we obtain the optimal sending rate and pre-drop rate in Theorem 1.

Theorem 1. *The optimal sending rate and pre-drop rate of problem* **IV-*P*1**-b *are*

given by

$$R_s^* = \left(\min\left\{\widehat{R}_s, \widetilde{R}_s\right\}\right)\Big|_{\psi=\psi^*} \qquad \forall\, \boldsymbol{\gamma}, \tag{4.32}$$

and

$$R_d^* = \left[-\frac{1}{\theta}\log\left(\frac{\psi^*\rho}{\theta}\right) - R_s^*\right]^+, \qquad \forall\, \boldsymbol{\gamma}, \tag{4.33}$$

respectively, for all $\boldsymbol{\gamma}$*, where* $[\cdot]^+ = \max\{\cdot, 0\}$*,* $\widehat{R}_s$ *is the solution to*

$$0 \in \partial L_{R_s}(\widehat{R}_s, 0; \psi), \qquad \forall\, \boldsymbol{\gamma}, \tag{4.34}$$

and $\widetilde{R}_s$ *is defined by Eq. (4.23)-(4.24). Moreover, the parameter* $\psi^* > 0$ *is a constant, which is selected such that the group loss rate satisfies:*

$$q_0|_{(R_s,R_d)=(R_s^*,R_d^*)} = q_{\text{th}} \tag{4.35}$$

Proof. The proof is provided in Appendix F. □

Based on Theorem 1, in each fading state the optimal policy can be determined through the following three-step procedures: 1) calculate the values of $\widehat{R}_s$ and $\widetilde{R}_s$ based on Eq. (4.34) and Eqs. (4.23)-(4.24), respectively; 2) set the sending rate equal to the minimum between $\widetilde{R}_s$ and $\widehat{R}_s$; 3) determine the optimal pre-drop rate by using Eq. (4.33). In the above results, $\widehat{R}_s$ and $\widetilde{R}_s$ used in Eq. (4.32) play the major role in determining the optimal sending rate. As mentioned previously, $\widetilde{R}_s$ maximizes the instantaneous sum goodput among all feasible sending rates. Based on Eq. (4.34), $\widehat{R}_s$ is the optimal sending rate on condition that no data is dropped. If $\widehat{R}_s < \widetilde{R}_s$, the system performance is optimized without applying the pre-drop strategy. By contrast, if $\widehat{R}_s > \widetilde{R}_s$, $\widehat{R}_s$ actually provides a service rate higher than $\widetilde{R}_s$ at the cost of

more data loss. In such a case, we need to set $R_s = \widetilde{R}_s$ to avoid unnecessary data loss, which also achieves the largest sum goodput. In the meanwhile, a positive pre-drop rate determined by Eq. (4.33) has to be adopted to compensate for the degradation total service rate. Having derived the optimal solution to **IV-*P*1**-*b*, we then obtain the optimal multicast policy for **IV-*P*1** in the following Theorem 2.

Theorem 2. *Problems* **IV-*P*1** *and* **IV-*P*1**-*a share the same optimal solution given by* $(\boldsymbol{\lambda}^\star, R_d^\star)$*, where* $(\boldsymbol{\lambda}^\star, R_d^\star)$ *is derived by Lemma 2.*

Proof. As discussed in Section D-1, the feasible solution set for **IV-*P*1** is a subset of that for **IV-*P*1**-*a*. Then, to prove Theorem 2 we only need to prove $q_n|_{(\boldsymbol{\lambda},R_d)=(\boldsymbol{\lambda}^\star,R_d^\star)} \leq q_{\text{th}}$ for all n. Since Lemma 2 shows that the policy $(\boldsymbol{\lambda}^\star, R_d^\star)$ can be uniquely characterized by using (R_s^*, R_d^*), we only need to prove

$$q_n|_{(R_s,R_d)=(R_s^*,R_d^*)} \leq q_{\text{th}}, \quad \forall\, n. \tag{4.36}$$

Given any ordered CSI vector $\boldsymbol{\gamma}_\pi$, there are total $N!$ different fading states corresponding to this $\boldsymbol{\gamma}_\pi$. We denote the set of unordered CSI vectors for these $N!$ fading states by $\mathcal{H}(\boldsymbol{\gamma}_\pi)$. Clearly, the point sets $\{A_{\pi(i)}\}_{i=1}^N$ (defined by Eq. (4.14)) in these $N!$ fading states are identical, resulting in the same function $\widetilde{g}_{\text{sum}}(R_s)$ for problem **IV-*P*1**-*b*. Therefore, in these $N!$ fading states **IV-*P*1**-*b* we get the identical (R_s^*, R_d^*). Further noting that the fading channels are i.i.d. across all multicast receivers, we obtain

$$\sum_{\boldsymbol{\gamma}\in\mathcal{H}(\boldsymbol{\gamma}_\pi)} g_1 = \sum_{\boldsymbol{\gamma}\in\mathcal{H}(\boldsymbol{\gamma}_\pi)} g_2 = \cdots = \sum_{\boldsymbol{\gamma}\in\mathcal{H}(\boldsymbol{\gamma}_\pi)} g_N, \qquad \forall\, \boldsymbol{\gamma}, \tag{4.37}$$

and thus

$$\mathbb{E}_{\boldsymbol{\gamma}}\{g_n\} = \mathbb{E}_{\boldsymbol{\gamma}_\pi}\left\{\frac{1}{N!}\sum_{\boldsymbol{\gamma}\in\mathcal{H}(\boldsymbol{\gamma}_\pi)} g_n\right\}, \quad \forall\, n = 1, 2, \ldots, N. \tag{4.38}$$

Eqs. (4.37) and (4.38) imply that under the policy $(\boldsymbol{\lambda}^\star, R_d^\star)$, we have

$$\mathbb{E}_{\boldsymbol{\gamma}}\Big\{g_1\Big\} = \mathbb{E}_{\boldsymbol{\gamma}}\Big\{g_2\Big\} = \cdots = \mathbb{E}_{\boldsymbol{\gamma}}\Big\{g_N\Big\}.$$

Since all multicast receivers have the same average goodput, their loss rates are equal, i.e., $q_1 = q_2 = \cdots = q_N$. Applying Eqs. (4.9) and (4.35), we get $q_n = q_0 = q_{\text{th}}$ for all n, which verifies Eq. (4.36), and thus Theorem 2 follows. □

E. The Optimal TS-Based Multicasting Policy under the Limiting Scenarios of Delay-QoS Constraints

While section D developed the general form of the optimal TS-based multicast policy with $\theta > 0$, in this section we aim at deriving the optimal policies to **IV-*P*1** under the special cases with $\theta \to 0$ and $\theta \to \infty$, respectively. As mentioned in Chapter II, $\theta \to 0$ corresponds to the scenario which can tolerate infinite delay, while any delay is intolerable as $\theta \to \infty$. Since the optimal multicast policy can be completely characterized by the solution (R_s^*, R_d^*) to **IV-*P*1**-b based on Lemma 2 and Theorem 2, we mainly focus on the characteristics of (R_s^*, R_d^*) in this section.

1. The Limiting Case of $\theta \to 0$

As $\theta \to 0$, there is no delay constraint imposed to the multicast transmissions. Accordingly, the effective capacity reduces to the average throughput of the service process [53]. Thus, problem **IV-*P*1** is to maximize the average multicast throughput given the loss-rate constraint. We first define

$$\widetilde{q}_0 = q_0\big|_{(R_s,R_d)=\left(\widetilde{R}_s,0\right)}. \tag{4.39}$$

We will then derive (R_s^*, R_d^*) with $q_{\rm th} < q_0$ and $q_{\rm th} \geq q_0$ in Theorems 3 and Theorem 4, respectively.

Theorem 3. *As $\theta \to 0$, the optimal solution (R_s^*, R_d^*) to* **IV-*P*1**-*b under the constraint $q_{\rm th} < \widetilde{q}_0$ reduces to the following form:*

$$R_d^* = 0; \tag{4.40}$$

$$R_s^* = \begin{cases} r_{\alpha(\overline{k})}, & \text{if } \eta_{\rm th} > \eta_{\alpha(\overline{k})}; \\ \varsigma_{\boldsymbol{\gamma}} r_{\alpha(\overline{k})} + (1-\varsigma_{\boldsymbol{\gamma}}) r_{\alpha(\overline{k}-1)}, & \text{if } \eta_{\rm th} = \eta_{\alpha(\overline{k})}. \end{cases} \tag{4.41}$$

for all $\boldsymbol{\gamma}$, where $\varsigma_{\boldsymbol{\gamma}} \in [0,1]$ and the integer $\overline{k}$ in each fading state is uniquely determined by the following inequality:

$$\eta_{\alpha(\overline{k}+1)} > \eta_{\rm th} \geq \eta_{\alpha(\overline{k})}. \tag{4.42}$$

In the above equations, $\eta_{\rm th} \geq 0$ is a constant parameter over all fading states but $\varsigma_{\boldsymbol{\gamma}}$ may vary with $\boldsymbol{\gamma}$. Moreover, $\eta_{\rm th} \geq 0$ and $\varsigma_{\boldsymbol{\gamma}}$ need to be selected such that the equation $q_0|_{(R_s,R_d)=(R_s^*,R_d^*)} = q_{\rm th}$ holds.

Proof. In order to prove Theorem 3, we first introduce Lemma 3 in the following. The rest of the proof for Theorem 3 is provided in Appendix G. □

Lemma 3. *Denote by $q_0(\eta_{\rm th}, \varsigma_{\boldsymbol{\gamma}})$ the group loss rate attained under the multicast policy characterized by Eqs. (4.40)-(4.42). For any $(\eta'_{\rm th}, \varsigma'_{\boldsymbol{\gamma}})$ and $(\eta''_{\rm th}, \varsigma''_{\boldsymbol{\gamma}})$ with $\eta'_{\rm th} > \eta''_{\rm th} > 0$, the inequality*

$$q_0(\eta'_{\rm th}, \varsigma'_{\boldsymbol{\gamma}}) \leq q_0(\eta''_{\rm th}, \varsigma''_{\boldsymbol{\gamma}}) \tag{4.43}$$

holds. Moreover, we have

$$\begin{cases} \lim_{\eta_{\rm th} \to N} q_0(\eta_{\rm th}, \varsigma_{\boldsymbol{\gamma}}) = 0; \\ \lim_{\eta_{\rm th} \to N} (R_s^*, R_d^*) = (r_{\pi(N)}, 0) \end{cases} \tag{4.44}$$

and

$$\begin{cases} \lim_{\eta_{\rm th}\to 0} q_0(\eta_{\rm th}, 1) &= \widetilde{q}_0; \\ \lim_{\eta_{\rm th}\to 0}(R_s^*, R_d^*) &= (\widetilde{R}_s, 0). \end{cases} \tag{4.45}$$

Proof. The proof of Lemma 3 is provided in Appendix H. □

If $\boldsymbol{\gamma}$ has continuous CDF, the probability $\Pr\left\{\exists\, \overline{k},\ \text{s.t.: } \eta_{\rm th} = \eta_{\alpha(\overline{k})}\right\}$ is equal to zero for any $\eta_{\rm th}$, suggesting that how to select $\varsigma_{\boldsymbol{\gamma}}$ does not affect the group loss rate and the average throughput. Consequently, when $\boldsymbol{\gamma}$ has continuous CDF, such as typical Rayleigh, Rician, and Nakagami-m channel fading models, we can set $\varsigma_{\boldsymbol{\gamma}} = 1$ without loss of generality. Eqs. (4.41) and (4.42) suggest that $\eta_{\rm th}$ is a subgradient of $\widetilde{g}_{\rm sum}(R_s)$ at $R_s = R_s^*$ for all $\boldsymbol{\gamma}$. Consequently, $\eta_{\rm th}$ can be interpreted as the increase rate of the average sum goodput w.r.t. the average throughput. Since $\widetilde{g}_{\rm sum}(R_s)$ is a concave function, the lower $\eta_{\rm th}$ corresponds to the higher loss rate, as demonstrated in Lemma 3.

As $\theta \to 0$, our effective-capacity optimization framework is similar to the problem studied in our previous work [76], where we aimed at maximizing the average multicast throughput under the specified loss-rate constraint. However, in [76] we did not incorporate the pre-drop strategy. Despite the differences between the two frameworks, it turns out that the optimal solution derived in Theorem 3 reduces to the multicast policy derived in [76]. This is expected based on the following reasons. When $\theta \to 0$, the multicast receivers can tolerate infinite long delay. Therefore, there is no need to drop data from the queue to decrease the queuing delay. Lemma 3 further suggests that given $q_{\rm th} \to \widetilde{q}_0$, the optimal policy converges to $(R_s^*, R_d^*) = (\widetilde{R}_s, 0)$, which already achieves the maximum average goodput with $q_0 = \widetilde{q}_0$. Then, when the tolerable loss-level gets lower (i.e., $q_{\rm th} < \widetilde{q}_0$), the only way to meet the loss-rate con-

straint is to decrease R_s, such that more multicast receivers can successfully decode the data in each fading state. Since the pre-drop strategy clearly cannot help suppress the loss rate, we need to set R_d^* equal to zero for this scenario. Following the above discussions, the multicast policy derived in [76] is a special case of our derived optimal policies for the multicast effective-capacity optimization framework. Lemma 3 also shows that $R_s^* \to r_{\pi(N)}$ as $q_{\text{th}} \to 0$ (see Eq. (4.44)), implying that the sending rate is determined by the worst-case SNR among all multicast receivers under $q_{\text{th}} = 0$.

Theorem 4. *As $\theta \to 0$, the optimal solution (R_s^*, R_d^*) to* **IV-*P*1-*b*** *under the constraint $q_{\text{th}} \geq \widetilde{q}_0$ is determined by*

$$(R_s^*, R_d^*) \;=\; \left(\widetilde{R}_s, [\kappa - \widetilde{R}_s]^+\right), \quad \forall \boldsymbol{\gamma}, \tag{4.46}$$

where κ is a constant and is selected such that $q_0|_{(R_s,R_d)=(R_s^*,R_d^*)} = q_{\text{th}}$ holds.

Proof. The proof of Theorem 4 is provided in Appendix I. □

Eq. (4.46) shows that the sending rate is always set equal to $\widetilde{R}_s$, and the nonzero pre-drop rate will be applied in some fading states. The above strategy is also expected based on the following arguments. Recall that if $q_{\text{th}} \to \widetilde{q}_0$, the optimal policy converges to $(R_s^*, R_d^*) = (\widetilde{R}_s, 0)$, which maximizes the average sum goodput. When the higher loss rate is tolerable, we can further increase the service rate. However, a sending rate higher than $\widetilde{R}_s$ will decrease the sum goodput. In contrast, increasing the pre-drop rate does not affect the sum goodput. As a result, the best strategy is to increase the pre-drop rate while letting R_s^* stay at $\widetilde{R}_s$ to avoid unnecessary data loss. Also note that this case is mainly for the mathematical completeness. For a multicast session without delay constraints, the loss rate threshold usually will not be set larger than $\widetilde{q}_0$ and the pre-drop strategy will not be applied. This is because the purpose of the pre-drop strategy is to decrease the queuing delay. Thus, when

there is no delay constraint, we do not need to improve the throughput at the cost of pure data loss caused by the pre-drop strategy.

2. The Limiting Case of $\theta \to \infty$

Theorem 5. *As $\theta \to \infty$, the optimal solution (R_s^*, R_d^*) to* **IV**-***P*1**-*b is determined by*

$$(R_s^*, R_d^*) = \begin{cases} (\widetilde{R}_s, \xi - \widetilde{R}_s), & \text{if } \widetilde{R}_s < \xi; \\ (\xi, 0), & \text{if } R_\rho \le \xi \le \widetilde{R}_s; \\ (R_\rho, 0), & \text{if } \xi < R_\rho \end{cases} \tag{4.47}$$

for all $\boldsymbol{\gamma}$, where ξ is a constant, which is selected such that $q_0|_{(R_s,R_d)=(R_s^*,R_d^*)} = q_{\text{th}}$ holds. In Eq. (4.47), R_ρ is given by

$$R_\rho = \begin{cases} r_{\alpha(\widehat{k})}, & \text{if } \eta_{\text{th}} > \eta_{\alpha(\widehat{k})}; \\ \chi_{\boldsymbol{\gamma}} r_{\alpha(\widehat{k})} + (1-\chi_{\boldsymbol{\gamma}}) r_{\alpha(\widehat{k}-1)}, \text{ for certain } \chi_{\boldsymbol{\gamma}} \in [0,1], & \text{if } \eta_{\text{th}} = \eta_{\alpha(\widehat{k})}, \end{cases} \tag{4.48}$$

where $\widehat{k}$ in each fading state is uniquely derived by solving

$$\eta_{\alpha(\widehat{k}+1)} > N(1-q_{\text{th}}) \ge \eta_{\alpha(\widehat{k})}, \tag{4.49}$$

and $\chi_{\boldsymbol{\gamma}}$ can be any real number within $[0,1]$.

Proof. The proof of Theorem 5 is provided in Appendix J. □

As $\theta \to \infty$, any queueing delay is intolerable. Accordingly, the effective capacity is determined by the minimum service rate over all fading states [53]. Theorem 5 shows that the total service rate (the sum of sending rate and pre-drop rate) is lower bounded by ξ. Clearly, given $q_{\text{th}} \neq 0$, ξ is a positive real number. Therefore, our derive adaptive multicast transmission policy can achieve a nonzero effective capacity even when the delay-QoS requirement is extremely stringent. This also demonstrates the importance of the pre-drop strategy, which can guarantee a departure rate for the

queue even when the channel is in deep fade status. The optimal TS policy under each specific case given in Eq. (4.47) will be further explained as follows.

a. If $\widetilde{R}_s < \xi$

With the similar arguments in the discussions for Theorem 4, in this case we need to set $R_s^* = \widetilde{R}_s$ to avoid unnecessary data loss; on the other hand, some data has to be dropped to guarantee the minimum service rate ξ.

b. If $\xi < R_\rho$

An interesting observation in this case is that the total service rate R_ρ is larger than the minimum requirement ξ, which is expected because of the following reasons. Note that $N(1 - q_{\text{th}})$ is a subgradient of $\widetilde{g}_{\text{sum}}(R_s)$ at $R_s = R_\rho$. Then, the increase rate of $\widetilde{g}_{\text{sum}}(R_s)$ w.r.t. R_s is always larger than $N(1-q_{\text{th}})$ for $R_s \leq R_\rho$ based on the concavity of $\widetilde{g}_{\text{sum}}(R_s)$. Because $N(1 - q_{\text{th}})$ is just the ratio of the average sum goodput to the average throughput under the optimal policy, setting $R_s = R_\rho$ maximizes the amount of correctly received data without degrading the loss-QoS satisfaction. In addition, since $R_s = R_\rho$ already exceeds the minimum required service rate, the data-drop operation is not needed.

c. If $R_\rho \leq \xi \leq \widetilde{R}_s$

Under this condition, we have to set $R_s = \xi$ to meet the delay constraint, although the subgradient of $\widetilde{g}_{\text{sum}}(R_s)$ becomes lower than $N(1-q_{\text{th}})$ for $R_s > R_\rho$, which causes negative impact to meet the loss-rate threshold q_{th}. Also, because the minimum required service rate has been satisfied, pre-drop rate will be set to zero to prevent further data loss.

F. Optimal SPC-Based Adaptive Multicast Transmission Policy

Unlike TS-based multicast, which controls time-slot lengths for diverse transmission rates in every fading states, SPC-based multicast uses dynamic power allocation to adapt the transmission rates to the heterogenous CSI across multicast receivers. However, we can solve the multicast effective-capacity optimization problems in these two scenarios through the similar way. In particular, the following the strategy used in Section D-1, we formulate problems **IV-*P2*-*a*** and **IV-*P2*-*b*** for SPC-based multicast as follows:

$$\begin{aligned}
\textbf{IV-}\boldsymbol{P2}\textbf{-}a: \quad & \min_{(\boldsymbol{\mu}, R_d)} \left\{ \mathbb{E}_{\boldsymbol{\gamma}} \left\{ e^{-\theta(R_s+R_d)} \right\} \right\} \\
& \text{s.t.:} \quad (i) \quad R_d \geq 0, \ \forall \boldsymbol{\gamma}; \\
& \qquad\quad (ii) \quad \mathbb{E}_{\boldsymbol{\gamma}} \Big\{ \rho(R_s + R_d) - g_{\text{sum}} \Big\} \leq 0; \\
& \qquad\quad (iii) \quad \mu_n \geq 0, \ n = 1, 2, \ldots, N, \ \textstyle\sum_{n=1}^{N} \mu_n = 1, \quad \forall \boldsymbol{\gamma}.
\end{aligned}$$

$$\begin{aligned}
\textbf{IV-}\boldsymbol{P2}\textbf{-}b: \quad & \min_{(R_s, R_d)} \left\{ \mathbb{E}_{\boldsymbol{\gamma}} \left\{ e^{-\theta(R_s+R_d)} \right\} \right\} \\
& \text{s.t.:} \quad (i) \quad \mathbb{E}_{\boldsymbol{\gamma}} \Big\{ \rho(R_s + R_d) - \widetilde{g}_{\text{sum}}(R_s) \Big\} \leq 0; \\
& \qquad\quad (ii) \quad R_d \geq 0, \ R_s \in \left[r_{\pi(N)}, r_{\pi(1)} \right], \quad \forall \ \boldsymbol{\gamma},
\end{aligned}$$

where

$$\widetilde{g}_{\text{sum}}(R_s) \triangleq \max_{\boldsymbol{\mu}: \, \mu_{\pi(i)} \geq 0, \, \forall \, i} \Big\{ g_{\text{sum}} \Big\} \tag{4.50}$$

$$\text{s.t.:} \qquad \sum_{i=1}^{N} r_{\pi(i)} = R_s, \quad \sum_{i=1}^{N} \mu_{\pi(i)} = 1. \tag{4.51}$$

Note that **IV-*P2*-*a*** and **IV-*P2*-*b*** for SPC-based multicast play the same roles as **IV-*P2*-*a*** and **IV-*P2*-*b*** for TS-based multicast, respectively. Specifically, **IV-*P2*-*a*** replaces constraint (*iii*) of **IV-*P2*** by using the relaxed loss-rate constraint. Problem

Table I. Greedy Algorithm to Determine $\widetilde{\boldsymbol{\mu}}(\varrho)$ under CSI $\boldsymbol{\gamma}$.

Step 1:	Calculate utility function:
	$U_{\pi(i)}(p) = \frac{(i-\varrho)\gamma_{\pi(i)}}{1+p\gamma_{\pi(i)}}$, for all $1 \leq i \leq N$.
Step 2:	Determine $U_{\pi}^{*}(p) = \max_{\varrho \leq i \leq N} \{U_{\pi(i)}(p)\}$.
Step 3:	Identify $M_{\pi(i)} = \{p \in [0,1] : U_{\pi(i)}(p) = U_{\pi}^{*}(p)\}$.
Step 4:	Obtain $\widetilde{\mu}_{\pi(i)}(\varrho) = \int_{M_{\pi(i)}} dp$.
Step 5:	End.

IV-*P2*-*b* combines constraints (i) and (iii) of **IV-*P2*-*a*** by using $\widetilde{g}_{\text{sum}}(R_s)$ to reduce the number of optimization variables. Accordingly, we can focus on problems **IV-*P2*-*b*** instead of solving problem **IV-*P2*** directly. Also, $\widetilde{g}_{\text{sum}}(R_s)$ defined by Eqs. (4.50)-(4.51) functions similarly as its counterpart of TS-based multicast. Among all SPC-based policies with the given sending rate R_s in a fading state, $\widetilde{g}_{\text{sum}}(R_s)$ represents the maximum achievable instantaneous sum goodput. The derivation and properties of $\widetilde{g}_{\text{sum}}(R_s)$ will be elaborated on in Section F-1 below.

1. Derivation of $\widetilde{g}_{\text{sum}}(R_s)$ and Its Properties

The Lagrangian function, denoted by $\nu(\boldsymbol{\mu}; \varrho, R_s)$, for the optimization problem formulated in Eqs. (4.50)-(4.51) is constructed as

$$\begin{aligned} \nu(\boldsymbol{\mu}; \varrho, R_s) &= \sum_{i=1}^{N} i r_{\pi(i)} + \varrho \left(R_s - \sum_{i=1}^{N} r_{\pi(i)} \right) \\ &= \sum_{i=1}^{w} (i - \varrho)\, r_{\pi(i)} + \varrho R_s, \end{aligned} \tag{4.52}$$

where $\sum_{i=1}^{N} \mu_{\pi(i)} = 1$. In the above equation, ϱ is the Lagrangian multiplier associated with the constraint $\sum_{i=1}^{N} r_{\pi(i)} = R_s$ given in Eq. (4.51). The Lagrangian dual function [77, 79], denoted by $z(\varrho; R_s)$, is then given by

$$z(\varrho; R_s) \triangleq \max_{\boldsymbol{\mu}:\ \sum_{i=1}^{N} \mu_{\pi(i)} = 1,\, \mu_{\pi(i)} \geq 0, \forall\, i} \Big\{ \nu(\boldsymbol{\mu}; \varrho, R_s) \Big\}. \tag{4.53}$$

Based on Eq. (4.52), the maximization results does not vary with R_s when ϱ is specified. We express the maximizer as a function of ϱ, which is denoted by $\widetilde{\boldsymbol{\mu}}(\varrho)$. Eq. (4.52) further implies that maximizing $\nu(\boldsymbol{\mu};\varrho,R_s)$ is equivalent to maximizing the sum $\sum_{i=1}^{N}(i-\varrho)r_{\pi(i)}$, which has the same form as the problem to optimize the weighted sum capacity over broadcast channels [13]. Accordingly, we can derive $\widetilde{\boldsymbol{\mu}}(\varrho)$ by using the *Greedy algorithm* developed by the authors of [13,14]. For completeness of this chapter, we describe the greedy algorithm in Table I. Please refer to [13] for the detailed information. The power-allocation vector $\widetilde{\boldsymbol{\mu}}(\varrho)$ is unique according to the results in [13]. Then, we denote the instantaneous sending rate and the instantaneous sum goodput achieved under $\widetilde{\boldsymbol{\mu}}(\varrho)$ by $\mathcal{R}_s(\varrho)$ and $\mathcal{G}_{\text{sum}}(\varrho)$, respectively, which are expressed by

$$\begin{cases} \mathcal{R}_s(\varrho) & \triangleq \left(\sum_{i=1}^{N} r_{\pi(i)}\right)\Big|_{\boldsymbol{\mu}=\widetilde{\boldsymbol{\mu}}(\varrho)}; \\ \mathcal{G}_{\text{sum}}(\varrho) & \triangleq g_{\text{sum}}\Big|_{\boldsymbol{\mu}=\widetilde{\boldsymbol{\mu}}(\varrho)} = \left(\sum_{i=1}^{N} i r_{\pi(i)}\right)\Big|_{\boldsymbol{\mu}=\widetilde{\boldsymbol{\mu}}(\varrho)}. \end{cases} \tag{4.54}$$

Correspondingly, we can rewrite $z(\varrho;R_s)$ as:

$$z(\varrho;R_s) = \mathcal{G}_{\text{sum}}(\varrho) + \varrho(R_s - \mathcal{R}_s(\varrho)), \quad \forall\, \varrho. \tag{4.55}$$

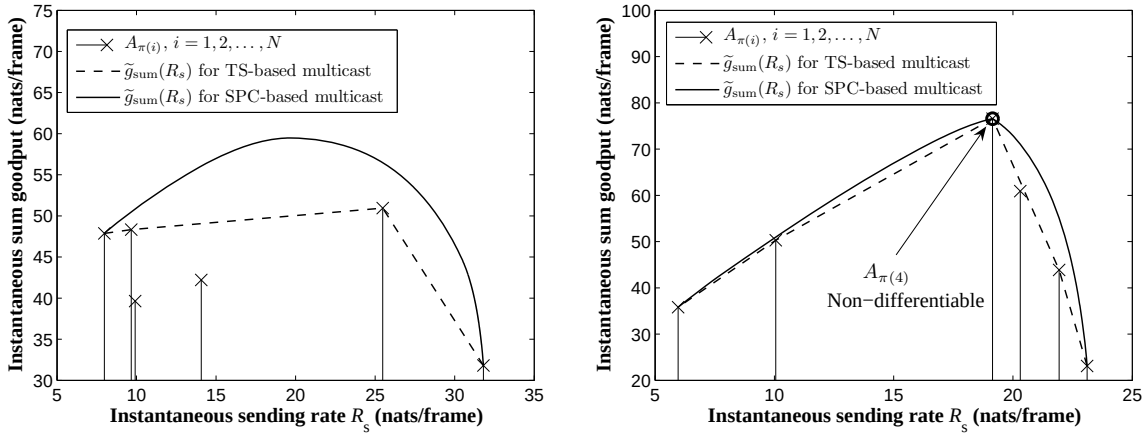
After obtaining $\widetilde{\boldsymbol{\mu}}(\varrho)$, Lemma 4 below solves for the expression of $\widetilde{g}_{\text{sum}}(R_s)$ of SPC-based multicast transmissions.

Lemma 4. *Given $R_s \in [BT\log(1+\gamma_{\pi(N)}), BT\log(1+\gamma_{\pi(1)})]$, there exists the real number $\overline{\varrho}$ such that*

$$\mathcal{R}_s(\overline{\varrho}) = R_s. \tag{4.56}$$

Then, the maximizer for Eqs. (4.50)-(4.51) is equal to $\widetilde{\boldsymbol{\mu}}(\overline{\varrho})$ and we have

$$\widetilde{g}_{\text{sum}}(R_s) = z(\overline{\varrho};R_s). \tag{4.57}$$



(a) (b)

Fig. 14. Instantaneous sum goodput $\widetilde{g}_{\rm sum}(R_s)$ versus instantaneous sum sending rate R_s for TS-based multicast and SPC-based multicast in example fading states, where $N = 6$. (a) The case with $\boldsymbol{\gamma} = (13.63, 10.71, 4.89, 2.29, 2.12, 0.87)$ dB. (b) The case with $\boldsymbol{\gamma} = (9.58, 9.01, 8.21, 7.62, 2.39, -0.88)$ dB.

Proof. The proof of Lemma 4 is provided in Appendix K. □

Figure 14 plots $\widetilde{g}_{\rm sum}(R_s)$'s for TS-based and SPC-based multicast, respectively, in the example fading states. As shown in Fig. 14, $\widetilde{g}_{\rm sum}(R_s)$ of the SPC-based multicast is always larger than or equal to that of the TS-based multicast. Then, it can be expected that the SPC-based multicast outperforms the TS-based multicast in terms of the effective-capacity performance. Further note that $\widetilde{g}_{\rm sum}(R_s)$ of the SPC-based multicast may be also non-differentiable at some point. An example is illustrated in Fig. 14(b), where $\widetilde{g}_{\rm sum}(R_s)$ is not differentiable at $A_{\pi(4)}$, which is marked with a hollow circle. Thus, we need to derive the subdifferential of $\widetilde{g}_{\rm sum}(R_s)$, which will be used to characterize the optimal solution of problem **IV-*P*2-*b***. Lemma 5 below proves the concavity of $\widetilde{g}_{\rm sum}(R_s)$ and derives the corresponding subdifferential.

Lemma 5. *For SPC-based multicast transmissions, $\widetilde{g}_{\rm sum}(R_s)$ is a concave function*

of R_s *within* $R_s \in [BT\log(1+\gamma_{\pi(N)}), BT\log(1+\gamma_{\pi(1)})]$. *Its subdifferential* $\partial\widetilde{g}_{\text{sum}}(R_s)$ *is given by*

$$\partial\widetilde{g}_{\text{sum}}(R_s) = \{\,\overline{\varrho}\,|\mathcal{R}_s\,(\overline{\varrho}) = R_s\}\,, \tag{4.58}$$

where $\mathcal{R}_s\,(\overline{\varrho})$ *is defined by Eq. (4.54).*

Proof. The proof of Lemma 5 is provided in Appendix L. □

We also use $\widetilde{R}_s$ to denote the maximizer of $\widetilde{g}_{\text{sum}}(R_s)$ as defined in Eq. (4.21). Because $\widetilde{g}_{\text{sum}}(R_s)$ for SPC-based multicast still concave, $0 \in \partial\widetilde{g}_{\text{sum}}(\widetilde{R}_s)$ must hold [79, pp. 128]. Further applying Lemma 5, we get

$$\widetilde{R}_s = \mathcal{R}_s(0). \tag{4.59}$$

2. The Optimal Solutions Obtained by Applying $\widetilde{g}_{\text{sum}}(R_s)$

Problem **IV-*P2*-*b*** can be also solved through the Lagrangian method. The Lagrangian of problem $P2$-b, denoted by $W(R_s, R_d; \phi)$, is constructed by

$$W(R_s, R_d; \phi) = \phi\big(\mathbb{E}\left\{\rho(R_s+R_d)\right\} - \mathbb{E}\left\{\widetilde{g}_{\text{sum}}(R_s)\right\}\big) + \mathbb{E}\left\{e^{-\theta(R_s+R_d)}\right\} = \mathbb{E}\left\{w\right\}, \tag{4.60}$$

where

$$w(R_s, R_d; \phi) \triangleq \phi\rho(R_s+R_d) - \phi\widetilde{g}_{\text{sum}}(R_s) + e^{-\theta(R_s+R_d)}. \tag{4.61}$$

and $\phi \geq 0$ is the Lagrangian multiplier associated with the group loss-rate constraint. Clearly, $W(R_s, R_d; \phi)$ is a convex function. Then, the optimal multicast policy of

IV-*P*2-*b*, denoted by (R_s^*, R_d^*), is the solution to the following equations:

$$\begin{cases} 0 \in \partial w_{R_s}(R_s^*, R_d^*; \phi^*); \\ 0 \in \partial w_{R_d}(R_s^*, R_d^*; \phi^*); \\ 0 = \mathbb{E}_{\boldsymbol{\gamma}}\Big\{\rho(R_s^* + R_d^*) - \widetilde{g}_{\rm sum}(R_s^*)\Big\}, \end{cases} \tag{4.62}$$

where $\partial w_{R_s}(R_s, R_d; \phi^*)$ and $\partial w_{R_s}(R_s, R_d; \phi^*)$ are the subdifferentials of $w(R_s, R_d; \phi^*)$ w.r.t. R_s and R_d, respectively. Based on Eqs. (4.61) and (4.58), we derive

$$\partial w_{R_d}(R_s^*, R_d^*; \phi^*) = \begin{cases} \left(-\infty, \phi^*\rho - \theta e^{-\theta(R_s^*+R_d^*)}\right], & \text{if } R_d^* = 0; \\ \left\{\phi^*\rho - \theta e^{-\theta(R_s^*+R_d^*)}\right\}, & \text{if } R_d^* > 0; \end{cases} \tag{4.63}$$

$$\partial w_{R_s}(R_s, R_d; \phi) = \Big\{u \,|\, u = \phi^*\rho - \theta e^{-\theta(R_s^*+R_d^*)} - \phi^*\widehat{\varrho},\ \forall\, \widehat{\varrho} \in \partial\widetilde{g}_{\rm sum}(R_s)\Big\}. \tag{4.64}$$

Plugging Eqs. (4.63)-(4.64) into Eq. (4.62) and solving these equations, we derive the optimal solution to **IV-*P*2-*b*** and summarize it in Theorem 6. Through (R_s^*, R_d^*), the optimal solutions of **IV-*P*2** and **IV-*P*2-*a*** is also derived in Theorem 6.

Theorem 6. *For SPC-based multicast transmissions, the following Claims hold.*

Claim 1: The optimal solution of **IV-*P*2-*b*** *is given by*

$$\begin{cases} R_s^* = \mathcal{R}_s\left([\widehat{\varrho}]^+\right); \\ R_d^* = \left[-\dfrac{1}{\theta}\log\left(\dfrac{\phi^*\rho}{\theta}\right) - R_s^*\right]^+ \end{cases} \tag{4.65}$$

for all $\boldsymbol{\gamma}$, where $\widehat{\varrho}$ in a given fading state is the solution to

$$\phi^*(\rho - \widehat{\varrho}) - \theta e^{-\theta\mathcal{R}_s(\widehat{\varrho})} = 0. \tag{4.66}$$

In the above equations, $\mathcal{R}_s(\widehat{\varrho})$ represents the sending rate under the policy $\widetilde{\boldsymbol{\mu}}(\widehat{\varrho})$, which is generated by the greedy algorithm given in Table I. Moreover, ϕ^* is a constant which needs selected such that $q_0|_{(R_s,R_d)=(R_s^*,R_d^*)} = q_{\rm th}$ holds.

Claim 2: Problems **IV-*P*2** *and* **IV-*P*2-*a*** *share the same optimal SPC-based multicast*

policy, which is given by

$$\begin{cases} \boldsymbol{\mu}^{\star} &= \widetilde{\boldsymbol{\mu}}\left([\widehat{\varrho}]^{+}\right); \\ R_d^{\star} &= R_d^{*} \end{cases} \tag{4.67}$$

for all $\boldsymbol{\gamma}$.

Proof. Note that the Lagrangian function of **IV-*P2*-*b***, which is given by Eqs. (4.60)-(4.61), has the same form as that of **IV-*P2*-*a***, which is characterized by Eqs. (4.25)-(4.25). Moreover, the function $\widetilde{g}_{\text{sum}}(R_s)$ is concave in both of these two cases. Then, following the same procedures used in the proof for Theorem 1, we obtain

$$\begin{cases} R_s^{*} &= \min\left\{\widetilde{R}_s, \widehat{R}_s\right\}; \\ R_d^{*} &= \left[-\frac{1}{\theta}\log\left(\frac{\phi^{*}\rho}{\theta}\right) - R_s^{*}\right]^{+}. \end{cases} \tag{4.68}$$

where $\widehat{R}_s$ is the solution to $0 \in \partial w_{R_s}(R_s^{*}, 0; \phi^{*})$. Applying Eq. (4.58) into Eq. (4.64), we can see that the condition $0 \in \partial w_{R_s}(R_s^{*}, 0; \phi^{*})$ is equivalent to finding a $\widehat{\varrho}$ to satisfy Eq. (4.66). With such a $\widehat{\varrho}$, we get $\widehat{R}_s = \mathcal{R}_s(\widehat{\varrho})$. Comparing $\widetilde{R}_s = \mathcal{R}_s(0)$ (see Eq. (4.59)) with $\widehat{R}_s = \mathcal{R}_s(\widehat{\varrho})$, we have

$$\min\left\{\widetilde{R}_s, \widehat{R}_s\right\} = \mathcal{R}_s\big(\max\{\widehat{\varrho}, 0\}\big), \tag{4.69}$$

which holds because $\mathcal{R}_s(\varrho)$ is a decreasing function of ϱ as shown in the proof of Lemma 5. Eq. (4.65) then follows by plugging Eq. (4.69) into Eq. (4.68). Furthermore, $q_0|_{(R_s,R_d)=(R_s^{*},R_d^{*})} = q_{\text{th}}$ is equivalent to the third condition in Eq. (4.62), and thus Claim 1 holds. Given $\mathcal{R}_s\left([\varrho]^{+}\right)$, the corresponding power allocation policy is determined by the greedy algorithm characterized in Table I and is thus expressed by $\widetilde{\boldsymbol{\mu}}\left([\varrho]^{+}\right)$. Then, following the same arguments used for the proofs of Lemma 2 and Theorem 2, we can prove the optimality of $\left(\widetilde{\boldsymbol{\mu}}\left([\varrho]^{+}\right), R_d^{*}\right)$ for **IV-*P2*** and **IV-*P2*-*a***, respectively. The proof of Theorem 6 is completed. □

Based on Eq. (4.68) in the proof of Theorem 6 and Eqs. (4.32)-(4.33) in Theorem 2, the optimal TS-based and SPC-based multicast polices can be characterized through a unified expression. This is because through the function $\widetilde{g}_{\rm sum}(R_s)$, we convert the effective-capacity optimization problem for these two scenarios to the same form (see **IV-*P*1**-a and **IV-*P*2**-b). However, due to the diverse properties of $\widetilde{g}_{\rm sum}(R_s)$ in TS-based and SPC-based multicast, the specific expressions of the optimal sending rates are different. Further using the similar derivations in Theorems 3-4 and 5, we obtain the optimal SPC-based multicast policies in limiting cases as $\theta \to 0$ and $\theta \to \infty$, respectively, as follows.

a. $\theta \to 0$ with $q_{\rm th} < \widetilde{q}_0$:

As $\theta \to 0$, the optimal solution (R_s^*, R_d^*) of **IV-*P*2**-b under $q_{\rm th} < \widetilde{q}_0$ is given by

$$\begin{cases} R_d^* &= 0; \\ R_s^* &= \mathcal{R}_s(\varrho_{\rm th}), \end{cases}$$

where $\widetilde{q}_0$ for SPC-based multicast is defined as $\widetilde{q}_0 \triangleq q_0|_{(R_s,R_d)=(\mathcal{R}_s(0),0)}$ and $\varrho_{\rm th}$ is a constant which needs to be selected to satisfy $q_{\rm th} = q_0|_{(R_s,R_d)=(R_s^*,R_d^*)}$.

b. $\theta \to 0$ with $q_{\rm th} \geq \widetilde{q}_0$:

As $\theta \to 0$, the optimal SPC-based multicast policy under $q_{\rm th} \geq \widetilde{q}_0$ reduces to

$$(R_s^*, R_d^*) \;=\; \left(\mathcal{R}_s(0), \left[\widetilde{\kappa} - \mathcal{R}_s(0)\right]^+\right),$$

where $\widetilde{\kappa}$ is a constant and is selected to guarantee $q_{\rm th} = q_0|_{(R_s,R_d)=(R_s^*,R_d^*)}$.

c. $\theta \to \infty$:

As $\theta \to \infty$, the optimal SPC-based multicast policy reduces to

$$(R_s^*, R_d^*) = \begin{cases} \left(\mathcal{R}_s(0), \widetilde{\xi} - \mathcal{R}_s(0)\right), & \text{if } \mathcal{R}_s(0) < \widetilde{\xi}; \\ \left(\widetilde{\xi}, 0\right), & \text{if } \mathcal{R}_s(\rho) \leq \widetilde{\xi} \leq \mathcal{R}_s(0); \\ (\mathcal{R}_s(\rho), 0), & \text{if } \widetilde{\xi} < \mathcal{R}_s(\rho) \end{cases}$$

where $\widetilde{\xi}$ is a constant and is selected such that $q_{\text{th}} = q_0|_{(R_s,R_d)=(R_s^*,R_d^*)}$ holds.

G. Simulation Evaluations

We use simulation experiments to evaluate the effective-capacity performances of our derived optimal TS-based and SPC-based adaptive multicast transmission policies. In simulations, the signal bandwidth B is equal to 5×10^4 Hz and the time-frame length T is set to 2 ms. We use Rayleigh-fading channel model as the typical example for simulations. For comparative analyses, we also investigate some straightforward adaptive multicast schemes as the baseline to demonstrate the superiority of our derived optimal policies. These baseline multicast schemes are described as follows.

1. Baseline Multicast Schemes

a. Fixed-dominating-position based policy (MFDP)

The fixed-dominating-position (FDP) based scheme, which was studied in [58, 76], determines the sending rate based on the jth largest SNR among multicast receivers in each time instant, where j is fixed for all $\boldsymbol{\gamma}$. In particular, the send rate is determined by $R_s = \log(1+\gamma_{\pi(j)})$. In order to satisfy the loss-rate constraint, the constant j is set equal to $\lceil N(1-q_{\text{th}})\rceil$. For fair comparisons, we also introduce the pre-drop strategy into the FDP scheme. Specifically, the pre-drop rate is equal to a constant $\overline{R}_d$, which

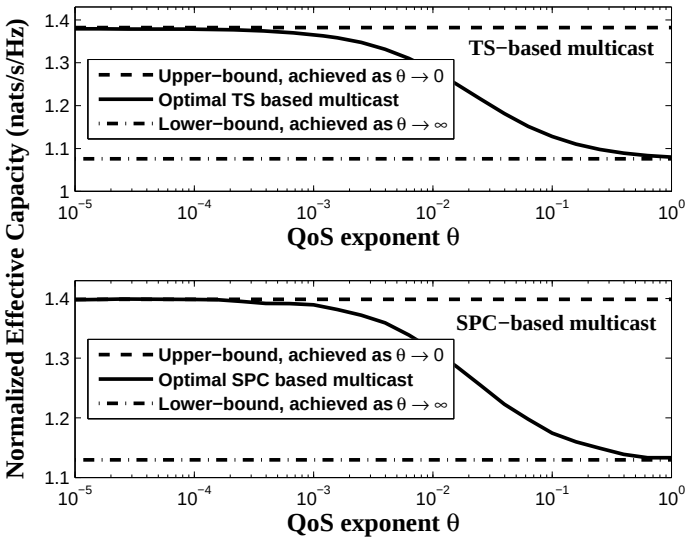


Fig. 15. Normalized effective capacity $(\mathscr{C}(\theta)/TB)$ versus QoS exponent θ in Raleigh fading channels, where $N = 5$, $\overline{\gamma} = 10$ dB, and $q_{\text{th}} = 0.1$.

is selected such that the loss rate just reaches q_{th}. We term this scheme the modified FDP (MFDP) policy.

b. Constant-rate policy

The constant rate policy is a non-adaptive transmission scheme, where the sending rate does not vary with instantaneous channel-fading status. The constant sending rate is denoted by $\overline{R}_s$, which needs to be determined such that the average loss rate is equal to q_{th}. Correspondingly, the pre-drop strategy will not be applied and the effective capacity of the constant-rate policy is just equal to $\overline{R}_s$.

2. Simulation Results

Figure 15 plots the normalized effective capacity $\mathcal{C}(\theta)$ versus the QoS exponent θ with $q_{\text{th}} = 0.1$. Fig. 15 shows that the effective capacities of both TS-based and SPC-based multicast policies decrease as θ increases. This is expected since with the fixed

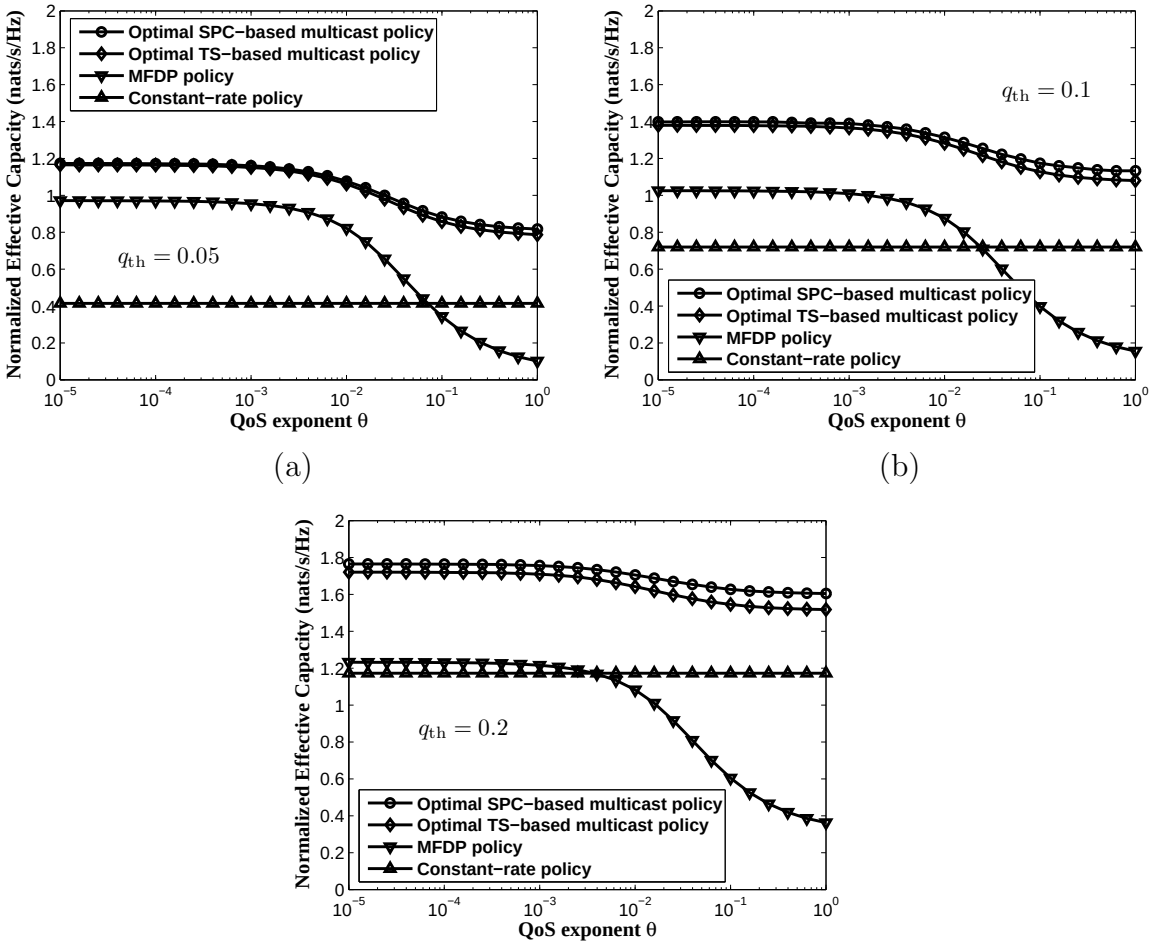
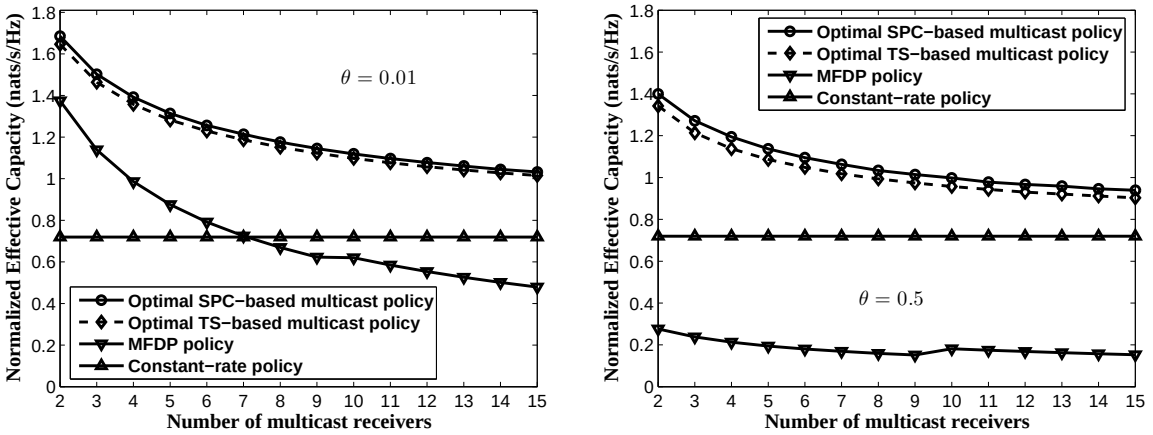


Fig. 16. Normalized multicast effective capacity $\mathscr{C}(\theta)/(BT)$ versus QoS exponent θ, where $\overline{\gamma} = 10$ dB and $N = 5$. (a) $q_{\rm th} = 0.05$. (b) $q_{\rm th} = 0.1$. (c) $q_{\rm th} = 0.2$.

amount of wireless resources, the more stringent QoS requirement (corresponding to larger θ) can only support the lower traffic load. Thus, the effective capacity achieves its upper-bound as $\theta \to 0$, implying that no delay constraint is imposed. Accordingly, we can obtain the upper-bounds for TS-based and SPC-based multicast through the optimal policies given by Eqs. (4.40)-(4.42) and Eq. (4.70), respectively, which are depicted in Fig. 15. On the other hand, the effective capacity will converge to its lower bound as $\theta \to \infty$, where any delay is intolerable. We can then determine

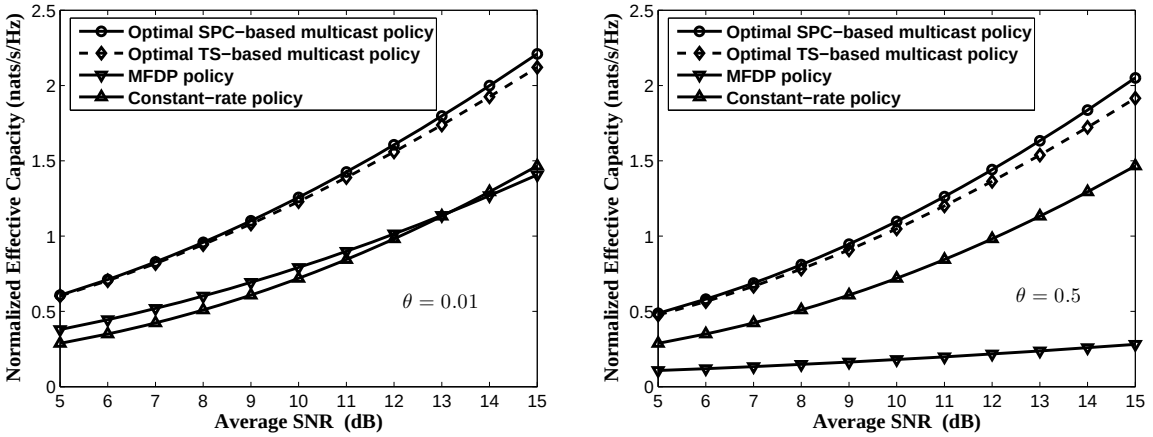


(a) (b)

Fig. 17. Normalized multicast effective capacity $\mathscr{C}(\theta)/(BT)$ versus multicast group size N, where $\overline{\gamma} = 10$ dB and $q_{\text{th}} = 0.1$.

the lower-bounds for TS-based and SPC-based multicast from Eqs. (4.47)-(4.49) and Eq. (4.70), respectively, which are also plotted in Fig. 15.

Figure 16 compares the normalized effective capacities among different adaptive multicast policies. As shown in Fig. 16, our derived optimal policies significantly outperform the MFDP and the constant-rate policies. The optimal SPC-based policy can achieve better effective-capacity performance than the optimal TS-based policy. The difference between their effective-capacity performances are not large. However, when either θ or q_{th} gets larger, the superiority of the SPC-based policy over the TS-based policy will become more significant. Fig. 16 also shows that the effective capacities of all policies are decreasing functions of θ except for the constant-rate policy. This is because the service rate of the constant-rate policy does not vary with either the instantaneous channel quality nor the delay-QoS requirement θ. However, comparing Fig. 16(b) with Fig. 15, which have the same simulation setup, we can see that even the lower-bounds for the optimal TS-based and SPC-based multicast policies significantly outperform the effective capacity achieved under the constant-rate policy.



(a) (b)

Fig. 18. Normalized multicast effective capacity $\mathscr{C}(\theta)/(BT)$ versus the average SNR $\overline{\gamma}$, where $N = 6$ and $q_{\text{th}} = 0.1$.

Figures 17(a) and 17(b) plot the dynamics of the effective capacity against the multicast group size N under $\theta = 0.01$ and $\theta = 0.5$, respectively. As shown in Fig. 17, our derived TS-based and SPC-based optimal policies also significantly outperform other policies. Fig. 17 shows that the effective capacity of the constant-rate policy does not vary with N, but the effective capacities of other policies decrease as N gets larger. We can also see from Fig. 17 that when N or θ gets large, the effective capacity of the MFDP policy degrades very quickly and become much lower than that of the constant-rate policy. In contrast, our derived optimal policies decrease slowly with the increase of N and θ and outperform the constant-rate policy under various conditions. Note that when $N \to \infty$, the constant-rate policy can be treated as a special form of the TS-based policy (also a special form of the SPC-based policy). Then, we can expect that our derived optimal TS-based and SPC-based multicast policies both dominate the constant-rate policy even as $N \to \infty$. Fig. 18 depicts the effective capacity as a function of the average SNR $\overline{\gamma}$. We observe from Fig. 18 that all schemes' effective capacities are increasing functions of the average SNR. When the average SNR gets larger, the superiority of our derived optimal policies over

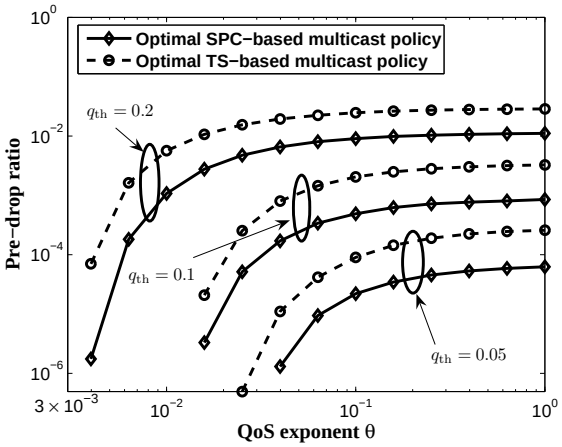


Fig. 19. Pre-drop ratio versus QoS exponent θ, where $\overline{\gamma} = 10$ dB, $N = 5$, and $q_{\rm th} = 0.1$.

other policies becomes more significant. Moreover, as the average SNR decreases, the difference between the effective capacities of the optimal TS-based and SPC-based policies gradually vanishes.

Figure 19 plots the pre-drop ratio of the optimal TS-based and SPC-based policies versus the QoS exponent θ, where the pre-drop ratio is defined as the ratio of the average pre-drop rate $\mathbb{E}\{R_d\}$ to the average throughput $\mathbb{E}\{R_s + R_d\}$. As shown in Fig. 19, the pre-drop ratios of both policies increases as θ becomes larger. This is because the major purpose to introduce the pre-drop rate is to meet the delay-QoS requirement when the instantaneous channel quality is poor. Thus, when θ gets larger, implying that the delay constraint becomes more stringent, more data then have to be dropped when channel quality is poor. Moreover, we can see that the ratio of the SPC-based policy is smaller than that of the TS-based policy. This is also expected since with the same loss-constraint, the SPC-based scheme can support higher sending rate than the TS-based scheme. As a result, lower pre-drop ratio is required to meet the delay-QoS constraint.

H. Summary

We proposed the efficient framework for mobile multicast over broadcast fading channels by integrating the effective-capacity theory, multicast rate adaptation, and loss-rate control. Subject to the QoS exponent and average loss-rate constraint, we formulated an effective-capacity maximization problem via channel-aware rate adaptation. For rate adaptation, we employed the time-sharing and superposition-coding techniques, respectively, to handle the heterogeneous qualities over channels across multicast receivers. We also developed a novel *pre-drop* scheme to implement the more efficient QoS-driven wireless multicasting. Under the developed framework, we derived the optimal time-sharing based and superposition-coding adaptive multicast policies. Simulation evaluations demonstrated the tradeoff between the effective capacity and QoS metrics and showed the superiority of our derived optimal policies over the fixed-dominating-position based policy and the constant-rate policies.

CHAPTER V

STATISTICAL DELAY-QOS PROVISIONINGS FOR WIRELESS UNICAST/MULTICAST OF MULTI-LAYER VIDEO STREAMS

A. Introduction

Recently, supporting real-time video services with diverse QoS constraints has become one of the essential requirements for wireless communications networks. Consequently, how to efficiently guarantee QoS for video transmission attracts more and more research attention [2, 3, 8, 19, 32, 35, 36, 53, 56, 76, 80, 81]. However, the unstable wireless environments and the popular layer-structured video signals [35, 36, 81] impose a great deal of challenges in delay QoS provisionings. As discussed in Chapters I and II of this dissertation, due to the highly-varying wireless channels, the *deterministic* delay QoS requirements are usually hard to guarantee. As a result, *statistical* delay QoS guarantees [2, 3, 8, 19, 53, 56, 76], in terms of effective bandwidth/capacity and queue-length-bound/delay-bound violation probabilities, have been proposed and demonstrated as the powerful way to characterize delay QoS provisionings for wireless traffics. While many related existing research works mainly focused on the scenarios with single-layer streams [8, 19, 53, 56], the modern video coding techniques usually generate layer-structured traffics [35, 36, 81]. Unfortunately, how to design efficient schemes to support statistical delay QoS for layered video traffics over wireless networks has been neither well understood nor thoroughly studied.

In video transmissions, video source is usually encoded into a number of data layers [35, 36, 81] in the application protocol layer. By applying layered video coding, the receivers under poorer channel conditions can get only lower video quality, while those under better channel conditions can achieve higher video quality. Although the

layered coding techniques are efficient in handling diverse channel conditions, they also raise new challenges for statistical delay QoS guarantees, which are not encountered in single-layer video transmissions. First, we need to keep the synchronous transmissions across different video layers, implying the same delay-bound violation probability for all layers. Second, for multi-layer video stream, it is a natural requirement that different video layers can tolerate different loss levels. Therefore, the scheduling and resource allocation need to be aware of the diverse loss constraints. Third, how to minimize the consumption of scarce wireless-resources while satisfying the specified delay QoS requirements is a widely cited open problem.

Besides the general challenges in statistical delay QoS guarantees for the unicast transmission of layered video, multicasting layered video over wireless networks further complicates the problem significantly due to the heterogeneous channel qualities across multicast receivers at each time instant. Unlike in the wireless multicast, there are relatively more research results for the multicast over wireline networks. A number of multicast protocols were proposed over wireline networks. The authors of [35] developed the efficient receiver-driven layered multicast over the Internet, where the video source is encoded to a hierarchical signal with different layers. Each layer corresponds to a multicast group and multicast receivers can join/leave the group based on their bandwidths. In [32], the authors proposed a novel flow control scheme for multicast services over the asynchronous transfer mode (ATM) networks. The kernel parts of this scheme include the optimal second-order rate control algorithm and the feedback soft-synchronization protocol [33, 34], which can achieve *scalable* and *adaptive* multicast flow control over bandwidth and buffer occupancies and utilizations. The above designs are shown to be efficient in the wireline networks. However, the multicast strategies in wireline networks cannot be directly applied into wireless networks. This is because highly and rapidly time-varying wireless-channels qualities

result in unstable bandwidths and thus unsatisfied loss and delay QoS. For wireless video multicast, at the multicast sender we need to design the transmission scheme to control the loss and/or delay performance for all multicast receivers at each video layer based on their instantaneous channel qualities.

In Chapter IV of this dissertation, we applied the effective capacity theory to propose and evaluate rate-adaptation schemes for statistical delay QoS guarantees in mobile multicast. However, the analyses only focused on single-layer stream. It remains one of the major challenges to extend the statistical QoS theory into multi-layer video multicast in developing QoS-driven transmission strategies. In [36], the authors proposed a cross-layer architecture for adaptive video multicast over multirate wireless LANs. In particular, two-layer video signals are considered, which include the base layer (more important) and enhancement layer (less important). The authors derived the transmission rate for the base layer according to the worst-case signal-to-noise ratio (SNR) among all receivers, while dynamically regulating the transmission rate for the enhancement layer based on the best-case SNR to benefit receivers with better channel qualities. However, under this strategy, the loss rate of the enhancement layer will vary significantly with the statistical characteristics of wireless channels, and thus is hard to control.

To overcome the above problems, in this chapter we propose an efficient framework to model the statistical delay QoS guarantees, in terms of QoS exponent, effective bandwidth/capacity, and delay-bound violation probability, for multi-layer video transmission over wireless networks. In particular, a separate queue is maintained for each video layer, and the same delay bound and the corresponding violation probability are set up for all video layers. We then develop a set of optimal adaptive transmission schemes to minimize the resource consumption while satisfying the diverse QoS requirements under various scenarios, including video unicast/multicast

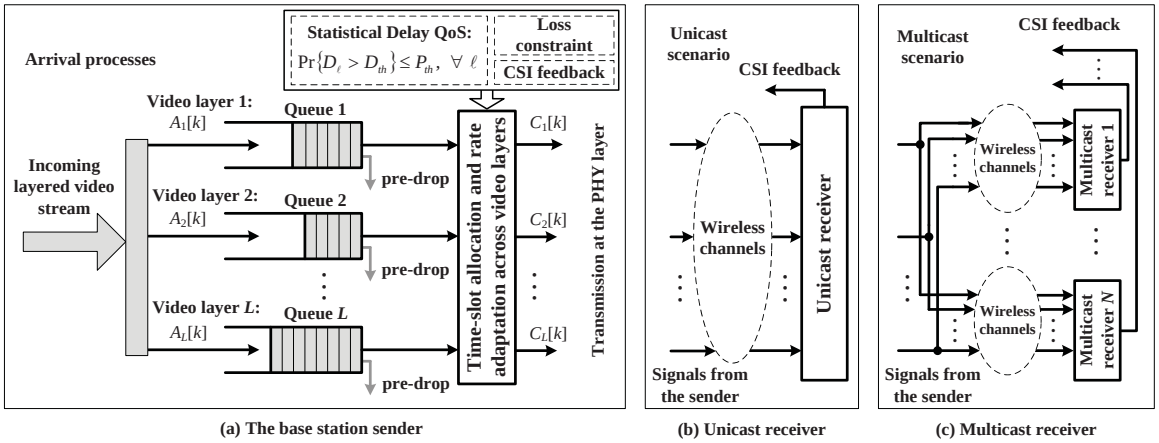


Fig. 20. The system modeling framework for layered-video transmission over wireless networks: (a) The layered-video arrival stream and the sender's processing. (b) Unicast scenario. (c) Multicast scenario.

with and/or without loss tolerance.

The rest of this chapter is organized as follows. Section B describes the system model. Section C proposes the framework of statistical delay QoS guarantees for multi-layer video unicast and multicast. Sections D and E derive the optimal adaptive transmission schemes for video unicast and multicast, respectively. Section F presents the simulation evaluations. The chapter concludes with Section G.

B. The System Model

We consider the unicast/multicast system model for multi-layer video distribution in wireless networks, as shown in Fig. 20. Specifically, the base station sender is responsible for transmitting a multi-layer video stream to a single receiver (unicast) or multiple receivers (multicast) over broadcast fading channels. The video stream generated by upper protocol layers (e.g., application layer) consists of L video layers, each having the specific QoS requirements. The L-layer video stream will be injected to the physical (PHY) layer. Then, as depicted in Fig. 20(a), we aim at develop-

ing strategies to efficiently allocate limited wireless resources for multi-layer video transmission while satisfying the specified QoS requirements for each video layer.

At the PHY layer, the sender uses a constant transmit power with the signal bandwidth equal to B Hz. The wireless broadcast channels are assumed to be flat fading. Then, we can use an SNR vector $\boldsymbol{\gamma} \triangleq (\gamma_1, \gamma_2, \ldots, \gamma_N)$ to characterize the channel state information (CSI) of receivers, where N denotes the number of unicast/multicast receivers, γ_n is the received SNR of the nth receiver for $n = 1, 2, \ldots, N$, and $\{\gamma_n\}_{n=1}^{N}$ are independent and identically distributed (i.i.d.) for the cases of $N > 1$. When N is equal to 1, the scenario reduces to video unicast,[1] as illustrated in Fig. 20(b); while N is larger than 1, we get the multicast scenario depicted in Fig. 20(c). The CSI $\boldsymbol{\gamma}$ is modeled as an ergodic and stationary block-fading process, where $\boldsymbol{\gamma}$ does not change within a time-frame with the fixed length T, but varies independently from frame to frame. Moreover, $\boldsymbol{\gamma}$ follows Rayleigh fading model, which is one the most generally used models to characterize wireless fading channels. In addition, we assume that $\boldsymbol{\gamma}$ can be perfectly estimated by the receivers and reliably fed back to the sender without delay through the dedicated feedback control channels.

C. Modeling Framework for Wireless Unicast/Muticast of Multi-Layer-Video

We propose the following framework for transmitting multi-layer video over fading channels by integrating the adaptive resource allocations, statistical QoS guarantees, and loss constraints.

[1]When $N = 1$, we write SNR as γ instead of γ_1 to simplify notation.

1. Multi-Queue Model for Multi-Layer Video Arrival Processes

The modern video coding techniques [81] usually encode the video source into a number of video layers with different relevance and importance. The most important layer is called based layer and the other layers are called enhancement layers. Because of the diverse importance, different strategies need to be proposed for the corresponding video layers in the PHY-layer transmission, depending on the specified QoS requirements (to be detailed in Sections C-2 and C-4). Then, to achieve the efficient video transmission, the sender manages a separate queue for each video layer. As shown in Fig. 20(a), the data arrival rate of the ℓth video layer is characterized by a discrete-time process, denoted by $A_\ell[k]$ (nats/frame), where $\ell = 1, 2, \ldots, L$, and $[k]$, $k = 1, 2, \ldots$, is the index of time frames; the service rate process (departure process) of the ℓth layer is denoted by $C_\ell[k]$ (nats/frame). Moreover, we determine $C_\ell[k]$ based on CSI, total available wireless resources, and QoS constraints.

2. Statistical Delay QoS Guarantees for Video Transmissions

For video transmissions, delay is one of the most important QoS metrics. However, due to the highly varying wireless channels, usually the hard delay bound cannot be guaranteed. Therefore, the statistical metric, namely *delay-bound violation probability* [2, 3, 8, 53], has been widely applied in QoS evaluations for real-time services. In our framework, we also use the delay-bound violation probability to statistically characterize the delay QoS provisionings for each video layer. In particular, a queueing delay bound, denoted by $D_{\rm th}$, is specified. Accordingly, over all video layers, the delay-bound violation probability cannot exceed the threshold denoted by $P_{\rm th}$:

$$\Pr\{D_\ell > D_{\rm th}\} \leq P_{\rm th}, \quad P_{\rm th} \in (0, 1), \tag{5.1}$$

for all $\ell \in \{1, 2, \ldots, L\}$, where D_ℓ denotes the queueing delay at the ℓth video layer. The delay-bound violation probability in Eq. (5.1) is evaluated over the entire transmission process, which is assumed to be long enough. Note that D_{th} and P_{th} are application-dependent parameters. Moreover, the pair of D_{th} and P_{th} is set to be the same for all video layers, because the synchronous transmission across different video layers is usually required. Also note that the delay in wireless video transmissions may result from multiple factors such as transmissions, queueing, and decoding. In this chapter, we mainly focus on queueing delay, which reflects the capability of the wireless channel (transmission bottleneck) in supporting video distributions.

3. Adaptive Resource Allocation and Transmissions

To efficiently use the limited wireless resources for video unicast/multicast, we employ the adaptive transmission strategy (based on the CSI), consisting of three folds: transmission rate adaptation, dynamic time-slot allocation, and adaptive pre-drop queue management strategy, as detailed below.

a. Time slot allocation for video layers

Each time frame is divided into L time slots, the lengths of which are denoted by $\{T_\ell[k]\}_{\ell=1}^{L}$, where $0 \leq T_\ell[k] \leq T$ and $\sum_{\ell=1}^{L} T_\ell[k] \leq T$. The time slot with length $T_\ell[k]$ is used for transmitting data of the ℓth video layer. For convenience of presentation, we further define time proportion $t_\ell[k] \triangleq T_\ell[k]/T$, and thus we have $\sum_{\ell=1}^{L} t_\ell[k] \leq 1$. Notice that our target is to minimize the wireless-resource consumption while satisfying the QoS requirements imposed by video qualities. Thus, $\sum_{\ell=1}^{L} t_\ell[k]$ may be smaller than 1 for some $\boldsymbol{\gamma}$.

b. Rate adaptation of unicast/multicast

We denote the total amount of transmitted data at the ℓth video layer in the kth time frame by $\mathcal{R}_\ell[k]$ (with the unit nats/frame). Moreover, we use the normalized transmission rate, denoted by $R_\ell[k]$ (nats/s/Hz), to characterize the transmission rate adaptation, where $R_\ell[k] \triangleq \mathcal{R}_\ell[k]/(BT)$. We assume that capacity-achieving codes are used for transmission at the PHY layer. Accordingly, for unicast, the normalized transmission rate of the ℓth video layer is set equal to the Shannon capacity under the current SNR γ:

$$R_\ell[k] = \log(1+\gamma) \qquad \text{(nats/s/Hz)}. \tag{5.2}$$

Clearly, $R_\ell[k]$ does not vary with ℓ, and thus we only focus on time-slot allocation for unicast.

For the multicast case, the rate adaptation becomes more complicated. In particular, the time slot for video layer ℓ is further partitioned into N sub-slots. The length of the nth sub-slot, denoted by $T_{\ell,n}[k]$, is equal to $T_\ell[k]t_{\ell,n}[k]$, where $0 \le t_{\ell,n}[k] \le 1$ and $\sum_{n=1}^{N} t_{\ell,n}[k] = 1$. Within the nth sub-slot, the transmission rate is set equal to the Shannon capacity under SNR γ_n, and thus the data transmitted in this sub-slot can be correctly decoded only by receivers with SNR higher than or equal to γ_n. Then, the normalized transmission rate $R_\ell[k]$ for the ℓth video layer becomes

$$\begin{aligned} R_\ell[k] &= \sum_{n=1}^{N} t_{\ell,n}[k] R_{\ell,n}[k] \\ &= \sum_{n=1}^{N} t_{\ell,n}[k] \log(1+\gamma_n) \quad \text{(nats/s/Hz)}, \end{aligned} \tag{5.3}$$

where $R_{\ell,n}[k] \triangleq \log(1+\gamma_n)$ is the normalized transmission rate for the nth sub-slot of the ℓth video layer. As a result, we need to not only adjust $t_\ell[k]$'s for each layer,

but also regulate $t_{\ell,n}[k]$'s within every time slot.

Unlike the wireline multicast networks, in this chapter we focus on the layered video transmissions over wireless networks, which has a single-hop cellular network structure. Due to the broadcast nature of wireless channels, the sender only needs to transmit a *single copy* of data and *all* multicast receivers can hear the transmitted signal for each video layer. Under this model, our scheme employs the *sender-oriented* multicast approach because the sender needs to dynamically adjust the transmissions rate in controlling the loss rate (to be detailed in Section C-4) and guaranteeing the delay-QoS requirements (see Section C-2) *across* different multicast receivers.

c. Pre-drop strategy

In [56], for multicasting single-layer-data we developed the pre-drop strategy to gain a more robust queueing behavior. In this chapter, we further extend the pre-drop strategy to multi-layer video transmission. Specifically, based on the CSI, in each time frame the sender can drop some data (see Fig. 20(a)) from the head of each queue, but treat them *as if* they were transmitted. We denote the amount of dropped data in ℓth video layer by $Z_\ell[k]$ (nats/frame) and define the normalized drop rate, denoted by $z_\ell[k]$, as $z_\ell[k] \triangleq Z_\ell[k]/(BT)$ (nats/s/Hz). Then, the service process $C_\ell[k]$ of the ℓth video layer is given by

$$C_\ell[k] = BT(t_\ell[k]R_\ell[k] + z_\ell[k]) \qquad \text{(nats/frame)}. \tag{5.4}$$

Clearly, the pre-drop strategy suppresses the growing speed of the queue for a more robust queueing behavior, but this strategy also causes data loss to all multicast receivers. As a result, $z_\ell[k]$ needs to be determined by not only the CSI, but also the loss constraints (see Section C-4).

4. Loss Rate Constraint

Although a certain loss is usually tolerable for delay-sensitive services, the loss level cannot be arbitrarily high. Consequently, we require the loss rate of the ℓth video layer for each receiver to be limited lower than or equal to an application-dependent threshold, denoted by $q_{\text{th}}^{(\ell)}$. The loss rate of the ℓth video layer for the nth receiver, denoted by $q_{\ell,n}$, is defined as the ratio of the amount of data correctly received by this receiver to that of the data transmitted at this video layer. Data loss for unicast will be caused only by the pre-drop strategy, while data loss for multicast will be introduced by both pre-drop operation and heterogeneous channel fading across multicast receivers.

Since various efficient forward-error control (FEC) codes [60, 67, 68, 82, 83] at upper protocol layers were proposed and widely applied to multicast communications in wired and/or wireless networks, in our framework we suppose that FEC mechanisms are already employed at the upper protocol layers. The error-control redundancies added in the FEC codes at different video layers are inherently related among video layers and are jointly determined by the targeted video-delivery qualities at different video layers. Correspondingly, the tolerable loss-rate levels $q_{\text{th}}^{(\ell)}$'s for different video layers (indexed by ℓ) are jointly specified based on the video delivery quality requirements and the error control redundancy degrees across different video layers. Under this framework, we then mainly focus on how to use the minimum wireless resources with QoS guarantees to unicast/multicast multi-layer video over wireless channels.

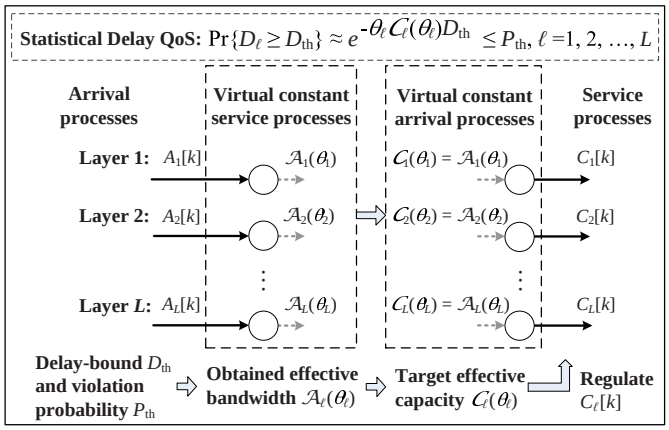


Fig. 21. Illustration of design procedures to guarantee statistical delay QoS by using effective capacity and effective bandwidth theories.

5. Design Procedures for Transmitting Layered Video with Statistical QoS Guarantees

As addressed in Chapter II, for a dynamic queueing system, in order to guarantee the QoS exponent θ given in Eqs. (2.1)-(2.2), the following equation need to be satisfied [3, 53]:

$$\mathcal{C}(\theta) = \mathcal{A}(\theta), \tag{5.5}$$

where $\mathcal{C}(\theta)$ denotes the effective capacity of the service-rate process under the specified QoS exponent θ and $\mathcal{A}(\theta)$ represents the effective bandwidth of the arrival-rate process of the queueing system. Inspired by this property, the statistical delay QoS guarantees can be characterized through the arrival process and service process separately. As shown in Fig. 21, the queueing system for the ℓth video layer can be decomposed to two virtual queueing systems. The one on the left-hand side of Fig. 21 is composed by the true arrival process $A_\ell[k]$ and one virtual constant-rate service process, the rate of which is equal to the effective bandwidth $\mathcal{A}_\ell(\theta_\ell)$ of $A_\ell[k]$; the right one consists of the true service process $C_\ell[k]$ and one constant-rate virtual arrival process, the rate

Table II. The Design Procedures to Provide Statistical Delay QoS Guarantees for Transmitting Multi-Layer Video Stream.

Step 1:	Determine effective bandwidth functions $\mathcal{A}_\ell(\theta)$ for the arrival processes $A_\ell[k]$, $\ell = 1, 2, \ldots, L$.
Step 2:	Apply Eq. (2.9) or Eq. (2.10) to find the solution θ_ℓ to the equation $\Pr\{D_\ell > D_{\rm th}\} = P_{\rm th}$ and get the corresponding effective bandwidth $\mathcal{A}_\ell(\theta_\ell)$.
Step 3:	Set the target effective capacity $\overline{\mathcal{C}}_\ell = \mathcal{A}_\ell(\theta_\ell)$ for the service processes of each video layer.
Step 4:	Jointly design adaptive service process $\mathcal{C}_\ell[k]$ for each video layer, such that $\mathcal{C}_\ell(\theta_\ell) \geq \overline{\mathcal{C}}_\ell$ is satisfied while minimizing the total consumed wireless resources.

of which is equal to the effective capacity $\mathcal{C}_\ell(\theta_\ell)$ of $C_\ell[k]$. Using the above concept, we develop the design procedures to provide statistical delay QoS guarantees for transmitting multi-layer video stream as shown in Table II.

Among the procedures in Table II, Steps 1 and 2 first identify the effective bandwidth $\mathcal{A}_\ell(\theta_\ell)$ and QoS exponent θ_ℓ required to satisfy the delay-bound $D_{\rm th}$ and its violation probability $P_{\rm th}$. Then, to satisfy the delay QoS in Eq. (5.1), we need to either satisfy Eq. (5.5) or guarantee that the effective capacity is larger than the effective bandwidth, which results in Steps 3 and 4.

The analytical expressions of effective bandwidth for many typical arrival processes, such as constant-rate process, autoregressive (AR) process, and Markovian process, can be found in [3]. Note that if $A_\ell[k]$ is time varying, in Step 2 we can determine θ_ℓ through Eq. (2.10). However, if $A_\ell[k]$ is a constant-rate process equal to $\overline{A}_\ell$, we have $\mathcal{A}_\ell(\theta) = \overline{A}_\ell$ for all θ, implying that the delay-bound violation probability of the virtual queueing system on the left-hand side of Fig. 21 is always equal to 0. Therefore, we cannot derive θ_ℓ directly through Eq. (2.10). In contrast, the QoS exponent θ_ℓ to guide the adaptive transmission needs to be determined by using Eq. (2.9) under the condition of $\mathcal{C}_\ell(\theta_\ell) = \mathcal{A}_\ell(\theta_\ell) = \overline{A}_\ell$.

D. Unicasting Multi-Layer Video Stream

Assuming that the target effective capacity $\left\{\overline{\mathcal{C}}_\ell\right\}_{\ell=1}^{L}$ and QoS exponent θ_ℓ have been determined, we next focus on developing the optimal adaptive time-slot allocation and pre-drop strategy to satisfy the QoS requirements while minimizing the wireless-resource consumption. Unless otherwise mentioned, we drop the time-frame index $[k]$ for the corresponding variables in the rest of this chapter to simplify notations. As discussed in Chapter II, if a stationary and ergodic service rate process C_ℓ is uncorrelated across different time frames, we can write its effective capacity as follows:

$$\mathcal{C}_\ell(\theta_\ell) = -\frac{1}{\theta_\ell}\log\left(\mathbb{E}\left\{e^{-\theta_\ell C_\ell}\right\}\right) \quad \text{nats/frame.} \tag{5.6}$$

Since the block-fading channel process described in Section B is time uncorrelated, we can apply Eq. (5.6) for our framework to derive the adaptive unicast/mutlicast schemes with the statistical QoS guarantees.

1. Unicasting Layered Video Stream Without Loss Tolerance

We first consider the cases without loss tolerance for multi-layer video transmissions, i.e., $q_{\text{th}}^{(\ell)} = 0$ for all ℓ and the pre-drop strategy will not be applied. Thus, we only need to focus on regulating the time-slot proportion $\{t_\ell\}_{\ell=1}^{L}$ for each video layer. Following the design target and QoS constraints characterized in Section C, we derive the adaptive transmission strategy by solving the following optimization problem.

$$\textbf{V-}\boldsymbol{P1}: \text{ Unicast without loss tolerance}$$

$$\min_{\boldsymbol{t}} \left\{\sum_{\ell=1}^{L}\mathbb{E}_\gamma\left\{t_\ell\right\}\right\} \tag{5.7}$$

$$\text{s.t.:} \quad \mathcal{C}_\ell(\theta_\ell) \geq \overline{\mathcal{C}}_\ell, \quad \forall\,\ell, \tag{5.8}$$

$$\sum_{\ell=1}^{L} t_\ell - 1 \leq 0,\ 0 \leq t_\ell \leq 1,\ \forall\,\gamma, \tag{5.9}$$

where $\boldsymbol{t} \triangleq (t_1, t_2, \ldots, t_L)$ and $\mathbb{E}_\gamma\{\cdot\}$ denotes the expectation over the random variable γ.

Using Eq. (5.6), the constraint in (5.8) can be equivalently rewritten as

$$\mathbb{E}_\gamma \left\{ e^{-\beta_\ell t_\ell \log(1+\gamma)} \right\} - V_\ell \leq 0, \tag{5.10}$$

where $\beta_\ell \triangleq \theta_\ell TB$ is termed normalized QoS exponent and $V_\ell \triangleq e^{-\theta_\ell \overline{\mathcal{C}}_\ell}$. It is not difficult to see: 1) the objective function in **V-*P*1** is convex over $\boldsymbol{t}$; 2) the functions on the left-hand side of all inequality constraints (Eqs. (5.9) and (5.10)) are convex over $\boldsymbol{t}$. Therefore, **V-*P*1** is a convex problem [77] and the optimal solution can be obtained by using the Lagrangian method and the Karush-Kuhn-Tucker (KKT) conditions [77], which is summarized in Theorem 7.

Theorem 7. *The optimal solution $\boldsymbol{t}^*$ to problem* **V-*P*1***, if existing, is determined by*

$$t_\ell^* = t_\ell(\gamma, \lambda_\ell^*, \psi_\gamma^*) \triangleq \left[-\frac{\log\left(\frac{1+\psi_\gamma^*}{\beta_\ell \lambda_\ell^* \log(1+\gamma)}\right)}{\beta_\ell \log(1+\gamma)} \right]^+, \tag{5.11}$$

where $[\cdot]^+ \triangleq \max\{\cdot, 0\}$. The parameters ψ_γ^* and $\{\lambda_\ell^*\}_{\ell=1}^L$ are the optimal Lagrangian multipliers associated with Eqs. (5.9) and (5.10), respectively. Given SNR γ in a fading state and $\{\lambda_\ell^*\}_{\ell=1}^L$, if

$$\sum_{\ell=1}^{L} t_\ell(\gamma, \lambda_\ell^*, 0) \geq 1 \tag{5.12}$$

holds, ψ_γ^* is the unique solution to

$$\sum_{\ell=1}^{L} t_\ell(\gamma, \lambda_\ell^*, \psi_\gamma^*) = 1, \quad \psi_\gamma^* \geq 0; \tag{5.13}$$

otherwise, we get

$$\psi_\gamma^* = 0. \tag{5.14}$$

Under the above strategy to determine $\boldsymbol{t}^*$ *and* ψ_γ^**, the optimal* $\{\lambda_\ell^*\}_{\ell=1}^L$ *are selected to satisfy*

$$\mathbb{E}_\gamma\left\{e^{-\beta_\ell t_\ell^* \log(1+\gamma)}\right\} - V_\ell = 0,\ \forall\, \ell. \tag{5.15}$$

Proof. We construct the Lagrangian function for **V-*P*1**, denoted by J, as follows:

$$\begin{aligned} J &= \mathbb{E}_\gamma\left\{\sum_{\ell=1}^L t_\ell\right\} + \mathbb{E}_\gamma\left\{\psi_\gamma\left(\sum_{\ell=1}^L t_\ell - 1\right)\right\} \\ &\quad + \sum_{\ell=1}^L \lambda_\ell\Big(\mathbb{E}_\gamma\left\{e^{-\beta_\ell t_\ell \log(1+\gamma)}\right\} - V_\ell\Big), \end{aligned} \tag{5.16}$$

where $\psi_\gamma \geq 0$ and $\lambda_\ell \geq 0$, $\ell = 1, 2, \ldots, L$, are Lagrangian multipliers associated with Eqs. (5.9) and (5.10), respectively. Then, the optimal $\boldsymbol{t}^*$ and Lagrangian multipliers of optimization problem **V-*P*1** satisfy the following KKT conditions [77]:

$$\begin{cases} \frac{\partial J}{\partial t_\ell}\big|_{t_\ell = t_\ell^*} = 0, \forall\, \ell, \forall\, \gamma \\ \psi_\gamma^* \geq 0 \text{ and } \lambda_\ell^* \geq 0; \\ \psi_\gamma^*\left(\sum_{\ell=1}^L t_\ell^* - 1\right) = 0,\ \forall\, \ell,\ \gamma; \\ \lambda_\ell^*\left(\mathbb{E}_\gamma\left\{e^{-\beta_\ell t_\ell^* \log(1+\gamma)}\right\} - V_\ell\right) = 0,\ \forall\, \ell. \end{cases} \tag{5.17}$$

Taking the derivative of J with respect to (w.r.t.) t_ℓ, we get

$$\frac{\partial J}{\partial t_\ell} = \Big(1 + \psi_\gamma - \lambda_\ell \beta_\ell (1+\gamma)^{-\beta_\ell t_\ell} \log(1+\gamma)\Big) f_\Gamma(\gamma) d\gamma \tag{5.18}$$

where $f_\Gamma(\gamma)$ is the probability density function (pdf) of γ. Plugging Eq. (5.18) into the first line of Eq. (5.17) and solving for t_ℓ^* under the boundary condition $t_\ell \geq 0$, we get Eq. (5.11).

According to Eq. (5.11), $t_\ell(\gamma, \lambda_\ell^*, \psi_\gamma)$ is a strictly decreasing function of ψ_γ for $t_\ell(\gamma, \lambda_\ell^*, \psi_\gamma) > 0$, and $\psi_\gamma^* \to \infty$ leads to $t_\ell^* \to 0$. Therefore, if Eq. (5.12) holds,

we can always find the unique ψ_γ^* to satisfy Eq. (5.14); otherwise, the inequality $\sum_{\ell=1}^{L} t_\ell(\gamma, \lambda^*, \psi_\gamma) - 1 < 0$ follows for any $\psi_\gamma \geq 0$, implying $\psi_\gamma^* = 0$ by applying the third line of Eq. (5.17). Through Eq. (5.11), $\lambda_\ell^* = 0$ results in $t_\ell = 0$ for all γ, and thus the constraint in Eq. (5.8) will be violated, implying an infeasible solution. Therefore, λ_ℓ^* has to be positive. Then, to satisfy the fourth line of Eq. (5.17), Eq. (5.15) must hold and thus Theorem 7 follows. □

Note that given $\{\lambda_\ell^*\}_{\ell=1}^{L}$, ψ_γ^* is easy to solve because $t_\ell(\gamma, \lambda_\ell^*, \psi_\gamma)$ is a decreasing function of ψ_γ. However, how to find $\{\lambda_\ell^*\}_{\ell=1}^{L}$ is still unknown. Moreover, Theorem 7 does not state whether the optimal solution exists. Next, we discuss how to get $\{\lambda_\ell^*\}_{\ell=1}^{L}$ and examine the existence of the optimal solution, which can be performed either off-line or on-line. Based on the optimization theory [77, 79], the Lagrangian dual problem to **V-*P*1** is given by

$$\max_{(\boldsymbol{\lambda}, \psi_\gamma)} \left\{ \widetilde{J}(\boldsymbol{\lambda}, \psi_\gamma) \right\}, \tag{5.19}$$

where $\boldsymbol{\lambda} \triangleq (\lambda_1, \lambda_2, \ldots, \lambda_L)$ and $\widetilde{J}(\boldsymbol{\lambda}, \psi_\gamma)$ is the Lagrangian dual function defined by $\widetilde{J}(\boldsymbol{\lambda}, \psi_\gamma) \triangleq \min_{\boldsymbol{t}}\{J\} = J|_{t_\ell = t_\ell(\gamma, \lambda_\ell, \psi_\gamma)}$. We can further convert Eq. (5.19) into $\max_{(\boldsymbol{\lambda}, \psi_\gamma)} \left\{ \widetilde{J}(\boldsymbol{\lambda}, \psi_\gamma) \right\} = \max_{\boldsymbol{\lambda}} \left\{ \widetilde{J}(\boldsymbol{\lambda}, \psi_\gamma(\boldsymbol{\lambda})) \right\}$, where $\psi_\gamma(\boldsymbol{\lambda})$ denotes the maximizer of $\widetilde{J}(\boldsymbol{\lambda}, \psi_\gamma)$ given $\boldsymbol{\lambda}$. Moreover, we can obtain $\psi_\gamma(\boldsymbol{\lambda})$ by using the same procedures as those used in determining ψ_γ^*, which are given by Eqs. (5.12)-(5.14) in Theorem 7.

Since problem **V-*P*1** is convex, there is no duality gap between **V-*P*1** and its dual problem given by Eq. (5.19) if the optimal solution exists. Thus, the optimal Lagrangian multipliers $\{\lambda_\ell^*\}_{\ell=1}^{L}$ and ψ_γ^* to problem **V-*P*1** also maximize the objective function $\widetilde{J}(\boldsymbol{\lambda}, \psi_\gamma)$ in Eq. (5.19). Consequently, we can obtain $\{\lambda_\ell^*\}_{\ell=1}^{L}$ through maximizing $\widetilde{J}(\boldsymbol{\lambda}, \psi_\gamma)$. Following convex optimization theory [77], $\widetilde{J}(\boldsymbol{\lambda}, \psi_\gamma(\boldsymbol{\lambda}))$ is a concave function over $\boldsymbol{\lambda}$, and thus we can track the optimal $\boldsymbol{\lambda}^*$ by using the subgradient

method [77]:

$$\lambda_\ell := \lambda_\ell + \epsilon \left(\mathbb{E}_\gamma \left\{e^{-\beta_\ell t_\ell \log(1+\gamma)}\right\} - V_\ell\right)\Big|_{t_\ell = t_\ell(\gamma, \lambda_\ell, \psi_\gamma(\boldsymbol{\lambda}))} \tag{5.20}$$

where ϵ is a positive real number close to 0. and $\left(\mathbb{E}_\gamma \left\{e^{-\beta_\ell t_\ell \log(1+\gamma)}\right\} - V_\ell\right)$ in the above equation is a subgradient of $\widetilde{J}(\boldsymbol{\lambda}, \psi_\gamma(\boldsymbol{\lambda}))$ w.r.t. λ_ℓ [79] (see the definition of subgradient in Definition 9). If the optimal solution to **V-*P*1** exists, the above iteration will converge to the optimal $\boldsymbol{\lambda}^*$ with properly selected ϵ because of the concavity of $\widetilde{J}(\boldsymbol{\lambda}, \psi_\gamma(\boldsymbol{\lambda}))$. Correspondingly, $\left(\mathbb{E}_\gamma \left\{e^{-\beta_\ell t_\ell \log(1+\gamma)}\right\} - V_\ell\right)$ will converge to 0. If the optimal solution to **V-*P*1** does not exist, we cannot support such a statistical QoS requirement even we use up all time slots. Then, $(\mathbb{E}_\gamma \left\{e^{-\beta_\ell t_\ell \log(1+\gamma)}\right\} - V_\ell)$ is always larger than 0 for some ℓ. As a result, λ_ℓ will approach infinity. So, if any λ_ℓ does not converge and keeps increasing, we can conclude that the optimal solution does not exist and current wireless resources are not enough to support the specified statistical delay QoS for the incoming multi-layer video stream.

To find the optimal $\boldsymbol{\lambda}^*$, we need the pdf of $\boldsymbol{\gamma}$. In realistic systems, although the pdf of $\boldsymbol{\gamma}$ is usually unknown, we can still apply Eq. (5.20) to implement online tracking. In particular, the iterative update of Eq. (5.20) will be performed in each time frame. However, the expectation of $e^{-\beta_\ell t_\ell \log(1+\gamma)}$ in Eq. (5.20) needs to be substituted by its estimation obtained based on the statistics from previous time frames. Denoting the estimation of $\mathbb{E}_\gamma \left\{e^{-\beta_\ell t_\ell \log(1+\gamma)}\right\}$ in the kth time frame by $\mathcal{S}_\ell[k]$, we obtain $\mathcal{S}_\ell[k+1]$ through a first-order autoregressive filter (low-pass filter) as follows:

$$\mathcal{S}_\ell[k+1] := (1-\alpha)\mathcal{S}_\ell[k] + \alpha e^{-\beta_\ell t_\ell[k+1] \log(1+\gamma[k+1])}, \tag{5.21}$$

where $\ell = 1, 2, \ldots, L$ and $\alpha \in (0, 1)$ is a small positive number close to 0. If the optimal solution exists, the online tracking method converges with properly selected α and ϵ. Section F will present some examples of tracking the optimal Lagrangian

multipliers through simulations.

2. Unicasting Layered Video Stream With Loss Tolerance

When $q_{\text{th}}^{(\ell)} > 0$, the transmission strategy becomes more complicated, but will use less wireless resources. After integrating the pre-drop strategy, the loss rate q_ℓ of the ℓth layer is derived as

$$q_\ell = 1 - \frac{BT\mathbb{E}_\gamma\{t_\ell R_\ell\}}{\mathbb{E}_\gamma\{C_\ell\}} = 1 - \frac{\mathbb{E}_\gamma\{t_\ell R_\ell\}}{\mathbb{E}_\gamma\{t_\ell R_\ell + z_\ell\}}. \tag{5.22}$$

Next, we identify the adaptive transmission policy by solving optimization problem **V-*P*2**:

$$\begin{aligned}
&\textbf{V-}\boldsymbol{P}\textbf{2}: \text{ Unicast with loss tolerance} \\
&\min_{(\boldsymbol{t},\boldsymbol{z})} \left\{ \sum_{\ell=1}^{L} \mathbb{E}_\gamma \{t_\ell\} \right\} && (5.23) \\
&\text{s.t.:} \quad \mathbb{E}_\gamma \left\{ e^{-\beta_\ell (z_\ell + t_\ell \log(1+\gamma))} \right\} - V_\ell \le 0, \ z_\ell \ge 0, \ \forall\, \ell, && (5.24) \\
&\qquad q_\ell \le q_{\text{th}}^{(\ell)}, \ \forall\, \ell && (5.25) \\
&\qquad \sum_{\ell=1}^{L} t_\ell \le 1, \ 0 \le t_\ell \le 1, \quad \forall\, \gamma, && (5.26)
\end{aligned}$$

where $\boldsymbol{z} \triangleq (z_1, z_2, \ldots, z_\ell)$.

Applying Eq. (5.22), we can rewrite Eq. (5.25) as follows:

$$\left(1 - q_{\text{th}}^{(\ell)}\right) \mathbb{E}_\gamma \{z_\ell\} - q_{\text{th}}^{(\ell)} \mathbb{E}_\gamma \{t_\ell \log(1+\gamma)\} \le 0, \ \forall\, \ell. \tag{5.27}$$

It is also not hard to prove that problem **V-*P*2** is still a convex problem and the Lagrangian method is still effective in finding the optimal solutions, which is summarized in Theorem 8.

Theorem 8. *The optimal solution* $(\boldsymbol{t}^*, \boldsymbol{z}^*)$*, if existing, is expressed by a set of func-*

tions of γ, $\{\lambda_\ell\}_{\ell=1}^L$, $\{\phi_\ell\}_{\ell=1}^L$, *and* ψ_γ *as follows:*

$$t_\ell^* = t_\ell(\gamma, \lambda_\ell^*, \phi_\ell^*, \psi_\gamma^*), \quad z_\ell^* = z_\ell(\gamma, \lambda_\ell^*, \phi_\ell^*, \psi_\gamma^*), \quad \forall\, \ell, \tag{5.28}$$

where

$$t_\ell(\gamma, \lambda_\ell, \phi_\ell, \psi_\gamma) \triangleq \begin{cases} \infty, & \text{if } -\infty < \frac{1+\psi_\gamma}{\log(1+\gamma)} \leq \phi_\ell q_{\text{th}}^{(\ell)}; \\ \left[-\frac{1}{\beta_\ell \log(1+\gamma)} \log\left(\frac{1+\psi_\gamma - q_{\text{th}}^{(\ell)} \phi_\ell \log(1+\gamma)}{\beta_\ell \lambda_\ell \log(1+\gamma)}\right)\right]^+, & \text{if } \phi_\ell q_{\text{th}}^{(\ell)} < \frac{1+\psi_\gamma}{\log(1+\gamma)} < \phi_\ell; \\ 0, & \text{if } \phi_\ell \leq \frac{1+\psi_\gamma}{\log(1+\gamma)} < \infty, \end{cases} \tag{5.29}$$

and

$$z_\ell(\gamma, \lambda_\ell, \phi_\ell, \psi_\gamma) \triangleq \begin{cases} 0, & \text{if } -\infty < \frac{1+\psi_\gamma}{\log(1+\gamma)} < \phi_\ell\,; \\ \left[-\frac{1}{\beta_\ell} \log\left(\frac{\phi\left(1-q_{\text{th}}^{(\ell)}\right)}{\beta_\ell \lambda_\ell}\right)\right]^+, & \text{if } \phi_\ell \leq \frac{1+\psi_\gamma}{\log(1+\gamma)} < \infty. \end{cases} \tag{5.30}$$

Given γ, $\{\lambda_\ell^*\}_{\ell=1}^L$, *and* $\{\phi_\ell^*\}_{\ell=1}^L$, *if* $\sum_{\ell=1}^L t_\ell(\gamma, \lambda_\ell^*, \phi_\ell^*, 0) \geq 1$, ψ_γ^* *is selected such that the equation* $\sum_{\ell=1}^L t_\ell(\gamma, \lambda_\ell^*, \phi_\ell^*, \psi_\gamma^*) = 1$ *holds; otherwise, we have* $\psi_\gamma^* = 0$. *The optimal* $\{\lambda_\ell^*\}_{\ell=1}^L$ *and* $\{\phi_\ell^*\}_{\ell=1}^L$ *need to be jointly selected such that "=" holds in both Eqs. (5.24) and (5.27).*

Proof. The proof of Theorem 8 can be readily obtained by using the standard Lagrangian-multiplier based method and KKT conditions. □

In order to search for the optimal Lagrangian multipliers and check the existence of the optimal solution, we can also design the adaptive tracking method similar to problem **V-*P*1**.

E. QoS Guarantees for Multicasting Layered-Video Stream

We consider the multicast scenario in this section. If no loss is tolerated, the transmission rate in each time frame is limited by the worst-case SNR among all multicast receivers. Thus, the system throughput will be degraded very quickly as the multicast group size increases. Therefore, we mainly focus on the multicast scenario with loss tolerance.

1. Problem Formulation for Multicast Scenario

Under the multicast rate-adaptation strategy given in Section C, the loss rate of the nth receiver at the ℓth video layer becomes

$$q_{\ell,n} = 1 - \frac{\mathbb{E}_{\boldsymbol{\gamma}}\left\{t_\ell \sum_{i=1}^{N} t_{\ell,i} \log(1+\gamma_i)\delta_{\gamma_n \geq \gamma_i}\right\}}{\mathbb{E}_{\boldsymbol{\gamma}}\left\{z_\ell + t_\ell R_\ell\right\}}, \tag{5.31}$$

where $\mathbb{E}_{\boldsymbol{\gamma}}\{\cdot\}$ denotes the expectation over all fading states of the random vector variable $\boldsymbol{\gamma}$, $\delta_{\gamma_n \geq \gamma_i}$ is the indication function (for a given statement ϖ, $\delta(\varpi) = 1$ if ϖ is true, and $\delta(\varpi) = 0$ otherwise), and $R_\ell = \sum_{i=1}^{N} t_{\ell,i} \log(1+\gamma_i)$ is the total normalized transmission rate in a time frame (see Eq. (5.3)). Accordingly, the following loss-rate constraint needs to be satisfied for each multicast receiver at every video layer, which is specified by the inequality as follows:

$$q_{\ell,n} \leq q_{\text{th}}^{(\ell)}, \quad \forall\, n,\ \forall\, \ell. \tag{5.32}$$

To simplify the derivations, we first use a *relaxed* constraint to replace Eq. (5.32) by:

$$q_{\ell,0} \triangleq \frac{1}{N} \sum_{n=1}^{N} q_{\ell,n} \leq q_{\text{th}}^{(\ell)}, \quad \forall\, \ell \tag{5.33}$$

where $q_{\ell,0}$ is called *group loss rate* (average loss rate over receivers) at the ℓth video layer. We will show later that the optimal adaptation policy derived under the group-

loss-rate constraint given in Eq. (5.33) does not violate Eq. (5.32), and thus is also optimal under the original loss-rate constraint given by Eq. (5.32). Plugging Eq. (5.31) into Eq. (5.33), we have

$$q_{\ell,0} = 1 - \frac{\mathbb{E}_{\boldsymbol{\gamma}} \left\{ t_\ell \sum_{i=1}^{N} t_{\ell,i} m_i \log(1+\gamma_i) \right\}}{N \mathbb{E}_{\boldsymbol{\gamma}} \left\{ z_\ell + t_\ell R_\ell \right\}}, \tag{5.34}$$

where m_i is the number of receivers with SNR higher than or equal to γ_i. In addition, it is clear that R_ℓ falls in the following range:

$$R_\ell \in \left[R_{\pi(N)}, R_{\pi(1)} \right], \tag{5.35}$$

where $R_{\pi(N)} \triangleq \min_{1 \le n \le N} \{\log(1+\gamma_n)\}$ and $R_{\pi(1)} \triangleq \max_{1 \le n \le N} \{\log(1+\gamma_n)\}$. Note that when we attempt to use a normalized transmission rate equal to R_ℓ in a time frame, there are many different choices for $\{t_{\ell,n}\}_{n=1}^{N}$ to get the same R_ℓ. In order to minimize the loss for the entire multicast group, among all these choices we need to select the one which maximizes the numerator of the second term on the right-hand side of Eq. (5.34), which represents the sum rate of data correctly received by each multicast receiver. Accordingly, we define

$$\begin{aligned} \widetilde{g}_s(R_\ell) \triangleq & \max_{\boldsymbol{t}_\ell:\ \sum_{i=1}^{N} t_{\ell,i}=1} \left\{ \sum_{i=1}^{N} t_{\ell,i} m_i \log(1+\gamma_i) \right\} \\ & \text{s.t.:} \sum_{i=1}^{N} t_{\ell,i} \log(1+\gamma_i) = R_\ell \end{aligned} \tag{5.36}$$

where $\boldsymbol{t}_\ell \triangleq (t_{\ell,1}, t_{\ell,2}, \ldots, t_{\ell,N})$.

Therefore, $\widetilde{g}_s(R_\ell)$ denotes the maximum sum of achieved rates over all multicast receivers under the given normalized transmission rate R_ℓ. In Chapter IV, we showed that $\widetilde{g}_s(R_\ell)$ can be derived through the concept of convex hull [77]. Using the properties of convex hull, in [76] we proved that $\widetilde{g}_s(R_\ell)$ is a *continuous*, *piecewise linear*,

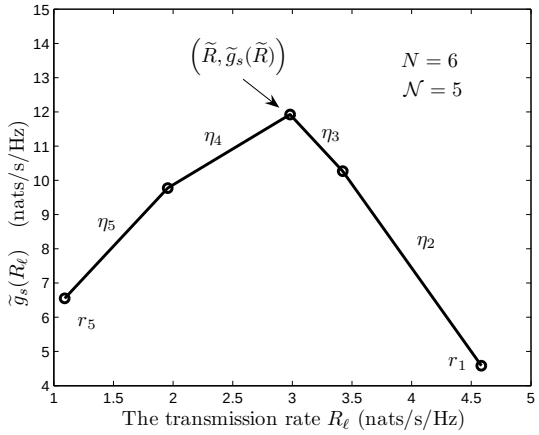


Fig. 22. An example for the function $\widetilde{g}_s(R_\ell)$ against R_ℓ, where $N = 6$ and $\boldsymbol{\gamma} = (1.98, 6.07, 18.71, 29.63, 45.43, 96.93)$.

and *concave function* over R_ℓ. Thus, we can obtain $\widetilde{g}_s(R_\ell)$ as follows:

$$\widetilde{g}_s(R_\ell) = \begin{cases} \widetilde{g}_s(r_i) + \eta_i(R_\ell - r_i), & \text{if } R_\ell \in [r_i, r_{i-1}),\ 2 \le i \le \mathcal{N}; \\ \widetilde{g}_s(r_2) + \eta_2(r_1 - r_2), & \text{if } R_\ell = r_1, \end{cases} \tag{5.37}$$

where $R_{\pi(1)} = r_1 > r_2 > \cdots > r_{\mathcal{N}} = R_{\pi(N)}$. Fig. 22 depicts an example for the function $\widetilde{g}_s(R_\ell)$. As shown in Fig. 22, within each interval $[r_i, r_{i-1})$, $\widetilde{g}_s(R_\ell)$ is a linear function of R_ℓ with the slope equal to η_i, and $(\mathcal{N}-1)$ is equal to the number of such intervals. Note that $\{(r_i, \widetilde{g}_s(r_i))\}_{i=1}^{\mathcal{N}}$ are actually the vertices on the upper boundary of the convex hull of the 2-dimensional point set $\left\{(\log(1+\gamma_i), m_i \log(1+\gamma_i))\right\}_{i=1}^{N}$ (see Chapter IV). For the complete procedures to identify $\widetilde{g}_s(R_\ell)$ and the corresponding time slot allocation policy, please refer to Chapter IV. The above discussions imply that we need to consider only the transmission policies yielding $\widetilde{g}_s(R_\ell)$, because only these policies can minimize the total loss for the entire multicast group. Moreover, Eqs. (5.36)-(5.37) suggest that we can focus on regulating the scalar R_ℓ instead of the N-dimension time-proportion vector $\boldsymbol{t}_\ell$. After R_ℓ is determined, we can use Eq. (5.36) to obtain $\boldsymbol{t}_\ell$.

Following previous analyses in this section, we formulate problem **V-*P*3** to derive the adaptation policy for multi-layer video multicast as follows:

$$\begin{aligned}
&\textbf{V-}\boldsymbol{P3}: \text{ Multicast with loss tolerance}\\
&\min_{(\boldsymbol{t},\boldsymbol{R},\boldsymbol{z})} \left\{ \sum_{\ell=1}^{L} \mathbb{E}_{\boldsymbol{\gamma}} \{t_\ell\} \right\} && (5.38)\\
&\text{s.t.: } \mathbb{E}_{\boldsymbol{\gamma}} \left\{ e^{-\beta_\ell (z_\ell + t_\ell R_\ell)} \right\} - V_\ell \le 0, \ z_\ell \ge 0, \ \forall\, \ell, && (5.39)\\
&\qquad N \left(1 - q_{\text{th}}^{(\ell)}\right) \mathbb{E}_{\boldsymbol{\gamma}} \{t_\ell R_\ell + z_\ell\} - \mathbb{E}_{\boldsymbol{\gamma}} \{t_\ell \widetilde{g}_{\text{s}}(R_\ell)\} \le 0, \forall\, \ell, && (5.40)\\
&\qquad \sum_{\ell=1}^{L} t_\ell - 1 \le 0, \ 0 \le t_\ell \le 1, \quad \forall\, \boldsymbol{\gamma}, && (5.41)
\end{aligned}$$

where $\boldsymbol{R} \triangleq (R_1, R_2, \ldots, R_L)$ and Eq. (5.40) is the group-loss-rate constraint (equivalent to Eq. (5.33)) for the policies corresponding to $\widetilde{g}_{\text{s}}(R_\ell)$.

2. Derivation of the Optimal Solution for Multicast Video

Notice that Problem **V-*P*3** is not convex, because the functions on the left-hand side of Eqs. (5.39) and (5.40) are not convex over $(\boldsymbol{t}, \boldsymbol{R}, \boldsymbol{z})$. However, we show in the following that the optimal solution can still be obtained through Lagrangian dual problem.

2.1 Lagrangian Characterization of Problem **V-*P*3**

The Lagrangian function of **V-*P*3**, denoted by W, is constructed as

$$W = \mathbb{E}_{\boldsymbol{\gamma}}\{w\} \tag{5.42}$$

where

$$\begin{aligned}
w \triangleq \sum_{\ell=1}^{L} t_\ell + \psi_{\boldsymbol{\gamma}} \left(\sum_{\ell=1}^{L} t_\ell - 1 \right) + \sum_{\ell=1}^{L} \lambda_\ell \left(e^{-\beta_\ell (z_\ell + t_\ell R_\ell)} - V_\ell \right)\\
+ \sum_{\ell=1}^{L} \phi_\ell \left(N \left(1 - q_{\text{th}}^{(\ell)}\right) (t_\ell R_\ell + z_\ell) - t_\ell \widetilde{g}_{\text{s}}(R_\ell) \right). \qquad (5.43)
\end{aligned}$$

In Eq. (5.43), $\psi_{\boldsymbol{\gamma}} \geq 0$, $\phi_\ell \geq 0$, and $\lambda_\ell \geq 0$ are the Lagrangian multipliers associated with the constraints given by Eqs. (5.41), (5.40), and (5.39), respectively. The Lagrangian dual function, denoted by U, is then determined by

$$U(\boldsymbol{\lambda}, \boldsymbol{\phi}, \psi_{\boldsymbol{\gamma}}) \triangleq \min_{(\boldsymbol{t},\boldsymbol{z},\boldsymbol{R})} \Big\{ W \Big\} = \mathbb{E}_{\boldsymbol{\gamma}} \Big\{ u(\boldsymbol{\lambda}, \boldsymbol{\phi}, \psi_{\boldsymbol{\gamma}}) \Big\}, \tag{5.44}$$

where $\boldsymbol{\phi} \triangleq (\phi_1, \phi_2, \ldots, \phi_L)$ and

$$u(\boldsymbol{\lambda}, \boldsymbol{\phi}, \psi_{\boldsymbol{\gamma}}) \triangleq \min_{(\boldsymbol{t},\boldsymbol{z},\boldsymbol{R})} \Big\{ w \Big\}. \tag{5.45}$$

It is clear that $u(\boldsymbol{\lambda}, \boldsymbol{\phi}, \psi_{\boldsymbol{\gamma}})$ is a concave function over $(\boldsymbol{\lambda}, \boldsymbol{\phi}, \psi_{\boldsymbol{\gamma}})$, and so is $U(\boldsymbol{\lambda}, \boldsymbol{\phi}, \psi_{\boldsymbol{\gamma}})$ Moreover, the Lagrange dual problem is defined as:

$$\textbf{V-}\boldsymbol{P}\textbf{3-Dual}: \; U^\star \triangleq U(\boldsymbol{\lambda}^\star, \boldsymbol{\phi}^\star, \psi_{\boldsymbol{\gamma}}^\star) = \max_{(\boldsymbol{\lambda},\boldsymbol{\phi},\psi_{\boldsymbol{\gamma}})} \Big\{ U(\boldsymbol{\lambda}, \boldsymbol{\phi}, \psi_{\boldsymbol{\gamma}}) \Big\}, \tag{5.46}$$

where $\big(\boldsymbol{\lambda}^\star, \boldsymbol{\phi}^\star, \psi_{\boldsymbol{\gamma}}^\star\big)$ is the maximizer. We then solve for the optimal adaptation strategy to **V-*P*3** through the dual problem. If the optimal solution to **V-*P*3** exists, we will show later that there is no duality gap between the primal problem **V-*P*3** and the dual problem **V-*P*3-Dual**. As a result, the optimal solution to **V-*P*1** must minimize the Lagrangian function W under the optimal Lagrangian multipliers $\big(\boldsymbol{\lambda}^\star, \boldsymbol{\phi}^\star, \psi_{\boldsymbol{\gamma}}^\star\big)$.

2.2 Derivation of the Lagrangian Dual Function $U(\boldsymbol{\lambda}, \boldsymbol{\phi}, \psi_{\boldsymbol{\gamma}}) = \mathbb{E}_{\boldsymbol{\gamma}}\big\{u(\boldsymbol{\lambda}, \boldsymbol{\phi}, \psi_{\boldsymbol{\gamma}})\big\}$

Since $\widetilde{g}_s(R_\ell)$ is nondifferentiable at some R_ℓ (as shown in Fig. 22), w is also nondifferentiable at some R_ℓ. Alternatively, we need to use the subgradient and subdifferential (see Definition 9) instead of gradient to derive the minimizer to Eq. (5.45), which is denoted by $(\boldsymbol{t}^*, \boldsymbol{z}^*, \boldsymbol{R}^*)$.

Then, using Eq. (5.36) and Definition 9, we obtain

$$\partial\widetilde{g}_{\mathrm{s}}(r) = \begin{cases} \big[\eta_{\mathcal{N}}, \infty\big], & \text{if } r = r_{\mathcal{N}}; \\ \big\{\eta_i\big\}, & \text{if } r_i < r < r_{i-1},\ 2 \le i \le \mathcal{N}; \\ \big[\eta_i, \eta_{i+1}\big], & \text{if } r = r_i,\ 2 \le i \le \mathcal{N}-1; \\ \big[-\infty, \eta_1\big], & \text{if } r = r_1. \end{cases} \tag{5.47}$$

Applying Eq. (5.43) into Definition 9, we get the subdifferential of w w.r.t. R_ℓ, denoted by ∂w_{R_ℓ}, as:

$$\partial w_{R_\ell} = \Big\{ y \,\Big|\, y = t_\ell \Big(\phi_\ell N \left(1 - q_{\mathrm{th}}^{(\ell)}\right) - \lambda_\ell \beta_\ell e^{-\beta_\ell (z_\ell + t_\ell R_\ell)} - \phi_\ell x \Big),\ \forall\, x \in \partial\widetilde{g}_{\mathrm{s}}(R_\ell) \Big\},\ \forall\, \ell, \boldsymbol{\gamma}. \tag{5.48}$$

It is clear that w is differentiable w.r.t. t_ℓ and R_ℓ. Taking the derivative of w w.r.t. t_ℓ and z_ℓ, respectively, we get

$$\frac{\partial w}{\partial t_\ell} = 1 + \psi_{\boldsymbol{\gamma}} - \lambda_\ell \beta_\ell R_\ell e^{-\beta_\ell (z_\ell + t_\ell R_\ell)} - \phi_\ell \widetilde{g}_{\mathrm{s}}(R_\ell) + \phi_\ell N \left(1 - q_{\mathrm{th}}^{(\ell)}\right) R_\ell,\ \forall\, \ell, \boldsymbol{\gamma}; \tag{5.49}$$

$$\frac{\partial w}{\partial z_\ell} = \phi_\ell N \left(1 - q_{\mathrm{th}}^{(\ell)}\right) - \lambda_\ell \beta_\ell e^{-\beta_\ell (z_\ell + t_\ell R_\ell)},\ \forall\, \ell, \boldsymbol{\gamma}. \tag{5.50}$$

Clearly, the minimization of w can be performed separately for each video layer. Now consider the ℓth video layer. Since the function w is not convex over the 3-tuple (t_ℓ, z_ℓ, R_ℓ), the equations $\partial w/t_\ell = 0$, $\partial w/z_\ell = 0$, and $0 \in \partial w_{R_\ell}$ are only the necessary conditions for $(t_\ell^*, z_\ell^*, R_\ell^*)$. However, if t_ℓ^* is given, w becomes a convex function over the 2-tuple (z_ℓ, R_ℓ). Using this property, we can decompose the minimization of w into several easier sub-problems. Applying the above principle, we discuss the cases with the fixed $t_\ell^* = 0$ and $t_\ell^* > 0$, respectively, as follows.

a. $t_\ell^* = 0$

The variable R_ℓ vanishes in Eq. (5.43) when $t_\ell = 0$. Then, we only need to find the minimizer z_ℓ^*. By solving $\partial w/\partial z_\ell = 0$ under the condition $z_\ell \geq 0$, we get

$$z_\ell^* = \left[-\frac{1}{\beta_\ell} \log \left(\frac{\phi_\ell N \left(1 - q_{\text{th}}^{(\ell)}\right)}{\lambda_\ell \beta_\ell} \right) \right]^+ . \tag{5.51}$$

b. $t_\ell^* > 0$

We jointly solve $\partial w/\partial z_\ell = 0$ and $0 \in \partial w_{R_\ell}$ under the condition $z_\ell \geq 0$ and $R_\ell \geq 0$, and then get the minimizer, which is summarized in Eqs. (5.52)-(5.53) as follows:

$$\begin{aligned} &\text{if } \widehat{R}_\ell > \widetilde{R}, \text{ then} \\ &\qquad \begin{cases} R_\ell^* = \widetilde{R}; \\ z_\ell^* = \left[-\frac{1}{\beta_\ell} \log \left(\frac{\phi_\ell N \left(1 - q_{\text{th}}^{(\ell)}\right)}{\lambda_\ell \beta_\ell} \right) - t_\ell^* R_\ell^* \right]^+; \end{cases} \end{aligned} \tag{5.52}$$

$$\begin{aligned} &\text{if } \widehat{R}_\ell \leq \widetilde{R}, \text{ then} \\ &\qquad \begin{cases} R_\ell^* = \widehat{R}_\ell; \\ z_\ell^* = 0, \end{cases} \end{aligned} \tag{5.53}$$

where $\widehat{R}_\ell$ is the unique solution to

$$0 \in \left(\partial w_{R_\ell} \right) |_{z_\ell = 0} \tag{5.54}$$

under the given t_ℓ, and

$$\widetilde{R} \triangleq \arg\max_r \left\{ \widetilde{g}_{\text{s}}(r) \right\}. \tag{5.55}$$

The detailed derivations for Eqs. (5.52) and (5.53) are provided in Appendix M.

Note that $\widetilde{R}$ depends only on $\widetilde{g}_{\text{s}}(r)$, but not on t_ℓ^*. Then, through Eqs. (5.52)-(5.53), we can see that with $t_\ell^* > 0$, the minimizer must satisfy either $z_\ell^* = 0$ or $R_\ell^* = \widetilde{R}$. Further note that the above results provide not only the mathematical

convenience, but also the insightful observations for the adaptive multicast transmission. For multicast, zero loss can be achieved only when setting transmission rate $R_\ell = R_{\pi(N)}$, which is determined by the worst-case SNR over all multicast receivers. The higher the transmission rate or the higher the drop rate we use, the higher the loss rate we get (observed from Eq. (5.34)). Therefore, when not violating the statistical delay QoS guarantees, we need to choose the transmission rate R_ℓ and the drop rate z_ℓ as small as possible. Moreover, R_ℓ and z_ℓ need to be jointly derived for performance optimization. Following the above strategies, we first derive $\widehat{R}_\ell$ which optimizes the system performance (equivalently, minimizes the Lagrangian function) given the zero drop rate. However, if $\widehat{R}_\ell > \widetilde{R}$, we can see that the achieved sum rate $\widetilde{g}_s(R_\ell)$ over all multicast receivers under $R_\ell = \widehat{R}_\ell$ decreases when $\widehat{R}_\ell$ increases, as depicted in Fig. 22. When this happens, we need to set $R_\ell = \widetilde{R}$ and apply the nonzero drop rate to avoid the degradation of $\widetilde{g}_s(R_\ell)$ while supporting the satisfied service rate.

Based on the above results for $t_\ell^* = 0$ and $t_\ell^* > 0$, the minimizer $(t_\ell^*, R_\ell^*, z_\ell^*)$ at the ℓth video layer must fall into one of the following three Sub-domains:

$$\begin{cases} \text{Sub-domain 1:} & t_\ell = 0,\ R_\ell \geq 0,\ z_\ell \geq 0; \\ \text{Sub-domain 2:} & t_\ell \geq 0,\ R_\ell = \widetilde{R},\ z_\ell \geq 0; \\ \text{Sub-domain 3:} & t_\ell \geq 0,\ R_\ell \geq 0,\ z_\ell = 0. \end{cases} \tag{5.56}$$

In Eq. (5.56), Sub-domain 1 is associated with the case of $t_\ell^* = 0$. For the case with $t_\ell^* > 0$, Sub-domains 2 and 3 correspond to the conditions $\widehat{R}_\ell \geq \widetilde{R}$ and $\widehat{R}_\ell < \widetilde{R}$, respectively. In order to get the minimizer $(t_\ell^*, R_\ell^*, z_\ell^*)$ of w, we can first find the minimizer within each Sub-domain, which is denoted by $\left(t_\ell^{(j)}, R_\ell^{(j)}, z_\ell^{(j)}\right)$, $j = 1, 2, 3$. After identifying the minimizers of each Sub-domain, $(t_\ell^*, R_\ell^*, z_\ell^*)$ can then be obtained

through

$$(t_\ell^*, R_\ell^*, z_\ell^*) = \left(t_\ell^{(j^*)}, R_\ell^{(j^*)}, z_\ell^{(j^*)}\right) \quad \forall\, \ell, \tag{5.57}$$

where

$$j^* = \arg\min_{j=1,2,3} \left\{ w\big|_{(t_\ell, R_\ell, z_\ell)=\left(t_\ell^{(j)}, R_\ell^{(j)}, z_\ell^{(j)}\right)} \right\}.$$

In Sub-domains 1, 2, and 3, the variables t_ℓ, R_ℓ, and z_ℓ are fixed, respectively, implying that there are only two optimization variables in each Sub-domain. Therefore, the minimization problem within each Sub-domain becomes tractable. For Sub-domain 1, the minimizer $\left(0, R_\ell^{(1)}, z_\ell^{(1)}\right)$ is given in Eq. (5.51). For Sub-domain 2, since R_ℓ is fixed, deriving the minimizer $\left(t_\ell^{(2)}, R_\ell^{(2)}, z_\ell^{(2)}\right)$ is equivalent to solving a convex problem. For Sub-domain 3, the optimization problem can be readily solved by applying the piecewise linear property of $\widetilde{g}(R_\ell)$. The detailed derivations for $\left(t_\ell^{(2)}, R_\ell^{(2)}, z_\ell^{(2)}\right)$ and $\left(t_\ell^{(3)}, R_\ell^{(3)}, z_\ell^{(3)}\right)$ are given in Appendix N.

2.3. The Optimal Solution to **V-*P3***

In Section E-2.2, we have obtained the minimizer $(\boldsymbol{t}^*, \boldsymbol{z}^*, \boldsymbol{R}^*)$ for w. Then, based on the optimization theory [79], the necessary and sufficient conditions for zero duality gap are as follows: there exists the feasible policy $(\boldsymbol{t}^*, \boldsymbol{z}^*, \boldsymbol{R}^*)|_{\{\phi_\gamma=\phi_\gamma^\star, \lambda_\ell=\lambda_\ell^\star, \phi_\ell=\phi_\ell^\star,\ \forall\, \ell,\gamma\}}$ such that

$$\begin{cases} \psi_{\boldsymbol{\gamma}}^\star \left(\sum_{\ell=1}^L t_\ell^* - 1\right) = 0,\ \forall\, \ell,\ \gamma; \\ \lambda_\ell^\star \left(\mathbb{E}_{\boldsymbol{\gamma}} \left\{ e^{-\beta_\ell (z_\ell^* + t_\ell^* R_\ell^*)} \right\} - V_\ell \right) = 0,\ \forall\, \ell; \\ \phi_\ell^\star \mathbb{E}_{\boldsymbol{\gamma}} \left\{ N \left(1 - q_{\text{th}}^{(\ell)}\right) (t_\ell^* R_\ell^* + z_\ell^*) - t_\ell^* \widetilde{g}_{\text{s}}(R_\ell^*) \right\} = 0,\ \forall\, \ell; \\ \psi_{\boldsymbol{\gamma}}^\star \geq 0,\ \lambda_\ell^\star \geq 0,\ \phi_\ell^\star \geq 0. \end{cases} \tag{5.58}$$

Then, the optimal policy to **V-*P*3** is given by

$$(\boldsymbol{t}^*, \boldsymbol{z}^*, \boldsymbol{R}^*)|_{\{\phi_\gamma=\phi^\star_\gamma, \lambda_\ell=\lambda^\star_\ell, \phi_\ell=\phi^\star_\ell,\ \forall\, \ell,\gamma\}}. \tag{5.59}$$

We can solve for the optimal Lagrangian multipliers by using the similar arguments in the proof of Theorem 7. Specifically, for each channel realization if $\sum_{\ell=1}^{L} t^*_\ell|_{\psi_{\boldsymbol{\gamma}}=0} \leq 1$, we have $\psi^\star_{\boldsymbol{\gamma}} = 0$; otherwise, $\psi^\star_{\boldsymbol{\gamma}}$ is the unique solution to $\sum_{\ell=1}^{L} t^*_\ell = 1$. Moreover, $\{\phi^\star_\ell\}_{\ell=1}^{L}$ and $\{\lambda^\star_\ell\}_{\ell=1}^{L}$ will be selected such that "=" holds for constraints given in Eqs. (5.39)-(5.40). Furthermore, we can design the adaptive tracking method similar to problem **V-*P*1** to examine the existence of the optimal solution and find the optimal Lagrangian multipliers.

Note that under the above optimal solution, different $\{\gamma_n\}_{n=1}^{N}$, which have the same ordered permutation, will generate the same function $\widetilde{g}_s(R_\ell)$ defined by Eq. (5.36) and thus the same adaptation policy. Then, since γ_n's are i.i.d. (as assumed in Section B), this policy will benefit all receivers evenly, implying $q_{\ell,0} = q_{\ell,1} = q_{\ell,2} = \cdots = q_{\ell,N} = q^{(\ell)}_{\text{th}}$. Therefore, the original loss-rate constraint is not violated for all multicast receivers. Moreover, since the group-loss-rate constraint given by Eq. (5.40) in problem **V-*P*3** is a relaxed version of the original loss-rate constraint for each multicast receiver (given by Eq. (5.32)), the optimal solution to problem **V-*P*3** is also optimal even if we replace Eq. (5.40) by using the original loss-rate constraint given by Eq. (5.32).

F. Simulation Evaluations

We use simulation experiments to evaluate the performances of our proposed optimal adaptive transmission schemes and to investigate the impact of QoS requirements on resource allocations. Note that the metric "delay" investigated/simulated in simu-

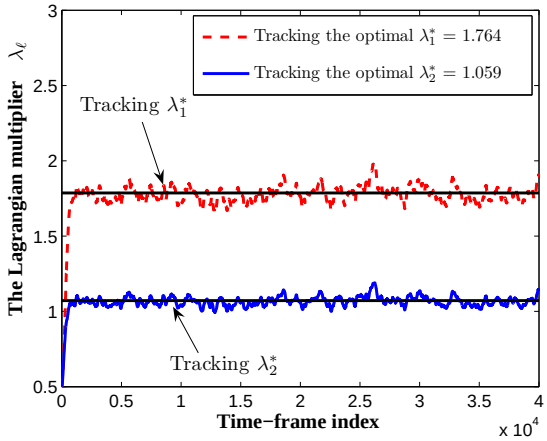


Fig. 23. Illustration of tracking the optimal Lagrangian multiplier λ_ℓ^* for unicast with zero loss, where the average SNR is $\overline{\gamma} = 10$ dB, the required delay bound is $P_{\rm th} = 10^{-4}$, and the required threshold for the delay-bound violation probability is $D_{\rm th} = 250$ ms.

lations represents the queueing delay, as addressed previously in Section C-2 for the framework of this chapter. In simulations, we set the signal bandwidth B and time-frame length T equal to 2×10^5 Hz and 10 ms, respectively. The arrival video stream includes two layers, both of which have constant arrival rates, where $A_1[k] = 250$ Kbps and $A_2[k] = 150$ Kbps. Then, the effective bandwidths of $A_1[k]$ and $A_2[k]$ are determined by $\mathcal{A}_1(\theta_1) = 1.733 \times 10^3$ nats/frame $\mathcal{A}_2(\theta_2) = 1.040 \times 10^3$ nats/frame, respectively. The values of θ_1 and θ_2 can be derived from solving Eq. (2.9), depending on the QoS requirements specified by $D_{\rm th}$ and $P_{\rm th}$, $\ell = 1, 2, \ldots, L$. The wireless channel follows the Rayleigh fading model and we denote the average SNR by $\overline{\gamma}$. Fig. 23 plots the iterative on-line tracking of the optimal Lagrangian multipliers λ_ℓ^*'s based on the method used in Section D-1 with $\epsilon = 0.01$ and $\alpha = 0.02$. As shown in Fig. 23, the Lagrangian multipliers quickly converge to the optimal value and oscillate slightly within the small dynamic ranges, which demonstrates the effectiveness of our tracking method.

We also investigate some straightforward time-slot allocation schemes as the

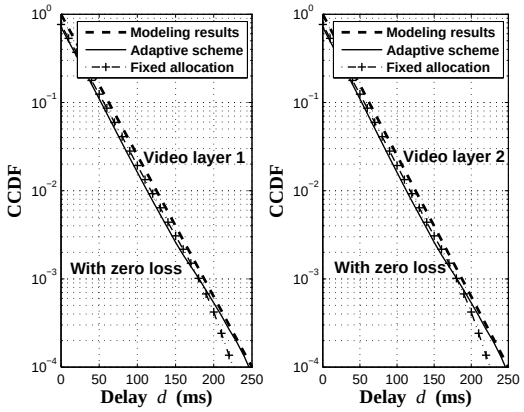


Fig. 24. The complementary cumulative distribution function (CCDF), denoted by $\Pr\{D_\ell > d\}$, of the queueing delay for the unicast scenario with zero loss for video layer 1 and video layer 2, respectively, where $\overline{\gamma} = 15$ dB, $P_{\rm th} = 10^{-4}$, and the required delay-bound is $D_{\rm th} = 250$ ms.

baseline schemes for comparative analyses. We will compare the average resource consumption between our derived optimal schemes and these baseline schemes under the same QoS satisfactions.

Fixed time-slot allocation for unicast without loss:

This scheme uses constant time-slot length $t_\ell[k] = \overline{t}_\ell$, $k = 1, 2, \ldots$, in all fading states. The normalized transmission rate is set to $R_\ell = \log(1+\gamma)$ nats/s/Hz for the ℓth video layer. The parameters $\overline{t}_\ell$, $\ell = 1, 2$, are selected such that the effective capacity $\mathcal{C}_\ell(\theta_\ell)$ of the ℓth video layer's service process is just equal to $\overline{C}_\ell$:

$$\mathcal{C}_\ell(\theta_\ell) = -\frac{1}{\theta_\ell}\log\left(\mathbb{E}\left\{e^{-\theta_\ell \overline{t}_\ell BT\log(1+\gamma)}\right\}\right) = \overline{C}_\ell, \tag{5.60}$$

where $\overline{C}_\ell = \mathcal{A}_\ell(\theta_\ell)$ and $\ell = 1, 2, \ldots, L$. We can obtain the unique $\overline{t}_\ell$'s by numerically solving the above equation. If we get $\sum_{\ell=1}^{L} \overline{t}_\ell > 1$, this scheme cannot guarantee the QoS requirements under current channel conditions, even using up all time-slot resources.

Fixed time-slot allocation for unicast with loss tolerance:

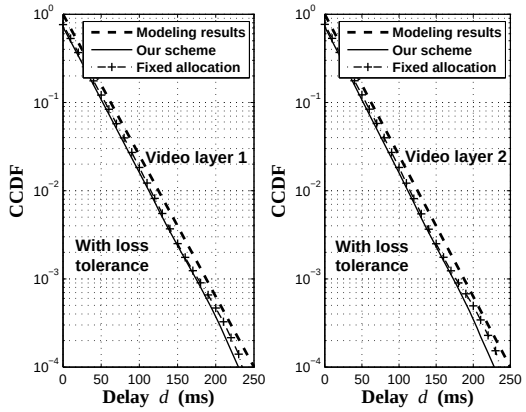


Fig. 25. The CCDF $\Pr\{D_\ell > d\}$ of the queueing delay for the unicast scenario with loss tolerance for video layer 1 and video layer 2, respectively, where $\overline{\gamma} = 15$ dB, $P_{\rm th} = 10^{-4}$, $D_{\rm th} = 250$ ms, and $\left(q_{\rm th}^{(1)}, q_{\rm th}^{(2)}\right) = (0.01, 0.02)$.

This scheme uses both the constant time-slot length $t_\ell[k] = \overline{t}_\ell$ and the constant per-drop rate $z_\ell[k] = \overline{z}_\ell$ in all fading states. The normalized transmission rate is also set to $R_\ell = \log(1+\gamma)$ nats/s/Hz for the ℓth video layer. The parameters $\overline{t}_\ell$ and $\overline{z}_\ell$ are can be obtained by solving

$$\mathcal{C}_\ell(\theta_\ell) = -\frac{1}{\theta_\ell}\log\left(\mathbb{E}\left\{e^{-\theta_\ell\left(\overline{t}_\ell BT\log(1+\gamma)+BT\overline{z}_\ell\right)}\right\}\right) = \overline{C}_\ell \tag{5.61}$$

and $q_\ell = q_{\rm th}^{(\ell)}$ for all video layers, where $\ell = 1, 2, \ldots, L$.

Fixed dominating position scheme for multicast with loss tolerance:

The fixed dominating position (FDP) scheme always sets $R_\ell = \log(1+\gamma_{\pi(i_\ell)})$ nats/s/Hz, where $\gamma_{\pi(i)}$ denotes the ith largest instantaneous SNR among all multicast receivers. The index i_ℓ is fixed at $i_\ell = \left\lceil N\left(1 - q_{\rm th}^{(\ell)}\right)\right\rceil$ such that the loss-rate QoS is not violated. Moreover, the FDP scheme also adopts the constant time-slot length $t_\ell[k] = \overline{t}_\ell$ and the constant per-drop rate $z_\ell[k] = \overline{z}_\ell$, which can be obtained by solving

$$\mathcal{C}_\ell(\theta_\ell) = -\frac{1}{\theta_\ell}\log\left(\mathbb{E}\left\{e^{-\theta_\ell\left(\overline{t}_\ell BT\log(1+\gamma_{\pi(i_\ell)})+BT\overline{z}_\ell\right)}\right\}\right) = \overline{C}_\ell \tag{5.62}$$

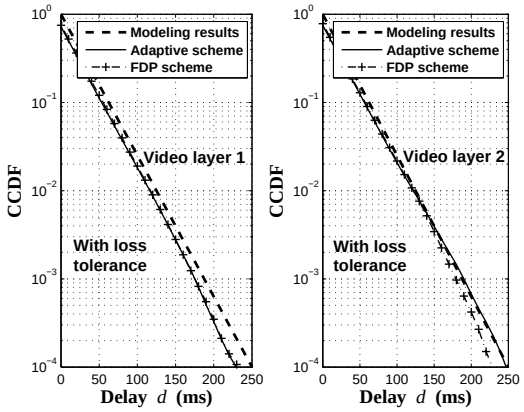


Fig. 26. The CCDF $\Pr\{D_\ell > d\}$ of the queueing delay for multicast scenario with loss tolerance for video layer 1 and video layer 2, respectively, where $N = 20$ receivers, $\overline{\gamma} = 20$ dB, $\left(q_{\rm th}^{(1)}, q_{\rm th}^{(2)}\right) = (0.01, 0.05)$, $P_{\rm th} = 10^{-4}$, and $D_{\rm th} = 250$ ms.

and $q_{\ell,0} = q_{\rm th}^{(\ell)}$ for all video layers, where $\ell = 1, 2, \ldots, L$.

Figures 24, 25, and 26 depict the complementary cumulative distribution function (CCDF) of the queueing delay, i.e., the probability $\Pr\{D_\ell > d\}$ given a threshold d, for unicast with zero loss, unicast with loss tolerance, and multicast with loss tolerance, respectively. We can observe from Figs. 24-26 that the CCDF's of all schemes agree well with the modeling results (see Eqs. (7)-(8)) at each video layer, where the delay-bound violation probability decreases exponentially against the delay bound. Moreover, for the required delay bound $D_{\rm th} = 250$ ms, the violation probability of all schemes can be upper-bounded by the targeted $P_{\rm th} = 10^{-4}$ for each video layer, which demonstrates the validity of all schemes in terms of statistical delay-QoS guarantees. Having shown that all schemes can meet the same QoS requirements, we then focus on the performance of the average time-slot consumption.

Figures 27 and 28 illustrate the impact of delay-bound $D_{\rm th}$ and its violation probability $P_{\rm th}$ on the resource consumption, respectively. As shown in Figs. 27 and 28, either smaller $D_{\rm th}$ or $P_{\rm th}$ will cause more resource consumption, which is expected

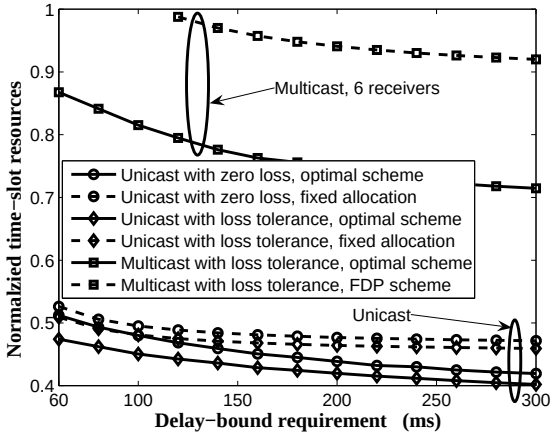


Fig. 27. The normalized time-slot resource consumption $\mathbb{E}_{\boldsymbol{\gamma}}\left\{\sum_{\ell=1}^{L} t_\ell\right\}$ versus delay-bound requirement $D_{\rm th}$, where $\overline{\gamma} = 15$ dB, $P_{\rm th} = 10^{-4}$, and $\left(q_{\rm th}^{(1)}, q_{\rm th}^{(2)}\right) = (0.01, 0.05)$.

because the smaller $D_{\rm th}$ or $P_{\rm th}$ implies the more stringent delay QoS requirement. Moreover, under various QoS conditions, our proposed optimal schemes always use much less wireless resources than the baseline schemes. For multicast services, our derived optimal scheme consumes at least 15% of total resources less than the FDP scheme. For unicast services, more resources can be saved by using the optimal scheme when delay QoS becomes looser, while less resources are saved under the more stringent QoS constraints.

After demonstrating the superiority of our proposed optimal schemes over the baseline schemes, Fig. 29(a) plots the average time-slot consumption versus the average SNR for multi-layer video unicast and multicast. We observe from Fig. 29(a) that under the same channel conditions, video unicast uses much less time-slot resources than multicast. When the average SNR is relatively low around 7-8 dB, video unicast does not need to consume all available time-slot resources. In contrast, video multicast almost uses up *all* resources even with $\overline{\gamma} = 13.5$ dB for the 6-receiver case. For the 10-receiver case, the average SNR needs to be larger than 15 dB to provide the

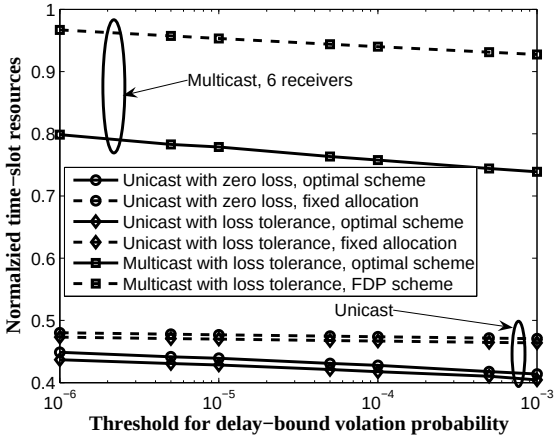
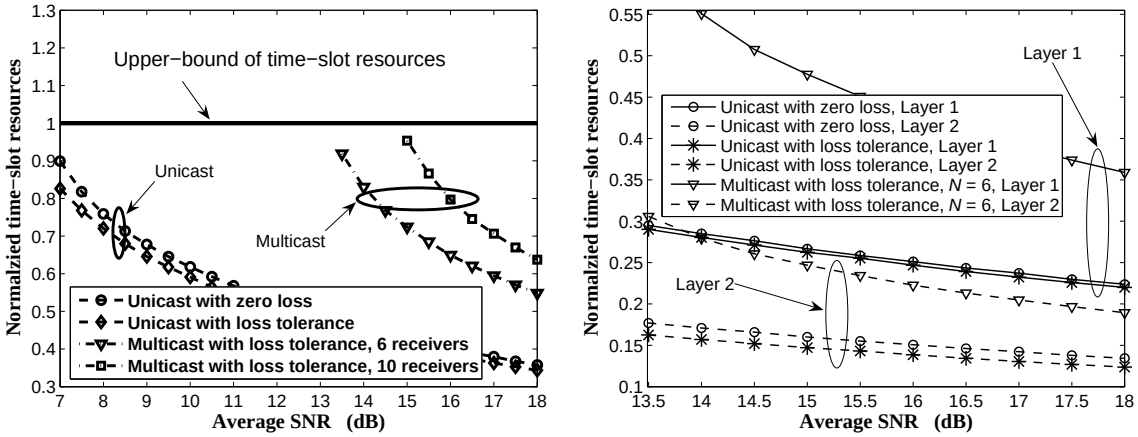


Fig. 28. The normalized time-slot resource consumption $\mathbb{E}_{\boldsymbol{\gamma}}\left\{\sum_{\ell=1}^{L} t_\ell\right\}$ versus the threshold $P_{\rm th}$ for the delay-bound violation probability, where $\overline{\gamma} = 15$ dB, $D_{\rm th} = 250$ ms, and $\left(q_{\rm th}^{(1)}, q_{\rm th}^{(2)}\right) = (0.01, 0.02)$.

QoS-guaranteed multicast services. The above observations reflect the key challenges on wireless multicast. In wireless broadcast channels, since all receivers can hear the sender, it is ideal that only one copy of data is transmitted such that sizable resources can be saved. However, due to the heterogeneous fading channels across multicast receivers, the transmission rate has to be limited within the relatively low range to avoid too much data loss for receivers with poorer instantaneous channel qualities. As a result, more time-slot resources are consumed to meet the QoS requirements. In addition, more multicast receivers result in more resource consumption, as depicted in Fig. 29(a). But note that although the wireless multicast faces many challenges, it still uses much less wireless resources than the strategy which uses multiple unicast links to implement wireless multicast. For example, if using multiple unicast links to implement multicast, we need the time-slot resources at least N times as much as the resource consumption for a unicast link. Clearly, for environments simulated in Fig. 29(a), even with $\overline{\gamma} = 18$ dB, we still do not have enough resources for such a unicast-based multicast scheme with just 6 receivers. Fig. 29(b) shows the resource



(a) (b)

Fig. 29. (a) The normalized time-slot resource consumption $\mathbb{E}_{\boldsymbol{\gamma}}\left\{\sum_{\ell=1}^{L} t_\ell\right\}$ versus $\overline{\gamma}$, where $P_{\rm th} = 10^{-4}$, $D_{\rm th} = 250$ ms, $\left(q_{\rm th}^{(1)}, q_{\rm th}^{(2)}\right) = (0.01, 0.05)$. (b) $\mathbb{E}_{\boldsymbol{\gamma}}\{t_\ell\}$ for each video layer versus $\overline{\gamma}$, where $P_{\rm th} = 10^{-4}$, $D_{\rm th} = 250$ ms, $\left(q_{\rm th}^{(1)}, q_{\rm th}^{(2)}\right) = (0.01, 0.05)$.

consumption for each video layer. We can see that video layer 1 requires more resources than video layer 2, which is because in our settings the traffic load of layer 1 is higher and the loss-rate QoS of layer 1 is more stringent. Furthermore, in multicast the difference of resource consumption between the two video layers is larger than the difference in unicast.

Figure 30 shows the impact from loss-rate constraints on video multicast. As shown in Fig. 30, even slightly increasing $q_{\rm th}^{(\ell)}$ can significantly reduce the total consumed wireless resources. This is because the higher loss-tolerance level will enable larger multicast transmission rate and thus consume less time-slot resources. The above observations suggest that there exists a tradeoff between loss-rate control at the physical layer and error recovery at the upper protocol layers. As mentioned previously, the loss-rate $q_{\rm th}^{(\ell)}$ depends largely on the capability of erasure-correction codes used at upper protocol layers, especially for multicast services. Thus, Fig. 30 sug-

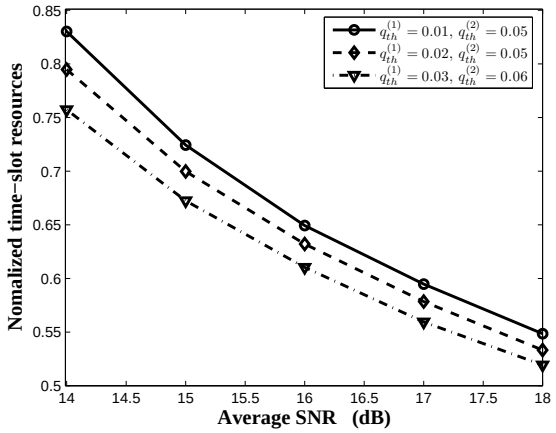


Fig. 30. The normalized time-slot resource consumption $\mathbb{E}_{\boldsymbol{\gamma}}\left\{\sum_{\ell=1}^{L} t_\ell\right\}$ versus $\overline{\gamma}$ for multicast with $N = 6$, $P_{\text{th}} = 10^{-4}$, and $D_{\text{th}} = 250$ ms.

gests that using more redundancy for forward error-control at upper protocol layers can effectively decrease the total wireless-resource consumption with QoS guarantees, while enabling the repair of more data losses at the physical layer.

G. Summary

We proposed a framework to model the wireless transmission of multi-layer video stream with statistical delay QoS guarantees. A separate queue is maintained for each video layer and the same statistical delay QoS-requirement needs to be satisfied by all video layers, where the statistical delay QoS is characterized by the delay-bound and its corresponding violation probability through the effective bandwidth/capacity theory. Under the proposed framework, we derived a set of optimal rate adaptation and time-slot allocation schemes for video unicast/multicast with and/or without loss tolerance, which minimizes the time-slot resource consumption. We also conducted extensive simulation experiments to demonstrate the impact of statistical QoS provisionings on wireless resource allocations by using our derived optimal adaptive transmission schemes.

CHAPTER VI

ADAPTIVE LOW-COMPLEXITY ERASURE-CORRECTING CODE-BASED PROTOCOLS FOR QOS-DRIVEN MOBILE MULTICAST SERVICES

A. Introduction

As in wired and/or unicast networks, error-control not only plays an important role for reliable mobile multicast services over wireless networks, but also provides an efficient means of supporting QoS diversities for different mobile multicast services over different mobile users. However, mobile-multicast imposes many new challenges in error-control for supporting diverse QoS, which are not encountered in wired and/or unicast networks. First, mobile multicast itself causes feedback implosion problems in error-control protocols [32–34]. Second, retransmission-based error-control is not scalable with multicast group size since the retransmission overhead and unnecessary retransmissions grow up quickly as the number of multicast receivers increases [84, 85]. Third, the packets-loss probabilities over wireless-channels vary dramatically when user mobility vary significantly and hand-offs occur frequently. Finally, wireless channels are highly asymmetric where the energy/processing power on uplink from mobile users are much less than that on downlink from the base station. Clearly, the problem on how to efficiently integrate the error control with supporting the QoS diversity for mobile multicast, despite its vital importance, has been neither well understood nor thoroughly studied.

There are mainly two categories of error-control techniques - Automatic Repeat request (ARQ) and Forward Error Correction (FEC) erasure coding. ARQ attempts to retransmit the lost packets while FEC adds the error-control redundancy into the packet flow such that the receivers recover from packet losses without sending error-

control feedback to the sender for retransmission. Clearly, FEC is more suitable to the error-control over mobile multicast services since it can avoid feedback implosion, scale well with multicast-tree size, and significantly reduce the feedback cost of precious energy-/processing- power at mobile users. In addition, with FEC any *one* repairing packet can repair the loss of different data packets at different multicast receivers [67] since FEC is a type of packet *sequence-number independent* error control. As a result, a significant amount of research for error-control in multicast has mainly focused on the FEC-based schemes [67].

Most previous FEC-based multicast error-control schemes for multicasting adaptations mainly focused on the use of the Reed-Solomon Erasure (RSE) codes [67, 86]. However, there are several severe problems inherently associated with RSE-based schemes when they are applied in mobile multicast. First, the error-control redundancy level needs to be dynamically regulated according to the variation of wireless-channels' qualities. Second, the maximum error-control redundancy is upper-bounded by the RSE-code symbol size, which may lead to decoding failures when the wireless-channel loss probabilities increases tremendously. Third, RSE codes' fixed code structures and decoding algorithm cannot be adjusted according to the QoS variations of multicast mobile users. Finally and more importantly, the implementation complexity of RSE coding is too high, particularly when RSE block and symbol sizes are large, to be applicable to the mobile multicast networks where both energy and processing powers are severely constrained at mobile users. To overcome these aforementioned problems, we propose a new adaptive low-complexity graph-code-based hybrid ARQ-FEC scheme for QoS-driven mobile multicast services. The main features of our proposed scheme are two-fold: the low complexity and dynamic adaptation to the variations of packet-loss level and QoS requirements of multicast mobile users. In addition, unlike the existing RSE-code-based schemes, our proposed scheme can

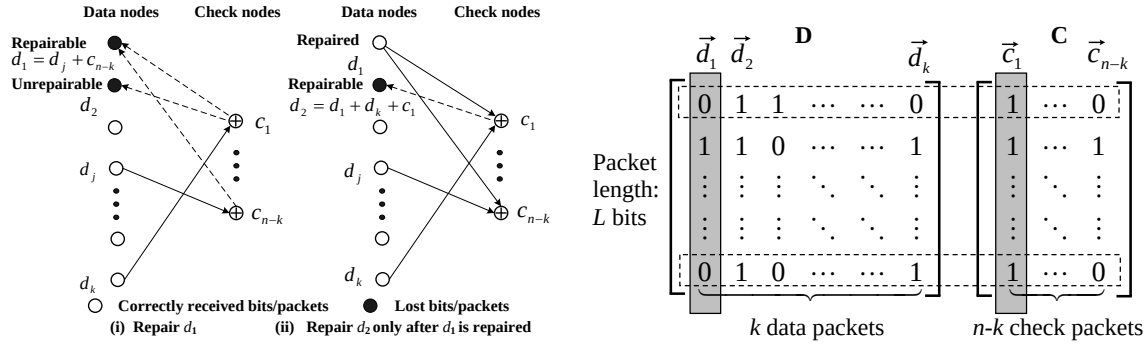


Fig. 31. (a) Iterative decoding for graph codes. (b) Employing graph codes in packet level, where **D** forms a transmission group (TG).

automatically adjust the error-control redundancy level according to different QoS requirements.

This chapter is organized as follows. Section B introduces the low-complexity graph-codes used for erasure-error corrections. Section C describes the system model for the QoS-driven Section D the QoS-driven Section D proposes the two-dimensional (2-D) adaptive mobile multicast error-control scheme and presents its analytical and numerical analyses. Section E evaluates the performance of our proposed schemes through simulations. The chapter concludes with Section F.

B. Low-Complexity Erasure Graph Codes

The principle and structure of the graph code [87] can be described by a bipartite graph shown in Fig. 31(a). A bipartite graph consists of two disjoint classes of nodes. Two nodes in different classes can be connected by an edge, but there are no edges connecting any two nodes within the same class. The number of edges connected to a node is called *degree* of that node. In a bipartite graph, each node on the left-hand side, representing a data bit, is called a data node. Each node on the right-hand

side, representing a parity-check bit, is called a check node. Consider a graph code of length n with k data nodes and $(n-k)$ check nodes in a bipartite graph. Let d_i denote the ith data bit and c_j denote the jth check bit. We call the edge-connection pattern between data nodes and check nodes in the bipartite graph the *mapping structure* of the graph code, each of which determines a specific graph code structure.

As shown in Fig. 31(a), each check bit is calculated as the sum in GF(2) (Galois Fields) of all the data bits connected to it. Graph codes can *iteratively* correct/repair the erasure errors by decoding through simple modulo-2 additions [87] (we use "+" to represent modulo-2 additions in all encoding/decoding operations throughout this chapter) as follows.

Step 1: Search for the check bits which are connected with *only one* lost data bit.

Step 2: Recover corresponding lost data bits according to this code mapping structure.

Step 3: Go back to Step 1 until all the lost data bits are repaired or no more can be repaired.

Fig. 31(a) shows an example of this procedure. First, assume d_1 and d_2 are the only lost bits as only lost bits as shown in Fig. 31(a)–(i). Thus, repaired by $d_1 = d_j + c_{n-k}$. Following this, d_2 can be iteratively repaired by $d_2 = d_1 + d_k + c_1$ as shown in Fig. 31(a)–(ii). Clearly, it is possible some lost Fig. 31(a)–(ii). Clearly, it is possible some lost data bits still cannot be repaired even after the iterative decoding procedure ends, depending on the code's mapping structure used and which/how many data bits are lost.

The graph code mapping structures can be algebraically expressed by the code-structure matrix $\mathbf{P} = (p_{ij})_{k\times(n-k)}$ with $p_{ij} \in \{0, 1\}$, where p_{ij} equals 1 (0) if the ith data bit is (not) connected to the jth check bit in the bipartite graph. Then, we can obtain the $(n-k)$-bit long check-bit vector $\mathbf{c}$ by the simple encoding procedure as

follows:

$$\mathbf{c} \triangleq [c_1\, c_2\, \ldots\, c_{n-k}] = \mathbf{d}\mathbf{P}_{k\times(n-k)} \tag{6.1}$$

in GF(2) from the k-bit long data-bit vector $\mathbf{d} \triangleq [d_1\, d_2\, \ldots\, d_k]$. Considering the systematic graph codes, the generator matrix of graph codes can be expressed as $\mathbf{G}_{k\times n} = [\mathbf{I}_{k\times k}\ \ \mathbf{P}_{k\times(n-k)}]$, and an n-bit long code word can be generated by $\mathbf{w} = \mathbf{d}\mathbf{G}_{k\times n}$. Then, the degrees of the ith data node, denoted by α_i, and jth check node, denoted by γ_j, are equal to the number of 1's in the ith row and jth column of $\mathbf{P}$, respectively. We also call α_i and γ_j the weights of the ith row and jth column, respectively. Generally, in order to increase the probability of successful decoding/repairing and reduce the computational complexity, α_i and γ_j are designed to be much smaller than k. This implies that a *sparse* $\mathbf{P}$ is usually required.

The most important advantage of the graph-code-based error-control schemes [87, 88] is that the encoding/decoding time-complexity is much lower as compared to the RSE-code-based schemes. Consequently, the graph-code-based error-control scheme has been applied into the asynchronous reliable multicast transmission [68] to achieve high efficiency while keeping the error-control complexity low. In addition, the decoding procedures for graph codes can be iteratively performed with any number of check packets correctly received instead of having to wait until at least k distinct packets (including both data and check packets) are correctly received, like in the decoding of RSE codes. This can help save a significant amount of bandwidth for QoS-driven mobile multicast services. Moreover, graph-code-based schemes enable the code structures to be adaptive for improving the error-control efficiency.

To extend graph codes to the packet level in implementing hybrid ARQ-FEC-based multicast services over wireless networks, we divide the source data packet stream into blocks each consisting of k consecutive data packets, which form *trans-*

mission groups (TG) [see Fig. 31(b)]. The number k is called TG size. Assuming the packet length is L bits, we denote a data packet by an $L \times 1$ column vector $\vec{d_i}$ where $i = 1, 2, \ldots, k$, as shown in the solid-lined box on the left-hand side in left-hand side in Fig. 31(b). Let k data packets form $\mathbf{D}_{L\times k}$, as shown in Fig. 31(b), where the jth column comes from the jth data packet and the ith row consists of the ith bit of all k data packets. Then, the encoding procedure given by Eq. (6.1) can be used to generate a $1 \times (n-k)$ check-bit vector in the ith row of the check matrix $\mathbf{C}$. The data bits in a row and corresponding check bits forms a *code word* as shown in a dash-lined box in Fig. 31(b). All the jth check bits in each row of $\mathbf{C}$ form the jth check packet with L bits long, denoted as $\vec{c_j}$, where $j = 1, 2, \ldots, (n-k)$, as shown in the solid-lined box on the right-hand side in Fig. 31(b). Fig. 31(b). The above encoding procedure at the packet level can be algebraically expressed in GF(2) by

$$\mathbf{C}_{L\times(n-k)} = \mathbf{D}_{L\times k}\mathbf{P}_{k\times(n-k)}, \tag{6.2}$$

which is virtually the same as the encoding procedure given by Eq. (6.1) at the bit level.

C. System Model of Hybrid ARQ–FEC-Based Mobile Multicast

1. The Hybrid ARQ–FEC-Based Mobile-Multicast Transmission Model

We model the mobile multicast transmission system by a multicast tree, which consists of one sender and a number of mobile multicast receivers. The sender multicasts a stream of data packets to each receiver with the required packet-loss-rate QoS, denoted by ξ [see Eq. (6.3)]. We assume that the packet losses are independent and identically distributed (i.i.d.) in terms of time (for different packets) and space (for different receivers). The assumption of i.i.d. loss for different packets is particularly

suitable for wireless networks, where the random loss often happens, unlike the wired networks, where the data loss usually occurs in the bursty fashion due to the congestion in bottlenecks. It should be also noted that FEC codes usually have much higher erasure-correcting capability for random loss than for bursty loss. The integrated ARQ-FEC error-control scheme is implemented through closed-loop information exchanges by using forward and feedback control packets between the sender and the receivers in the mobile multicast tree. Also, we assume that all control information can be reliably transmitted. The error-control information is exchanged in each *transmission round* (TR), which is defined as follows. To implement the adaptive error-control, a TG of data packets are usually transmitted through a number of TR's. Each TR begins with the sender multicasting k data packets (i.e., data-packet TR or retransmission round) or a certain number of check packets (i.e., check-packet TR) and ends with the sender having received consolidated feedbacks from all multicast receivers. So, TR is also the basic control period of adaptation, where TR is indexed by $t = 1, 2, \ldots$.

The packet stream from the data source is divided into a number of TG's each with k data packets. For each TG, the sender multicasts the k data packets in the first TR. Then, the sender waits until all feedback packets arrive, which carry the error-control information from the mobile receivers. Based on the feedback error-control information (e.g., the packet-loss level, to be detailed later), the sender determines to transmit either a new next TG or a number of parity-check packets to repair losses for the current TG. Specifically, unless the reliability-QoS [to be detailed later in Eq. (6.3) and Section C-2] is satisfied by all receivers, the sender must generate a number of check packets from the k data packets of the current TG and then multicast them to all mobile receivers for loss repairing. This loss-repairing procedure repeats until the reliability-QoS requirement is satisfied by all mobile receivers. However, if

the reliability-QoS fails to be satisfied after all the available check packets have been generated and transmitted, the retransmission of the current TG must be executed by the sender. In addition, we assume that the control information such as the packet sequence number and the packet-loss transmitted between the sender and receivers. To achieve excellent performance, several parameters need to be selected carefully. A set of parameter selection algorithms are presented in Section D.

2. Different QoS Requirements for Mobile Multicast Services

While there are a wide range of QoS metrics, we mainly focus on the QoS metrics closely associated with the error-control for mobile multicast, which include the reliability and transmission delays. To efficiently use the limited resources in mobile wireless networks while supporting QoS requirements, the error-control parameters need to be adjusted dynamically according to the different QoS requirements for different mobile multicast services. In particular, the real-time (e.g., video/audio) mobile multicast services must upper bound the transmission delay, but can tolerate a certain packet losses, implying that a relatively higher packet-loss rate is allowed than that for reliable services. Furthermore, this required loss-rate QoS threshold can be increased (or decreased) as the required quality of the received audio/video streams decreases (or increases). On the other hand, the data mobile multicast services must have zero loss while tolerating a certain transmission delay. As a result, the various QoS requirements of interest in this chapter can be characterized by the reliability QoS. We define the required reliability-QoS by packet-loss rate, denoted by ξ. To complete the transmission of a TG with the required packet-loss rate QoS ξ in the t-th TR, the following condition must be satisfied by all receivers:

$$\frac{f_r(t)}{k} \leq \xi, \quad \forall\ 1 \leq r \leq R, \tag{6.3}$$

where $f_r(t)$ is the number of lost/unrepaired data packets of a TG for the rth receiver after completing decoding procedures in the t-th ($t = 1, 2, \ldots$) TR. Note that reliability-QoS is not the only QoS measure in this chapter. On the condition that reliability-QoS requirements must be satisfied, we also consider other QoS metrics such as the average delay and so on, which are defined in Section C-4. For our proposed error-control scheme, once the condition given in Eq. (6.3) is Eq. (6.3) is satisfied for a certain number t of loss repairing TR's, the sender stops sending check/repairing packets for the current TG and then immediately starts transmitting the next new TG. As a result, a significant amount of bandwidth can be saved for graph-code-based error-control schemes where the decoding procedure can proceed iteratively and cumulatively with *any* number of correctly received check packets. By contrast, RSE-code-based schemes do not have this advantage because the decoding procedure cannot start for a mobile multicast receiver until at least k distinct data/check packets have been correctly received at this receiver. Note that throughout this chapter, we use two similar terms which have different meanings, namely, 1) *packet-loss rate*, denoted by ξ, represents the required reliability-QoS; (2) *packet-loss probability* p, denoted by p, represents the channel quality.

3. The Cost-Effective Feedback Signaling Algorithms

To solve the feedback explosion and synchronous problems, we propose to use the Soft-Synchronous Protocol developed by [32–34] (SSP) in this adaptive protocol for mobile multicast services, which consolidates the numbers $f_r(t)$, $r \in \{1, 2, \ldots, R\}$, of lost data packets for the rth receiver in the t-th TR by selecting/feeding back the maximum number $\theta_{\max}(t)$ of lost packets among all receivers as:

$$\theta_{\max}(t) \triangleq \max_{r \in \{1,2,\ldots,R\}} \{f_r(t)\}. \tag{6.4}$$

in the t-th TR with $t = 1, 2, \ldots$. Note that the feedback consolidation procedure given by Eq. (6.4) is just the general procedure, which in fact is iteratively implemented at each branch-node within that multicast sub-trees. Thus, Eq. (6.3) can be equivalently rewritten as

$$\frac{\theta_{\max}(t)}{k} = \frac{\max\limits_{r\in\{1,2,\ldots,R\}}\{f_r(t)\}}{k} \leq \xi, \quad \forall\ 1 \leq r \leq R. \tag{6.5}$$

By using SSP, the packet-sequence-independent error-control schemes can be efficiently applied. The feedbacks only contain information of the *number* of lost packets rather than a series of the *sequence numbers* of lost packets during each TR. Consequently, the feedback bandwidth overhead is significantly reduced. Note that by using SSP, the sender adjusts error-control parameters for each next TR only based on the worst packet-loss level among all receivers. For the detailed SSP, see [32–34].

4. Performance Metrics

For the FEC based error-control protocols/schemes used in mobile multicast, we use the following metrics to evaluate their performance.

4.1. Bandwidth Efficiency η:

To complete the transmission for a TG with k data packets, the sender usually needs to transmit a *random* number M ($M \geq k$) of packets until Eq. (6.5) is satisfied. We define the bandwidth efficiency η by

$$\eta \triangleq \frac{k}{E\{M\}}, \tag{6.6}$$

where $E\{M\}$ is the expectation of M. Clearly, we have $0 \leq \eta \leq 1$. The high bandwidth efficiency implies the high error-control efficiency for the error-control schemes. Since the RSE code has almost the highest error-control efficiency/capability for erasure channels (the RSE code is a type of maximum-distance separable (MDS)

code [88]), the performance of a new FEC-based protocol (not RSE-code-based) can be evaluated by comparing its η with that of the RSE code in terms of following criteria: 1) For reliable services, η should be close to η_{RS}, which is the bandwidth efficiency for RSE-code-based error-control schemes; 2) η dose not decrease quickly when packets loss probability increases; 3) η does not decrease quickly when the number of receivers increases and thus the protocol has good scalability.

4.2. Average number $E\{Q\}$ of TR's to Reach the Reliability-QoS Requirement ξ:

We denote the number of TR's to complete the transmission of a TG and its expectation by Q and $E\{Q\}$, respectively. Clearly, to obtain the feedbacks in each TR, the sender needs to wait at least a round-trip-time (RTT), which is the major contributor to the delay. Thus, a multicast protocol needs to keep a low $E\{Q\}$ to achieve the low delay. Also, a low $E\{Q\}$ represents low overhead introduced to multicast services.

4.3. Average delay QoS to Reach the Reliability-QoS Requirement ξ:

The average delay, denoted by τ, to complete the transmission of a TG between the sender and the receivers is expressed by using Eq. (6.6) as

$$\tau = \frac{LE\{M\}}{B} + (RTT)E\{Q\} = \frac{kL}{\eta B} + (RTT)E\{Q\}, \tag{6.7}$$

where L is the packet length (we assume fixed packet length throughout the chapter), B is the bottleneck bandwidth among all receivers, and RTT is the maximum end-to-end round trip time among all the sender-receiver pairs. From Eq. (6.7), our error-control scheme has two factors affecting the delay QoS. One is bandwidth efficiency η and the other is total average number $E\{Q\}$ of TR's. Either increasing η or decreasing $E\{Q\}$ will improve the delay QoS. However, increasing η may lead to a higher $E\{Q\}$. Thus, this introduces a tradeoff between η and $E\{Q\}$.

D. The Two-dimensional Adaptive Error-Control Design Based on Graph Codes

Unlike RSE based FEC multicast error control, where the sender only dynamically adjusts the code redundancy according to the packet-loss levels while the coding scheme (RSE codes) stays the same, to further improve error-control efficiency and support the QoS diversity, we propose the two-dimensional graph-code-based multicast error-control schemes that regulate not only the code redundancy, but also the code structures, dynamically, based on different packet-loss levels fed back from multicast mobile receivers. This is motivated by our analyses of the graph-code-based schemes, which indicate that besides adapting error-control redundancy in each TR, the loss-repairing efficiency can also be significantly improved by using the *nonuniformed* code mapping structures corresponding to different packet-loss levels. The key components and principles of our proposed two-dimensional adaptive graph-code-based scheme for providing the QoS-driven mobile multicast services are detailed below in terms of *code mapping-structure adaption* and *error-control redundancy adaption*, respectively.

In particular, for the transmission of each TG, the matrix $\mathbf{P}$ characterizing the graph code (see Section B) is composed of $(Q-1)$ sub-matrices denoted by $\mathbf{P}_1, \mathbf{P}_2, \ldots, \mathbf{P}_{Q-1}$, where $\mathbf{P} = [\mathbf{P}_1\ \mathbf{P}_2\ \cdots\ \mathbf{P}_{Q-1}]$. The sub-matrix $\mathbf{P}_{t-1}$ represents the mapping structure for the check packets generated in the t-th TR (the 1st TR is the data transmission round). In the t-th TR, $t \geq 2$, the sender dynamically generates $\mathbf{P}_{t-1}$ for loss-repairing according to the packet-loss level $\theta_{\max}(t)$. Also, the error-control redundancy in the t-th TR (the number of check packets or, equivalently, the number of columns of $\mathbf{P}_{t-1}$) is dynamically determined according to $\theta_{\max}(t)$. How to determine the mapping structure of $\mathbf{P}_{t-1}$ and the error-control redundancy in each TR will be elaborated on in Sections D-1 and D-2, respectively.

1. Code Mapping Structure Adaptation

The construction of the graph code's mapping structure for one check packet (one column of $\mathbf{P}_{t-1}$) includes two parts. One is the selection of check-node degree (the numbers of 1's in each column of $\mathbf{P}_{t-1}$), denoted by γ. The other is the selection of which γ data packets are connected to the check packet (edge-connection pattern) in the bipartite graph. Consider one single receiver. We denote the packet-loss level by θ. Because losses are i.i.d. for different packets, then given the packet-loss level θ, the probabilities of the occurrence for each *loss pattern* (loss pattern refers to which θ data packets are lost) are equal. Consequently, the probability of repairing one lost data packet by one single check packet does not depends on edge-connection pattern, but only on the check-node degree γ. Thus, we select the check-node degree and edge-connection pattern separately.

In this chapter, we propose to use the *random mapping structure* for the construction of each check packet. In particular, for each check packet, we randomly choose γ distinct data packets and then connect them with this check packet in the bipartite graph. Note that each data packet is equally likely to be chosen. In addition, because a TR is the adaptation cycle, we let all check packets in a TR have the same check-node degree. Also, we assume that the random selections of edge-connection pattern for different check packets are independent. The random mapping structure described above has the following characteristics. First, it is easy for implementation. Second, the supported maximum error-control redundancy is virtually not upper-bounded. Third, by using the same random-number generating algorithm and setting the same initial random-number seed, both the sender and all receivers can construct the exactly same mapping structure in each TR based on the same control information, e.g., the packet-loss level. Thus, the sender needs to transmit

Table III. Parameters and Metrics to Evaluate the Repairing Efficiency

k	Size of a transmission group.
θ	The number of lost/unrepaired data packets out of k data packets, also called packet-loss level.
γ	Check-node degree.
ℓ	The number of data packets which are successfully repaired by m received check packets. $0 \leq \ell \leq \min\{m, \theta\}$.
$\psi_m(k, \theta, \gamma, \ell)$	Given k, θ and γ, the probability that total ℓ data packets are successfully repaired by m ($m \geq 1$) received check packets.
$N_m(k, \theta, \gamma)$	Given k, θ and γ, the average number of successfully repaired data packets by m ($m \geq 1$) received check packets, which is defined to characterize the loss-repairing efficiency.
$\gamma_m^*(k, \theta)$	Given the number of received check packets m and packet-loss level θ, the optimal check-node degree maximizing the loss-repairing efficiency.

only a small amount of control information instead of the entire mapping structure to all receivers.

Next, we discuss how to select the check-node degree in each TR to achieve high error-control efficiency. Note that in this section, the derived parameter selection algorithms are based on the single receiver case. However, these algorithms are also efficient for multiple receiver cases. Because the consolidated $\theta_{\max}(t)$ represents the highest packet-loss level among all receivers', thus the derived algorithms actually aim at efficiently improving the error-control efficiency for the receiver with the worst-case losses.

For the given check-node degree γ, packet-loss level θ, and m correctly received check packets, we derive the *average number* $N_m(k, \theta, \gamma)$ of successfully repaired data packets to characterize the loss-repairing efficiency, which is expressed as

$$N_m(k, \theta, \gamma) = \sum_{\ell=1}^{\min\{\theta, m\}} \ell \, \psi_m(k, \theta, \gamma, \ell), \tag{6.8}$$

where $\psi_m(k, \theta, \gamma, \ell)$ is the probability that total ℓ data packets are successfully repaired by m ($m \geq 1$) received check packets under given k, θ and γ. All the related parameters are defined in Table III. Also, we define the *optimal check-node degree* by

$$\gamma_m^*(k, \theta) \triangleq \arg \max_{1 \leq \gamma \leq k} N_m(k, \theta, \gamma), \tag{6.9}$$

which maximizes the average number of successfully repaired data packets.

a. Single check packet case $m = 1$

We first discuss the single check packet case ($m = 1$). Note that with a single check packet ($m = 1$), at most one lost data packet can be repaired. Thus, loss-repairing efficiency becomes $N_1(k, \theta, \gamma)$ which actually equals the loss-repairing probability $\psi_1(k, \theta, \gamma, 1)$. Theorem 9 introduced below derives the equations and criteria to determine the optimal check-node degree γ^* for any given code size k and the number θ of lost packets with the *single* ($m = 1$) check packet.

Theorem 9. *If a graph code has k data packets in which θ data packets are lost randomly with* i.i.d. *distributions, then the following claims hold for $k \geq 1$ and $\theta = 1, 2, \ldots, k$.*

Claim 1. The probability, denoted by $\psi_1(k, \theta, \gamma, 1)$, that one ($\ell = 1$) lost packet can be repaired by one ($m = 1$) received parity-check packet with check-node degree γ is determined by

$$\psi_1(k, \theta, \gamma, 1) = N_1(k, \theta, \gamma) = \begin{cases} \frac{\theta\gamma(k-\theta)\,!(k-\gamma)!}{(k-\gamma-\theta+1)\,!k\,!}, & \text{if } \gamma \leq k - \theta + 1; \\ 0, & \text{if } \gamma > k - \theta + 1. \end{cases} \tag{6.10}$$

Claim 2. For any given (k, θ) satisfying $k \geq 1$ and $1 \leq \theta \leq k$, there exists the maximum for $N_1(k, \theta, \gamma)$ as the function of γ, and the maximizer $\gamma_1^*(k, \theta)$ is determined by

$$\begin{aligned} \gamma_1^*(k, \theta) &\triangleq \arg \max_{1 \leq \gamma \leq k} N_1(k, \theta, \gamma) \\ &= \arg \max_{1 \leq \gamma \leq k} \psi_1(k, \theta, \gamma, \ell)\Big|_{\ell=1} \\ &= \left\lceil \frac{(k+1) - \theta}{\theta} \right\rceil, \end{aligned} \tag{6.11}$$

where $\lceil w \rceil$ denotes the least integer number that is larger than or equal to w.

Claim 3. The dynamics of $\psi_1(k,\theta,\gamma,1)$ is symmetric with respect to θ and γ such that $\psi_1(k,\theta,\gamma,1) = \psi_1(k,\gamma,\theta,1)$, and if and only if $(\theta = 1, \gamma_1^*(k,1) = k)$ or $(\theta = k, \gamma_1^*(k,k) = 1)$, $\psi_1(k,\theta,\gamma,1)$ attains its least upper bound $\psi_1^*\big(k,\theta,\gamma_1^*(k,\theta),1\big)$ determined by

$$\begin{aligned}\psi_1^*\big(k,\theta,\gamma_1^*(k,\theta),1\big) &= \sup_{\substack{1\le\theta\le k\\ 1\le\gamma\le k}} \{\psi_1(k,\theta,\gamma,1)\}\\ &= \psi_1\big(k,\theta,\gamma_1^*(k,\theta),1\big)\big|_{\left(\theta=1,\gamma_1^*(k,1)=k\right)\ or\ \left(\theta=k,\gamma_1^*(k,k)=1\right)}\\ &= 1. \end{aligned} \tag{6.12}$$

Proof. The detailed proof is provided in Appendix O. □

Remarks on Theorem 9. Claim 1 derives general expressions for loss-repairing probability/efficiency with a single check packet. Claim 2 states the existence and gives the closed-form expression of $\gamma_1^*(k,\theta)$. For any given (k,θ), a γ either much larger, or much smaller, than $\gamma_1^*(k,\theta)$ is undesired. This is expected since a γ much larger than $\gamma_1^*(k,\theta)$ can increase the cases of having *two* or *more than two* edges of the same check packet to be connected to the lost data packets, while a γ much smaller than $\gamma_1^*(k,\theta)$ can yield more cases where all edges of the check packet are only connected to the correctly received data packets. Equation (6.11) makes the critical observation packets. Equation (6.11) makes the critical observation decreasing function of the number θ of lost data packets. More importantly, Eq. (6.11) provides the network designers with a closed-form analytical expression to calculate the optimal value $\gamma_1^*(k,\theta)$ of check-node degree according to feedback of packet-loss level θ for any given graph code block size k. Claim 3 implies that variables θ and γ are function-

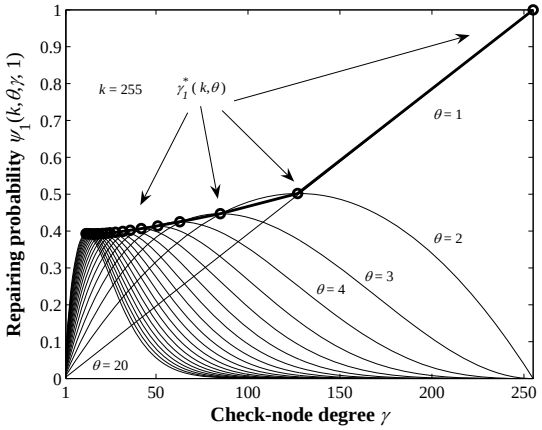


Fig. 32. Repairing probability $\psi_1(k, \theta, \gamma, 1)$ versus check-node degree γ. $\theta = 1, 2, \ldots, 20$ and $k = 255$.

ally equivalent or exchangeable. In addition, this claim derives the conditions when $\psi_1(k, \theta, \gamma, 1)$ attains its globally absolute maximum. When $\theta = 1$, i.e., at most one data packet is lost for any multicast receivers, the optimal check-node degree satisfies $\gamma_1^*(k, \theta) = k$ based on Claim 2. Thus, the check packet actually is the modulo-2 addition of all the data packets (in this case, the code reduces to the well-known single parity check code [89, Chapter 3.8.1]) and its loss-repairing probability attains its upper bound 1 according to Claim 3. It is clear that this mapping structure can repair the lost packet for any loss pattern with $\theta = 1$. Since this case corresponds to the possible last mapping structure to be selected for $\theta = 1$ immediately before all lost packets are repaired, we call this mapping structure the *final protocol*, which has the highest loss-repairing efficiency with a single check packet. Under this condition, the multicast system reaches a special state, where the sender only needs to keep on transmitting the check packet generated by the *final protocol* until all the lost data packets have been repaired. On the other hand, if $\theta = k$ (all data packets are lost), $\gamma_1^*(k, k) = 1$ should be selected to guarantee repairing one lost packet. Thus, the protocol effectively reduces to the retransmission protocol.

Fig. 32 numerically plots the loss-repairing probability $\psi_1(k, \theta, \gamma, 1)$ against check-

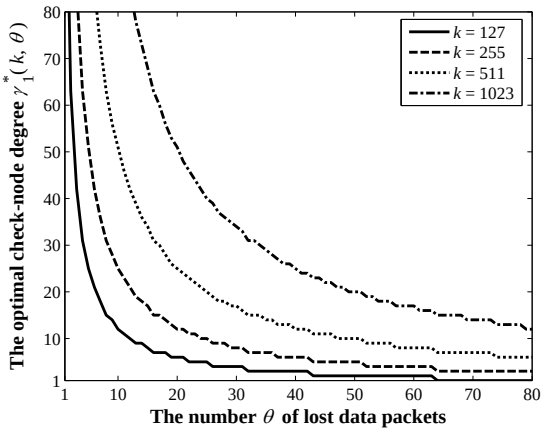


Fig. 33. Check-node degree $\gamma_1^*(k,\theta)$ versus the number θ of lost-data packets. $k = 127, 255, 511, 1023$.

node degree check-node degree γ. We can see from Fig. 32 packet-loss level θ, there is the optimal $\gamma_1^*(k,\theta)$ which maximizes $\psi_1(k,\theta,\gamma,1)$, as marked with a circle in Fig. 32, verifying the as marked with a circle in Fig. 32, verifying the Eq. (6.11), Fig. 33 plots the optimal check-node degree $\gamma_1^*(k,\theta)$ against the packet losses θ with different code-block sizes $k = 127, 255, 511, 1023$, which show $\gamma_1^*(k,\theta)$ is a decreasing function of θ. So, we should select small check-node degree if packet-loss level is high and vice versa. Also, we observe that the smaller θ is, the faster $\gamma_1^*(k,\theta)$ increases as θ decreases. All the above observations suggest that the *nonuniformed* code structures should be used to achieve the high error-control efficiency. In addition, for any given θ, Fig. 33 shows that the larger the block-size k, the higher the optimal check-node degree $\gamma_1^*(k,\theta)$. This is also expected since a large k implies that we need to have more repairing edges from the check nodes connected to the data packets to cover the lost data packets, and vice versa.

b. Multiple check packet case ($m > 1$)

In realistic systems, we usually need to send multiple check packets in each TR rather than only a single check packet. However, the derivations of $N_m(k,\theta,\gamma)$ and $\gamma_m^*(k,\theta)$

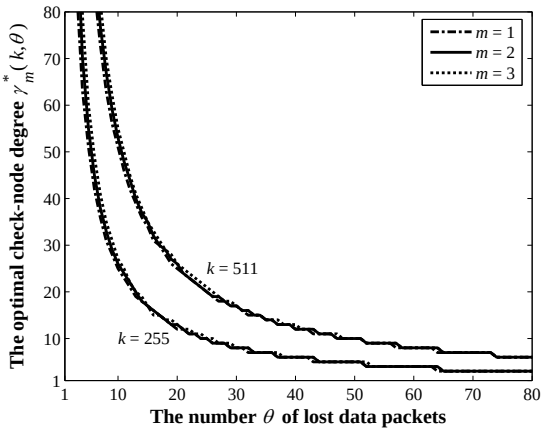


Fig. 34. Check-node degree $\gamma_m^*(k,\theta)$ versus packet-loss level θ. $m = 1, 2, 3$ and $k = 255, 511$.

become much more complicated as m increases. For simplicity, we still select the check-node degree as $\gamma = \gamma_1^*(k,\theta)$ even for $m \geq 2$. To investigate the impact of m on the selection of $\gamma_m^*(k,\theta)$, we derive the analytical expressions of $\psi_m(k,\theta,\gamma,\ell)$'s for $m = 2, 3$, which are summarized by Eqs. (6.13) through (6.17). Correspondingly, $N_m(k,\theta,\gamma)$ can be derived by using Eq. (6.8) and $\psi_m(k,\theta,\gamma,\ell)$ given in Eqs. (6.13)-(6.17). Then, we can obtain $\gamma_m^*(k,\theta)$ through Eq. (6.9).

Fig. 34 plots the numerical results of $\gamma_m^*(k,\theta)$ against θ for $m = 1, 2, 3$. From $m = 1, 2, 3$. From Fig. 34, we observe that the very close to each other for all θ. Thus, it can be expected that setting the check-node degree equal to $\gamma_1^*(k,\theta)$ will not cause significant performance loss for multiple check packet cases. Based on the above discussions, we only use Eq. (6.11) to select the check-node degree in our proposed adaptive protocol.

$$\psi_2(k,\theta,\gamma,1)$$
$$= \begin{cases} \frac{\theta\binom{k-\theta}{\gamma-1}}{\binom{k}{\gamma}^2}\left[2\binom{k}{\gamma}+(1-2\theta)\binom{k-\theta}{\gamma-1}-2(\theta-1)\binom{k-\theta}{\gamma-2}\right], & \text{if } 2\le\gamma\le k-\theta+1, \theta\ge 1; \\ \frac{(1+2k)\theta-2\theta^2}{k^2}, & \text{if } \gamma=1, \theta\ge 1; \\ 0, & \text{otherwise,} \end{cases} \tag{6.13}$$

$$\psi_2(k,\theta,\gamma,2)$$
$$= \begin{cases} \frac{2(k-\theta+\gamma)\binom{\theta}{2}\binom{k-\theta}{\gamma-1}\binom{k-\theta+1}{\gamma-1}}{(k-\theta+1)\binom{k}{\gamma}^2}, & \text{if } 2\le\gamma\le k-\theta+1, \theta\ge 2; \\ \frac{2}{k^2}\binom{\theta}{2}, & \text{if } \gamma=1, \theta\ge 2; \\ 0, & \text{otherwise,} \end{cases} \tag{6.14}$$

$$\psi_3(k,\theta,\gamma,1)$$
$$= \begin{cases} \frac{\theta\binom{k-\theta}{\gamma-1}}{\binom{k}{\gamma}^3}\left\{3\left[\binom{k}{\gamma}-\theta\binom{k-\theta}{\gamma-1}-(\theta-1)\binom{k-\theta}{\gamma-2}\right]\cdot\left[\binom{k}{\gamma}-(\theta-1)\binom{k-\theta+1}{\gamma-1}\right]+\binom{k-\theta}{\gamma-1}^2\right\}, & \\ & \text{if } 2\le\gamma\le k-\theta+1, \theta\ge 1; \\ \frac{\theta}{k^3}\Big[3(k-\theta)^2+3(k-\theta)+1\Big], & \text{if } \gamma=1, \theta\ge 1; \\ 0, & \text{otherwise,} \end{cases} \tag{6.15}$$

$$\psi_3(k,\theta,\gamma,2)$$
$$= \begin{cases} \frac{6\binom{\theta}{2}\binom{k-\theta}{\gamma-1}}{\binom{k}{\gamma}^3}\Bigg\{\binom{k-\theta+1}{\gamma-1}\left[\frac{k-\theta+\gamma}{k-\theta+1}\left(\binom{k}{\gamma}-\theta\binom{k-\theta}{\gamma-1}-(2\theta-3)\binom{k-\theta}{\gamma-2}-(\theta-2)\binom{k-\theta}{\gamma-3}\right)\right. \\ \qquad \left.+(2-\frac{\gamma-1}{k-\theta+1})\binom{k-\theta}{\gamma-2}\right]+\binom{k-\theta}{\gamma-1}^2\Bigg\}, & \text{if } 3\le\gamma\le k-\theta+1,\theta\ge 2; \\ \frac{6\binom{\theta}{2}}{\binom{k}{2}^3}\Big\{(1-\theta)(k-\theta)^3+\Big[\binom{k}{2}-4\theta+5\Big](k-\theta)^2+\Big[2\binom{k}{2}-4\theta+7\Big](k-\theta)\Big\}, & \\ & \text{if } 2=\gamma\le k-\theta+1,\theta\ge 2; \\ \frac{6}{k^3}(k-\theta+1)\binom{\theta}{2}, & \text{if } \gamma=1,\theta\ge 2; \\ 0, & \text{otherwise,} \end{cases} \tag{6.16}$$

$$\psi_3(k,\theta,\gamma,3)$$
$$= \begin{cases} \frac{6\binom{\theta}{3}\binom{k-\theta}{\gamma-1}}{\binom{k}{\gamma}^3}\left[\binom{k-\theta}{\gamma-1}^2+6\binom{k-\theta}{\gamma-1}\binom{k-\theta}{\gamma-2}+3\binom{k-\theta}{\gamma-1}\binom{k-\theta}{\gamma-3}+9\binom{k-\theta}{\gamma-2}^2+6\binom{k-\theta}{\gamma-2}\binom{k-\theta}{\gamma-3}\right], & \\ & \text{if } 3\le\gamma\le k-\theta+1,\theta\ge 3; \\ \frac{6\binom{\theta}{3}}{\binom{k}{2}^3}\Big[(k-\theta)^3+6(k-\theta)^2+9(k-\theta)\Big], & \text{if } 2=\gamma\le k-\theta+1,\theta\ge 3; \\ \frac{6}{k^3}\binom{\theta}{3}, & \text{if } \gamma=1,\theta\ge 3; \\ 0, & \text{otherwise.} \end{cases} \tag{6.17}$$

2. Error-Control Redundancy Adaptation

After the check-node degree is selected in each TR, we need to determine an appropriate error-control redundancy (the number of check packets constructed and transmitted) in each TR based on the current packet-loss level θ. We denote the error-control redundancy in a TR by T. Consider the case where we select a very large T for the current TR. During the iterative decoding/repairing procedures in the current TR, the packet-loss level θ decreases gradually such that the selected $\gamma_1^*(k,\theta)$ cannot achieve near optimal loss-repairing probability with the changed θ. That is,

if T is too large, the loss-repairing efficiencies of a majority of check packets received in the corresponding TR drop with the gradually decreasing packet-loss level. Consequently, more check packets are required because of the low loss-repairing efficiency, which severely degrades the bandwidth efficiency. If T is too small, although we can avoid the problems mentioned above, the improvement of the bandwidth efficiency is achieved at the cost of a higher Q, which may lead a long delay. Thus, we need to select a balanced T in each TR.

We develop a loss-covering strategy to determine T. For a given graph code, if a data node/packet is connected to one or more check nodes/packets, we say that this data node/packet is *covered*. In order for a lost data packet to be repaired, it must be covered. Under this principle, we develop the following *covering criterion* to obtain a balanced T with the given TG size k, check-node degree γ, and packet-loss level θ.

Covering Criterion: Using the random mapping structure, we let T in a TR equal the average number $T(k, \theta, \gamma)$ of check packets required to cover *at least one* lost data packet or, equivalently, to cover *at least* $(k - \theta + 1)$ data packets.

Clearly, under the above covering criterion, the error-control redundancy $T(k, \theta, \gamma)$ is affected by both γ and θ. The following Theorem 10 derives the closed-form solution to $T(k, \theta, \gamma)$ for the above developed covering criterion.

Theorem 10. *Using the* random mapping structure*, if the TG size is equal to k, the check-node degree is equal to γ, $1 \leq \gamma \leq k$, and the packet-loss level is equal to θ, $1 \leq \theta \leq k$, then the average number $T(k, \theta, \gamma)$ of check packets required to cover* at least one *lost packet or, equivalently,* at least $(k - \theta + 1)$ *data packets is given by,*

$$T(k, \theta, \gamma) = \begin{cases} 1, & \text{if } \gamma \geq k - \theta + 1; \\ h_0, & \text{if } \gamma < k - \theta + 1, \end{cases} \tag{6.18}$$

where h_0 is determined by the following iterative equations:

$$\begin{cases} h_i = \dfrac{1}{1-\rho_{i,i}} \left(1 + \displaystyle\sum_{j=i+1}^{k-\theta+1} \rho_{i,j} h_j \right), \text{ if } 0 \le i \le k-\theta; \\ h_{k-\theta+1} = 0, \end{cases} \tag{6.19}$$

and $\rho_{i,j}$, for $0 \le i, j \le k-\theta+1$, is given by

$$\rho_{i,j} = \begin{cases} \binom{i}{\gamma-j+i}\binom{k-i}{j-i} / \binom{k}{\gamma}, & \text{if } 0 \le j-i \le \gamma \le j \text{ and } j < k-\theta+1; \\ \displaystyle\sum_{v=k-\theta+1}^{\min\{i+\gamma,k\}} \binom{i}{\gamma-v+i}\binom{k-i}{v-i} / \binom{k}{\gamma}, & \text{if } j = k-\theta+1 \text{ and } i+\gamma \ge j; \\ 0, & \text{otherwise.} \end{cases} \tag{6.20}$$

Proof. This Theorem is proved by using the Markov Chain model as described in Appendix P. □

Note that $\binom{u}{v} = u!/((u-v)!v!)$ for nonnegative integers u and v, $u \ge v \ge 0$. Also, $T(k,\theta,\gamma)$ may not be an integer. Then, we let $T = \lceil T(k,\theta,\gamma) \rceil$ to determine the error-control redundancy in each TR. Fig. 35 numerically plots the error-control redundancy $T(k,\theta,\gamma_1^*(k,\theta))$ in a TR against the packet-loss level θ. Through Fig. 35, we level θ. Through Fig. 35, we have the $T(k,\theta,\gamma_1^*(k,\theta))$ increases (decreases) with the increasing of packet-loss level θ when θ is relatively small (large). This is because packet-loss level θ and check-node degree $\gamma_1^*(k,\theta)$ jointly determine $T(k,\theta,\gamma_1^*(k,\theta))$. On the one hand, if θ becomes large, the check packets need to cover a smaller number $(k-\theta+1)$ of data packets such that fewer check packets can satisfy the covering criterion. On the other hand, a smaller $\gamma_1^*(k,\theta)$ is selected if θ becomes large. Thus, each check packet covers fewer data packets and thus more check packets are required to satisfy the covering criterion. When θ is relatively small, $\gamma_1^*(k,\theta)$ decreases quickly (see Fig. 33) and then the change of $\gamma_1^*(k,\theta)$ dominates the variation of $T(k,\theta,\gamma_1^*(k,\theta))$.

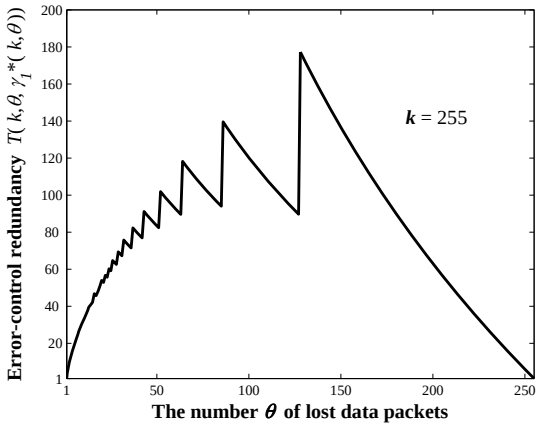


Fig. 35. Error-control redundancy $T(k, \theta, \gamma_1^*(k, \theta))$ in a TR versus packet-loss level θ under the covering criterion.

As a result, the envelop of $T(k, \theta, \gamma_1^*(k, \theta))$ increases as θ increases. In contrast, when θ is relatively large, $\gamma_1^*(k, \theta)$ decreases very slowly, then, the change of packet-loss level θ dominates the variation of $T(k, \theta, \gamma_1^*(k, \theta))$. Thus, the envelop of $T(k, \theta, \gamma_1^*(k, \theta))$ decreases as θ increases when θ is large. (ii) We observe that $T(k, \theta, \gamma_1^*(k, \theta))$ oscillates as θ increases, which can be explained as follows. From (6.11), all the packet-loss levels can be divided into a number of regions resulted from the $\lceil \cdot \rceil$ operation, within each of which $\gamma_1^*(k, \theta)$ remains the same. Consequently, $T(k, \theta, \gamma_1^*(k, \theta))$ is a decreasing function of θ within each region because with more losses, we need fewer check packets to satisfy the covering criterion. However, because $\gamma_1^*(k, \theta)$ is the decreasing function of θ (see Fig. 33), the value of $\gamma_1^*(k, \theta)$ drops between the boundary points of two neighboring regions. Then, more check packets are required in a TR to satisfy the covering criterion because each check packet covers fewer data nodes. According to the above analyses, the covering criterion is jointly controlled by θ and γ such that we can achieve the balanced error-control redundancy.

Table IV. Variables Used in Pseudo Codes

k	The number of data packets in each TG.
γ	Check-node degree which is dynamically adjusted in each TR.
T	Error-control redundancy which is dynamically adjusted in each TR.
ξ	Reliability-QoS requirement. See (6.3).
$\mathbf{D}$	$L \times k$ matrix denoting data packets of a TG. See Section B.
$\mathbf{C}_t$	$L \times T$ matrix denoting the check packets generated and multicast by the sender in the t-th TR. See Section B.
$\mathbf{P}_{t-1}$	$k \times T$ mapping structure matrix, which is used to generate $\mathbf{C}_t$ in the t-th TR, where $\mathbf{C}_t = \mathbf{D}\mathbf{P}_{t-1}$.
state	The random-number seed which is used to randomly construct mapping structures. It is initialized with the same value in the sender and all receivers.
$\mathbf{P}$	$\mathbf{P} = [\mathbf{P}_1 \ \ \mathbf{P}_2 \ \ \cdots \ \ \mathbf{P}_{t-1}]$ for the t-th TR, $t \geq 2$.

Table V. Pseudo Code for the Sender.

00. **Initial data transmission for a new TG**:
01. $t := 1$; Initialize random-number seed *state*;
    ! Both the sender and all receivers initialize *state* to the same value.
02. Update $\mathbf{D}$; Multicast $\mathbf{D}$.
03. **On receipt of feedbacks of the t-th TR from all receivers**:
04. $\theta_{\max}(t) := \max_{r=1,2,\ldots,R}\{f_r(t)\}$;
05. **if** $(\theta_{\max}(t)/k \leq \xi)$ **goto** line-00; ! QoS requirement is satisfied.
06. **else** { $t := t + 1$; ! Next TR.
07. $(\mathbf{P}_{t-1}, state) :=$ Construct$(\theta_{\max}(t-1), , state)$;
        ! Adaptively construct mapping structure matrix
08. $\mathbf{C}_t := \mathbf{D}\mathbf{P}_{t-1}$; Multicast $\theta_{\max}(t-1)$ and $\mathbf{C}_t$; ! Loss repairing.}

3. The Adaptive Graph-Code-Based Hybrid ARQ-FEC Protocol for Error-Control of Multicast

We describe our proposed adaptive two-dimensional hybrid ARQ-FEC protocol for error control of multicast by using the pseudo codes presented in Tables V-VII. The variables used in pseudo codes are defined in Table IV. We explain the pseudo codes as follows.

1) Protocol for the sender:

The sender multicasts a data TG $\mathbf{D}$ in the first TR. Then, the sender waits for feedbacks $f_r(t)$ from all receivers, where $r = 1, 2, \ldots, R$. After having received all feedbacks, the sender gets the maximum number of lost data packets $\theta_{\max}(t)$.

Table VI. Pseudo Code for the rth Receiver.

```
00. Initialization, on receipt of a new TG:
01.  t := 1; θ_max(0) := k; f_r(0) := k;
02.  Initialize random-number seed state.
     ! Both the sender and all receivers initialize state to the same value.
03. On receipt of packets from the sender in the t-th TR:
04.  if (f_r(t−1)/k ≤ ξ) {Update f_r(t) := f_r(t−1); goto line-13; }
                    ! Required reliability-QoS is satisfied
05.  if (t = 1) {Save correctly received data packets; Update f_r(t);}
                    ! Initial data transmission
06.  else {   ! Receipt of check packets
07.       Save θ_max(t−1) and correctly received check packets;
08.       (P_{t−1}, state) := Construct(θ_max(t−1), state);
                    ! Adaptively construct the mapping structure matrix
09.       if (t = 2)  P := P_1;
10.       else P := [P   P_{t−1}];
11.       Decode based on P and all correctly received packets;
12.       Update f_r(t); }
13.  Feed back f_r(t) to the sender; t := t + 1;
```

Table VII. Pseudo Code for the Mapping Structure Construction Function.

```
00. Function Construct(θ_max(t−1), state);
01.  if (θ_max(t−1) := 1) {P_{t−1} := (1, 1, ..., 1)^τ; T := 1;
02.   γ := k; }  ! Final protocol, P_{t−1} is a k × 1 column vector
03.  else {
04.   Set γ := γ*_1(k, θ_max(t−1)) by using (6.11);
05.   T := ⌈T(k, θ_max(t−1), γ)⌉ by using (6.18)-(6.20);
                ! Select check-node degree and error-control redundancy
06.   Randomly build P_{t−1} with γ, T, and state; Update state; }
07.   return (P_{t−1}, state); }
```

If $\theta_{\max}(t)/k \leq \xi$, the reliability-QoS requirement is satisfied and the sender starts to multicast the next new TG. If $\theta_{\max}(t)/k > \xi$, the sender needs to execute loss-repairing procedures in the next TR. Set $t := t + 1$. The sender constructs the mapping structure $\mathbf{P}_{t-1}$ in the t-th TR according to packet-loss level $\theta_{\max}(t-1)$. After that, the sender multicasts $\mathbf{C}_t = \mathbf{D}\mathbf{P}_{t-1}$ and $\theta_{\max}(t-1)$ to all receivers. Then, the sender goes into the state waiting for feedbacks.

2) Protocol for the rth receiver where $r = 1, 2, \ldots, R$:

The rth receiver receives a data TG $\mathbf{D}$ in the first TR. Then, the rth receiver

calculates $f_r(t)$, feeds it back to the sender, and set $t := t + 1$. On condition that $\theta_{\max}(t-1)/k > \xi$, the rth receiver will receive $\theta_{\max}(t-1)$ and a number of check packets in the current t-th TR. If the reliability-QoS requirement for the rth receiver is already satisfied, i.e., $f_r(t-1)/k \leq \xi$, the rth receiver will ignore the received packets, simply set $f_r(t) := f_r(t-1)$ and feed $f_r(t)$ back. If the reliability-QoS requirement is not satisfied, i.e., $f_r(t-1)/k > \xi$, the rth receiver will construct the mapping structure $\mathbf{P}_{t-1}$ for the current TR and start the iterative decoding (repairing) procedures. Note that although $\mathbf{P}_{t-1}$ is constructed according to the loss status in last TR, the decoding is performed based on the all $\mathbf{P}_u$, $u = 1, 2, \ldots, t-1$, and all packets correctly received for the current TG to fully make use of the received redundancy. After the repairing procedure, the rth receiver feeds the updated $f_r(t)$ back to the sender. Having sent the feedback information $f_r(t)$, the rth receiver sets $t := t+1$ and goes into the state waiting for new packets from the sender.

3) Protocol for the mapping structure construction function:

In the data TR, no mapping structure will be constructed. In loss-repairing TR's, if $\theta_{\max}(t-1) = 1$, the *final protocol* will be selected. If $\theta_{\max}(t-1) > 1$, the check-node degree γ and error-control redundancy T are selected based on (6.11) and (6.18)-(6.20). By using the same random number generating algorithm, the sender and corresponding mapping structures for the t-th TR with the selected parameters γ and T. Note that the sender and all receiver initialize the random-number seed *state* to the same value in the first TR as described in Tables V and VI. Also, as assumed in the packet-loss level $\theta_{\max}(t-1)$ can be reliably transmitted between the sender and receivers in each TR. Thus, the sender and all receivers can always get the same parameters T and γ in each TR and then construct the exactly same mapping structure.

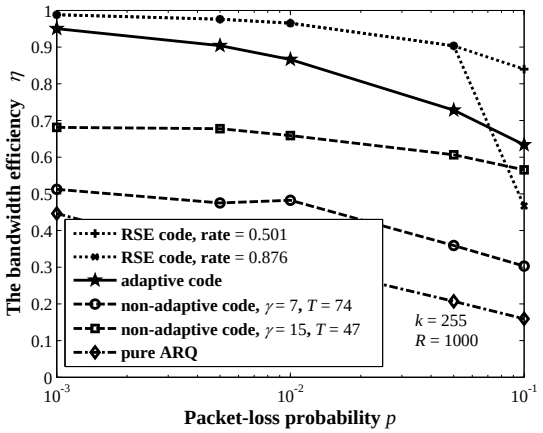


Fig. 36. Bandwidth efficiency η versus packet-loss probability p for reliable services.

E. Performance Evaluations

Using simulations, we evaluate the performance of our proposed adaptive graph-code-based multicast protocol for mobile multicast services. We also compare the performances of the adaptive graph-code-based protocol with those using RSE codes, non-adaptive graph codes (also using random mapping-structure) and pure ARQ-based approach. The TG size k is set to 255. For RSE-code-based schemes, the sender sends $\theta_{\max}(t)$ check packets in each repairing TR. We simulate the RSE codes (509,255) and (291,255) with the symbol size of 10 bits, the corresponding code rates of which are 0.501 and 0.876, respectively. Note that the two RSE codes can support maximum 254 and 36 check packets, respectively. For non-adaptive graph-code-based schemes, the sender uses the constant γ and T in each repairing TR. We simulate two sets of parameters, $(\gamma = 7, T = 74)$ and $(\gamma = 15, T = 47)$. In the simulation, we consider the packet-loss probability p equal to 0.001 through 0.1, which typically covers a wide range of channel quality for mobile wireless networks.

Figure 36 compares bandwidth efficiency for reliable services ($\xi = 0$) under dif-

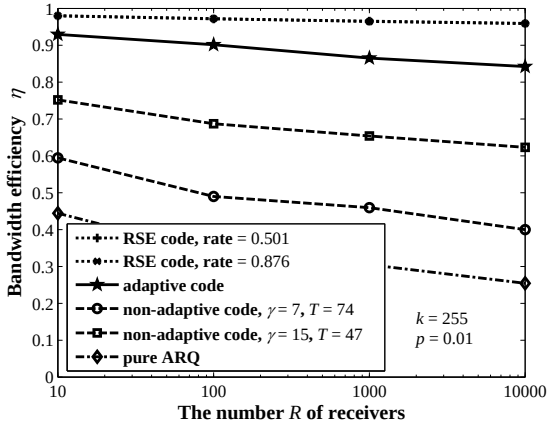


Fig. 37. Bandwidth efficiency η with different number R of receivers for reliable services.

ferent packet-loss probabilities. As shown in Fig. 36, the Fig. 36, the adaptive graph-code-based protocol can gain at least 10% higher bandwidth efficiency than those using non-adaptive graph codes. Moreover, for low packet-loss probability, bandwidth efficiency of the adaptive scheme is very close to that of RSE-code-based schemes. Under high packet-loss probability, RSE codes with high code rate (e.g., 0.876) cannot provide enough error-control redundancy and thus lead to decoding failure, retransmission, and the very low η. In contrast, our proposed adaptive scheme can support sufficient error-control redundancy to avoid this problems by using the random mapping structure for graph codes. Fig. 37 shows that the bandwidth efficiency of the adaptive graph-code-based scheme is not sensitive to the increasing of the number R of receivers. This indicates that our proposed adaptive scheme has good scalability. Fig. 38 gives the average number $E\{Q\}$ of TR's to complete the transmission of a TG for each scheme. We can see that $E\{Q\}$ of our proposed adaptive scheme is usually lower as compared to the non-adaptive graph-code-based schemes and the pure ARQ-based approach. This implies that the adaptive scheme imposes a relatively low overhead to multicast services.

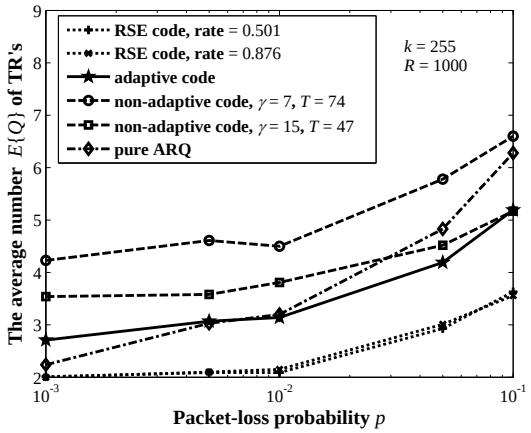


Fig. 38. Average number $E\{Q\}$ of transmission rounds versus packet-loss probability p for reliable services.

Figure 39 compares the bandwidth efficiency for various protocols under different reliability-QoS requirements ξ from 0.0 to 0.1. As shown in shown in Fig. 39, when the requirement ξ becomes larger (more losses are tolerated), the bandwidth efficiency of the adaptive graph-code-based protocol improves significantly and becomes much closer to the performance of RSE-code-based approaches. This phenomenon can be explained as follows. Using the adaptive graph codes, receivers can dynamically update packet-loss status in each TR because the iterative decoding procedure can be executed as long as any number of check packets are received. Thus, for various reliability-QoS requirements, our proposed adaptive scheme can efficiently avoid unnecessary repairing packet transmission for perfect reliability. In contrast, because the decoding of RSE codes can be performed only after k or more distinct data/check packets having been correctly received, RSE codes cannot further improve bandwidth efficiency η when reliability-QoS requirement ξ increases.

Figure 40 shows the average number $E\{Q\}$ of TR's with different ξ. We can see that our proposed adaptive scheme has much lower $E\{Q\}$ than those of RSE-based schemes when ξ is high. Fig. 41 illustrates a comprehensive Fig. 41 illustrates a

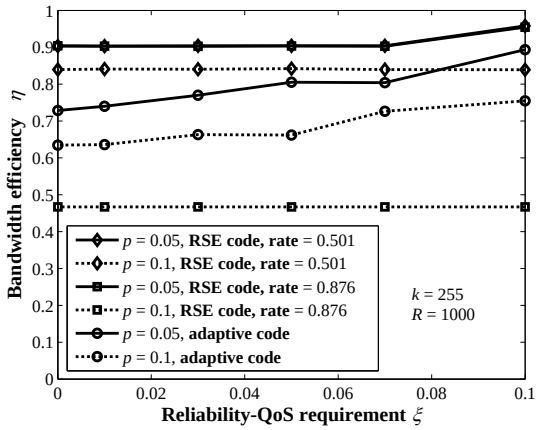


Fig. 39. Bandwidth efficiency η versus the reliability-QoS requirement ξ. $k = 255$ and $p = 0.05, 0.1$.

comprehensive effect of the reliability-QoS requirement on the average delay. In the simulation, we assume that packet length $L = 1000$ bits, bandwidth $B = 1$ Mbps and the maximum RTT among all the sender-receiver pairs equal 80 ms. With the same channel quality, we can observe that our proposed adaptive scheme can achieve even lower average delay than those of RSE-code-based schemes as ξ increases. This again verifies that our proposed adaptive protocol can efficiently avoid unnecessary repairing packet transmission such that performances can be further improved. In contrast, although RSE codes have the best erasure-correcting capability, its inflexible structure and high complexity severely limit its applicability to QoS-driven mobile multicast services. By contrast, our proposed adaptive scheme can flexibly and dynamically adjust coding structures to achieve high error-control efficiency for highly-diverse QoS requirements.

F. Conclusion

To provide flexible and efficient error-control schemes for QoS diverse multicast services, we developed and analyzed an adaptive hybrid ARQ-FEC graph-code-based

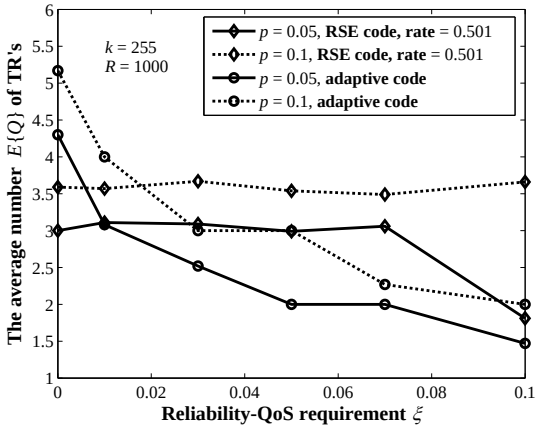


Fig. 40. Average number $E\{Q\}$ of transmission rounds versus the reliability-QoS requirement ξ. $k = 255$ and $p = 0.05, 0.1$.

erasure-correcting protocol for the QoS-driven multicast services over mobile wireless networks. The key features of our proposed scheme are two-fold: the low complexity and dynamic adaptation to packet-loss levels. The low complexity is achieved by using the graph code. In addition, the accumulatively iterative decoding procedures of graph codes can flexibly adapt to the variations of reliability-QoS requirements. To increase the error-control efficiency, we proposed a two-dimensional adaptive error-control scheme, which dynamically adjusts both the error-control redundancy and code-mapping structures in each adaptation step according to packet-loss levels. By deriving and identifying the closed-form nonlinear analytical expression between the optimal check-node degree and the packet-loss level for any given code-block length, we proposed the nonuniformed adaptive coding structures to achieve high error-control efficiency. Furthermore, we developed a loss covering strategy to determine the error-control redundancy in each transmission round and derive the corresponding analytical expressions of the error-control redundancy. Using the proposed two-dimensional nonuniformed adaptive error-control scheme, we developed an efficient hybrid ARQ-FEC protocol for multicast. We evaluated the proposed proto-

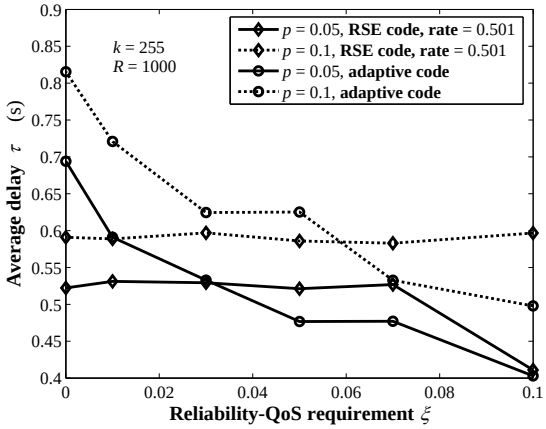


Fig. 41. Average delay τ for the transmission of a TG versus the reliability-QoS requirement ξ. $k = 255$ and $p = 0.05, 0.1$. B=1 Mbps, L=1000 bits, RTT=80ms.

col through simulation experiments. The simulation results show that our proposed adaptive scheme can achieve high error-control efficiency for QoS-driven multicast services while introducing low computational complexity and implementation overhead.

CHAPTER VII

RESOURCE ALLOCATION FOR DOWNLINK STATISTICAL MULTIUSER DELAY-QOS PROVISIONINGS IN CELLULAR WIRELESS NETWORKS

A. Introduction

The rapidly increasing demands on wireless orientated applications accelerate the evolution of next-generation wireless networks. In this chapter, we focus on downlink multiuser delay quality-of-service (QoS) provisionings via dynamic resource allocation in cellular networks. Although there have been a great deal of research on resource allocation for this scenario from multiuser information theory perspective [10, 15, 18, 39], these works are not comprehensive to characterize a wide range of *delay* QoS requirements. In [10, 15], the authors derived the ergodic capacity region for distributing independent information to multiple users over broadcast fading channels, which corresponds to the scenario without any delay constraints. On the other hand, the authors of [18] proposed the optimal power allocation for outage capacities over broadcast fading channels, where the service rate for each user is a constant with a certain outage probability. This framework imposes extremely stringent delay requirements for wireless services. Likewise, there have been many references focusing only on the above two extreme cases. However, because one of the most attractive features of the next-generation networks is the capability of supporting the diverse delay QoS requirements, we need to take various delay QoS constraints into consideration to develop the new resource allocation schemes.

To achieve efficient wireless communications while supporting diverse delay QoS requirements, we employ the *effective capacity* as the main performance metric in this chapter. The effective capacity was defined in [8] to evaluate the capability of a

wireless service process in supporting data transmission subject to a *statistical* delay QoS requirement metric, called *QoS exponent* and denoted by θ. The higher θ corresponds to the more stringent delay constraint. Also, θ can continuously vary from 0 to ∞, and thus a wide spectrum of QoS constraints can be readily characterized by a general model. Moreover, the authors of [19] showed that the effective capacity model can also yield insightful observation from the information theory perspective. In [19], by integrating the information theory with the effective capacity approach, we proposed a framework to maximize the effective capacity of a point-to-point wireless link via power adaptation, where we adopts the QoS exponent θ as the delay constraint. We showed that the effective capacity builds connections between the ergodic capacity and zero-outage capacity through θ. As θ approaches 0, our study becomes to obtain the ergodic capacity. When θ goes to ∞, the framework reduces to deriving the zero-outage capacity.

However, incorporating the effective capacity model into multiuser communications still faces significant challenges, which are not encountered in the single wireless links. First, the effective-capacity optimization over a single link addressed in [19] only deals with the fixed time slots and/or bandwidths, but the multiuser systems often have to dynamically allocate these resources based on mobile users' channel state information (CSI). Second, the multiuser systems usually need to balance the performances among all mobile users where the single-link based strategies are not applicable. How to allocate resources among mobile users depends closely on the users' QoS requirements and their priorities to receive the wireless services.

To overcome the above problems, we formulate the sum *effective capacity* maximization problem via power and time-slot length adaptation subject to the *proportional-effective-capacity* constraint and the diverse *statistical* delay QoS requirements. To effectively deal with the non-additive objective function, we decompose the original

optimization problem into two sub-problems and then derive the optimal power and time-slot adaptation policy. We also develop a suboptimal equal-length TD policy. We simulate our proposed schemes to show the impact of QoS provisionings on the resource allocation across different users and on the network performance.

The rest of the chapter is organized as follows. Section B describes the system model. Section C formulates the optimization problem for adaptive resource allocation. Section D derives the optimal power and time-slot length adaption policy. Section E evaluates our proposed schemes through simulations. The chapter concludes with Section F.

B. The System Model

We consider the scenario where the base station (BS) transmits independent messages to N different mobile users over broadcast fading channels in a cellular wireless network, as shown in Fig. 42. The fading channels are assumed to be ergodic, stationary, and flat block-fading processes. In particular, the complex channel gains for the link between the BS and mobile users are invariant within a time frame with length T, but vary independently from frame to frame. We index the time frames by k for $k = 1, 2, \ldots$. The BS employs time division (TD) for multiple user access, where each time frame of length T is divided into N time slots. In the kth time frame, we denote the length of the nth time-slot by $T_n[k]$, which is allocated to the nth mobile user, where $\sum_{n=1}^{N} T_n[k] = T$. In our derivations, we use the more general constraint $\sum_{n=1}^{N} T_n[k] \leq T$ by taking the outage transmission states into account. Also, we denote the time proportion of the nth time slot by $\varphi_n[k]$, where $\varphi_n[k] = T_n[k]/T$. Then, in the nth mobile user's time slot, the physical-layer model for the above system can be expressed by $y_n[k] = h_n[k]x_n[k] + z_n[k], n = 1, 2, \ldots, N$, where $x_n[k]$ is

the complex signal transmitted to the nth mobile user with signal bandwidth B, $h_n[k]$ is the complex channel gain between the BS and the nth mobile user, $y_n[k]$ is the corresponding received signal, and $z_n[k]$ is the complex additive white Gaussian noise (AWGN) with power spectral density $\sigma_0/2$. The transmit power used in the nth time slot, denoted by $\mu_n[k]$, is given by $\mu_n[k] \triangleq \mathbb{E}\{|x_n[k]|^2\}$, where $\mathbb{E}\{\cdot\}$ denotes the expectation. Furthermore, we define $\nu_n[k] \triangleq \mu_n[k]\varphi_n[k]$, which is the total power consumed by the nth mobile user within the kth frame. For the convenience of presentation, we further define the vectors: $\boldsymbol{\varphi}[k] \triangleq (\varphi_1[k], \varphi_2[k], \ldots, \varphi_N[k])$ and $\boldsymbol{\nu}[k] \triangleq (\nu_1[k], \nu_2[k], \ldots, \nu_N[k])$.

Without loss of generality, we characterize the channel state information (CSI) by $\gamma_n[k] \triangleq |h_n[k]|^2\mathcal{P}/(\sigma_0 B)$, which is called the *reference* SNR, where $\mathcal{P}$ is the average power threshold (to be detailed later). Correspondingly, we define the CSI vector as $\boldsymbol{\gamma}[k] \triangleq (\gamma_1[k], \gamma_2[k], \ldots, \gamma_N[k])$ to represent a *fading state*. Inspired by the stationary properties of the fading channels, we use $\overline{\gamma}_n$ to represent the mean of $\gamma_n[k]$, and denote the joint probability density function (pdf) of $\boldsymbol{\gamma}[k]$ by $f_{\boldsymbol{\Gamma}}(\boldsymbol{\gamma})$. Throughout this chapter, we assume that CSI $\gamma_n[k]$, $n = 1, 2, \ldots, N$, can be estimated accurately at the receivers and reliably fed back to the BS without delay. When the context is clear, we drop the frame index $[k]$ to simplify the notation.

The BS maintains a separate queue for each mobile user, as shown in Fig. 42, and regulates the resource allocation, including the time-slot and power allocation, based on the CSI, such that the QoS requirements for each user can be efficiently supported. Clearly, the power and time-slot allocation can be characterized by the pair $(\boldsymbol{\varphi}, \boldsymbol{\nu})$, where $\boldsymbol{\varphi}$ is the time-proportion vector and $\boldsymbol{\nu}$ is the power vector. In addition, the adaptive transmission needs to satisfy the average power constraint, which is given by $\mathbb{E}_{\boldsymbol{\gamma}}\{\sum_{n=1}^{N} \nu_n\} \leq \mathcal{P}$, where $\mathbb{E}_{\boldsymbol{\gamma}}\{\cdot\}$ is the expectation over all $\boldsymbol{\gamma}$ and $\mathcal{P}$ is the average power threshold. Moreover, we assume that the adaptive modulation

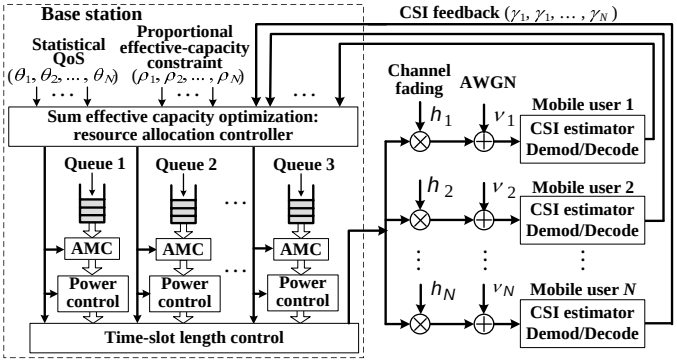


Fig. 42. The downlink communication model in a cellular wireless network.

and coding (AMC) technique is used with the capacity-achieving codes, such that the Shannon capacities for all mobile users in each frame can be achieved. Accordingly, given $(\boldsymbol{\varphi}, \boldsymbol{\nu})$, we set the data transmission rate (bits/frame) for the nth mobile user, denoted by R_n, equal to

$$R_n = \begin{cases} BT\varphi_n \log_2\left(1 + \frac{\nu_n \gamma_n}{\varphi_n}\right), & \text{if } \varphi_n > 0,\ \nu_n > 0; \\ 0, & \text{if } \varphi_n = 0,\ \nu_n = 0, \end{cases} \tag{7.1}$$

where $n = 1, 2, \ldots, N$.

C. The Optimization Problem Formulation

1. Statistical QoS Requirements

As discussed in Chapter II, the *effective capacity* [8] is a powerful approach to evaluate and devise the capability of a wireless channel to support data transmissions with diverse *statistical* quality of service (QoS) guarantees. In this chapter, we use QoS exponent θ_n as the QoS requirements from the nth mobile users.

2. Framework for Adaptive Resource Allocation with Statistical QoS Provisionings

We denote the QoS exponent for the nth mobile user by θ_n, $0 < \theta_n < \infty$, $n = 1, 2, \ldots, N$. Based on Eq. (2.8), the nth mobile user's effective capacity, denoted by $C_n(\theta_n)$, is equal to

$$C_n(\theta_n) = -\frac{1}{\theta_n} \log \left(\mathbb{E}_{\boldsymbol{\gamma}} \left\{ e^{-\theta_n R_n} \right\} \right), \quad \text{bits/frame.} \tag{7.2}$$

Correspondingly, the normalized effective capacity (bits/s/Hz) of the nth receiver, denoted by $\mathscr{C}_n(\beta_n)$, is defined as $\mathscr{C}_n(\beta_n) \triangleq C_n(\theta_n)/(TB)$, where $\beta_n \triangleq \theta_n TB/\log 2$ is the normalized QoS exponent. To optimize the overall network throughput achievable by the entire cellular system, we propose to optimize the *sum effective capacity* over all mobile users, which is denoted by $C_{\text{sum}}(\boldsymbol{\theta})$ and expressed as

$$C_{\text{sum}}(\boldsymbol{\theta}) \triangleq \sum_{n=1}^{N} TB\mathscr{C}_n(\beta_n), \tag{7.3}$$

where $\boldsymbol{\theta} \triangleq (\theta_1, \theta_2, \ldots, \theta_N)$. Also, the normalized sum effective capacity, denoted by $\mathscr{C}_{\text{sum}}(\boldsymbol{\theta})$ is defined by $\mathscr{C}_{\text{sum}}(\boldsymbol{\beta}) \triangleq C_{\text{sum}}(\boldsymbol{\theta})/(TB)$, where $\boldsymbol{\beta} \triangleq (\beta_1, \beta_2, \ldots, \beta_N)$. Moreover, we have the following system constraints:

Constraint 1. *Time slot constraint:* in any fading state, $\sum_{n=1}^{N} \varphi_n \leq 1$ needs to be satisfied.

Constraint 2. *Average power constraint:* as aforementioned in the above, the average sum power over all mobile users needs to be upper-bounded by a specified threshold $\mathcal{P}$, i.e., $\sum_{n=1}^{N} \mathbb{E}_{\boldsymbol{\gamma}} \{\nu_n\} \leq \mathcal{P}$.

Constraint 3. *Proportional-effective-capacity constraint:* each user is allocated a service-class label, denoted by ρ_n ($0 < \rho_n < \infty$, $n = 1, 2, \ldots, N$), under which $\mathscr{C}_n(\beta_n)$'s need to satisfy $\mathscr{C}_1(\beta_1)/\rho_1 = \mathscr{C}_2(\beta_2)/\rho_2 = \cdots = \mathscr{C}_N(\beta_N)/\rho_N$. Clearly, the higher ρ_n implies the higher priority and the user with higher ρ_n achieves the larger

effective capacity. We further define $\boldsymbol{\rho} \triangleq (\rho_1, \rho_2, \ldots, \rho_N)$.

To optimize the sum effective capacity, the time-slot and power allocation policy $(\boldsymbol{\varphi}, \boldsymbol{\nu})$ needs to be the function of not only the CSI $\boldsymbol{\gamma}$, but also the QoS parameters $\boldsymbol{\theta}$ and $\boldsymbol{\rho}$. Then, we derive the optimal resource allocation scheme by solving the following Problem **VII-*A***.

VII-*A* : Sum effective capacity optimization

$$\max_{(\boldsymbol{\varphi}, \boldsymbol{\nu})} \left\{ \mathscr{C}_{\text{sum}}(\boldsymbol{\beta}) \right\} \tag{7.4}$$

$$\text{s.t. :} \quad 1). \quad \frac{\mathscr{C}_1(\beta_1)}{\rho_1} = \frac{\mathscr{C}_2(\beta_2)}{\rho_2} = \cdots = \frac{\mathscr{C}_N(\beta_N)}{\rho_N}, \tag{7.5}$$

$$2). \quad \sum_{n=1}^{N} \mathbb{E}_{\boldsymbol{\gamma}} \{\nu_n\} \leq \mathcal{P}, \quad \nu_n \geq 0, \tag{7.6}$$

$$3). \quad \sum_{n=1}^{N} \varphi_n \leq 1, \quad \varphi_n \geq 0. \tag{7.7}$$

Accordingly, the optimal adaptation policy to Problem **VII-*A*** is denoted by $(\boldsymbol{\varphi}^{\text{opt}}, \boldsymbol{\nu}^{\text{opt}})$.

D. The Optimal Power and Time-Slot Length Adaptation Policy

1. Decomposition of Problem **VII-*A***

Since $\mathscr{C}_{\text{sum}}(\boldsymbol{\beta})$ is not an additive objective function over all fading states, it is intractable to process each fading state separately, and thus it is hard to directly derive the optimal solution for Problem **VII-*A***. To overcome this problem, we decompose

Problem **VII-*A*** into two sub-problems (SP) as follows.

$$\text{SP I}: \quad \min_{(\boldsymbol{\varphi},\boldsymbol{\nu})} \left\{ \sum_{n=1}^{N} \mathbb{E}_{\boldsymbol{\gamma}} \{\nu_n\} \right\} \tag{7.8}$$

$$\text{s.t.} : 1). \ A_n\big(\varphi_n, \nu_n\big) - 2^{-\beta_n \rho_n \mathscr{C}_0} \leq 0, \ \forall\, n, \tag{7.9}$$

$$2). \ \textstyle\sum_{n=1}^{N} \varphi_n - 1 \leq 0, \quad \varphi_n \geq 0, \ \forall\, \boldsymbol{\gamma}, \tag{7.10}$$

where

$$A_n\big(\varphi_n, \nu_n\big) \triangleq \mathbb{E}_{\boldsymbol{\gamma}} \left\{ e^{-\theta_n R_n} \right\} = \mathbb{E}_{\boldsymbol{\gamma}} \left\{ \left(1 + \tfrac{\nu_n \gamma_n}{\varphi_n} \right)^{-\beta_n \varphi_n} \right\} \tag{7.11}$$

and $\mathscr{C}_0 \geq 0$ is a specified constant. Clearly, SP I minimizes the average power required to guarantee $\mathscr{C}_n(\beta_n) \geq \rho_n \mathscr{C}_0$, $n = 1, 2, \ldots, N$, with the given $\mathscr{C}_0$. We denote the optimal adaptation policy to SP I by $(\boldsymbol{\varphi}^*, \boldsymbol{\nu}^*)$, and denote the average power consumed under $(\boldsymbol{\varphi}^*, \boldsymbol{\nu}^*)$ by $\overline{P}^*(\mathscr{C}_0)$. Next, consider SP II give by

$$\text{SP II}: \ \text{Solve for } \mathscr{C}_0^* \text{ such that}$$

$$\overline{P}^*(\mathscr{C}_0^*) = \mathcal{P}. \tag{7.12}$$

With the obtained $\mathscr{C}_0^*$, we then get the solution to Problem **VII-*A*** as $\boldsymbol{\nu}^{\text{opt}} = \boldsymbol{\nu}^*|_{\mathscr{C}_0 = \mathscr{C}_0^*}$ and $\boldsymbol{\varphi}^{\text{opt}} = \boldsymbol{\varphi}^*|_{\mathscr{C}_0 = \mathscr{C}_0^*}$. The proof for the optimality of solution obtained through the above decomposition method is provided in the following sections.

2. The Optimal Solution to Sub-Problem I

Define $\mathcal{S}_n \triangleq \{(0,0)\} \cup \mathbb{R}_{++}^2$, $n = 1, 2, \ldots, N$, which is a convex set, where $\mathbb{R}_{++}$ denotes the set of positive real numbers. According to the definition of convex function [77, Chapter 3.14], we can easily prove that $(1 + \nu_n \gamma_n / \varphi_n)^{-\beta_n \varphi_n}$ is strictly convex over $(\varphi_n, \nu_n) \in \mathcal{S}_n$. Then, SP I is a convex problem based on the criteria in [77, Chapter 4.2.1]. Thus, we can solve SP I by applying the standard Lagrangian method and

the Karush-Kuhn-Tucker (KKT) conditions, and derive the optimal adaptation policy in Theorem 11 as follows.

Theorem 11. *The optimal adaptation policy* $(\boldsymbol{\varphi}^*, \boldsymbol{\nu}^*)$ *to SP I is given by*

$$\begin{cases} \varphi_n^* = \dfrac{1}{\beta_n}\left[\dfrac{\log\left(\beta_n\gamma_n\psi_n^*\right)}{1+W\left(\frac{\gamma_n\epsilon_{\boldsymbol{\gamma}}^*-1}{e}\right)}-1\right]^+; \\ \nu_n^* = \dfrac{\varphi_n^*}{\gamma_n}\left\{\exp\left(1+W\left(\dfrac{\gamma_n\epsilon_{\boldsymbol{\gamma}}^*-1}{e}\right)\right)-1\right\}, \end{cases} \tag{7.13}$$

where $W(z)$ *is the Lambert-W function [74],* ψ_n^**, for* $n = 1, 2, \ldots, N$*, is fixed in all fading states, and* $\epsilon_{\boldsymbol{\gamma}}^*$ *varies with* $\boldsymbol{\gamma}$*. In fading state* $\boldsymbol{\gamma}$*, if* $\gamma_n \leq 1/(\beta_n\psi_n^*)$ *for all* $1 \leq n \leq N$*, we have* $\epsilon_{\boldsymbol{\gamma}}^* = 0$*; otherwise,* $\epsilon_{\boldsymbol{\gamma}}^*$ *is the solution to the equation*

$$\textstyle\sum_{n=1}^{N} \varphi_n^* = 1. \tag{7.14}$$

Moreover, ψ_n^**, for* $n = 1, 2, \ldots, N$*, needs to be selected such that*

$$A_n\big(\varphi_n, \nu_n\big) - 2^{-\beta_n\rho_n\mathscr{C}_0} = 0, \ \forall\, 1 \leq n \leq N. \tag{7.15}$$

Proof. Construct the Lagrangian function, denoted by $\mathcal{L}(\boldsymbol{\varphi}, \boldsymbol{\nu}; \boldsymbol{\psi}, \epsilon_{\boldsymbol{\gamma}})$, as

$$\begin{aligned} \mathcal{L}(\boldsymbol{\varphi}, \boldsymbol{\nu}; \boldsymbol{\psi}, \epsilon_{\boldsymbol{\gamma}}) = &\sum_{n=1}^{N} \mathbb{E}_{\boldsymbol{\gamma}}\{\nu_n\} + \mathbb{E}_{\boldsymbol{\gamma}}\left\{\epsilon_{\boldsymbol{\gamma}}\left(\sum_{n=1}^{N}\varphi_n - 1\right)\right\} \\ &+ \sum_{n=1}^{N} \psi_n\Big(A_n\big(\varphi_n, \nu_n\big) - 2^{-\beta_n\rho_n\mathscr{C}_0}\Big), \end{aligned} \tag{7.16}$$

where $\boldsymbol{\psi} \triangleq (\psi_1, \psi_2, \ldots, \psi_N)$, $\psi_n \geq 0$ is the Lagrangian multiplier associated with the constraint of Eq. (7.9), and $\epsilon_{\boldsymbol{\gamma}} \geq 0$ is Lagrangian multiplier associated with the constraint of Eq. (7.10) in fading state $\boldsymbol{\gamma}$. Based on the convex optimization theory, the optimal adaptation policy $(\boldsymbol{\varphi}^*, \boldsymbol{\nu}^*)$ and the corresponding Lagrangian

multipliers, denoted by $\boldsymbol{\psi}^*$ and $\epsilon^*_{\boldsymbol{\gamma}}$, are the solution to the Karush-Kuhn-Tucker (KKT) conditions [77] summarized in the following Eq. (7.17) through Eq. (7.20).

For all $1 \leq n \leq N$, if $\varphi_n^* > 0$ and $\nu_n^* > 0$, (φ_n^*, ν_n^*) satisfies

$$\begin{cases} \dfrac{\partial \mathcal{L}(\boldsymbol{\varphi}^*, \boldsymbol{\nu}^*; \boldsymbol{\psi}^*, \epsilon^*_{\boldsymbol{\gamma}})}{\partial \varphi_n} = 0; \\ \dfrac{\partial \mathcal{L}(\boldsymbol{\varphi}^*, \boldsymbol{\nu}^*; \boldsymbol{\psi}^*, \epsilon^*_{\boldsymbol{\gamma}})}{\partial \nu_n} = 0; \\ \varphi_n^* > 0, \nu_n^* > 0; \end{cases} \tag{7.17}$$

if Eq. (7.17) does not have solutions, we get

$$\varphi_n^* = 0, \quad \nu_n^* = 0. \tag{7.18}$$

Furthermore, the KKT conditions require the following equations:

$$\epsilon^*_{\boldsymbol{\gamma}} \left(\textstyle\sum_{n=1}^N \varphi_n^* - 1 \right) = 0, \ \epsilon^*_{\boldsymbol{\gamma}} \geq 0, \ \forall \boldsymbol{\gamma} \ , \tag{7.19}$$

$$\psi_n^* \left(A_n\big(\varphi_n^*, \nu_n^*\big) - 2^{-\beta_n \rho_n \mathscr{C}_0} \right) = 0, \ \psi_n^* \geq 0, \ \forall n \ . \tag{7.20}$$

Plugging Eqs. (7.11) and (7.16) into Eq. (7.17), we obtain the following equations:

$$0 = \epsilon_{\boldsymbol{\gamma}} - \psi_n \beta_n \left(1 + \frac{\gamma_n \nu_n}{\varphi_n}\right)^{-1-\varphi_n \beta_n} \times \left[\left(1 + \frac{\gamma_n \nu_n}{\varphi_n}\right) \log\left(1 + \frac{\gamma_n \nu_n}{\varphi_n}\right) - \frac{\gamma_n \nu_n}{\varphi_n} \right]; \tag{7.21}$$

$$0 = 1 - \psi_n \beta_n \gamma_n \left(1 + \frac{\gamma_n \nu_n}{\varphi_n}\right)^{-1-\varphi_n \beta_n}, \tag{7.22}$$

whose solutions are φ_n^* and ν_n^*, if $\varphi_n^* > 0$ and $\nu_n^* > 0$. Solving Eqs. (7.21)-(7.22) and applying Eq. (7.18), we obtain Eq. (7.13) in Theorem 11. If $\gamma_n \leq 1/(\beta_n \psi_n^*)$ for all $n = 1, 2, \ldots, N$, through Eq. (7.13) we have $\sum_{n=1}^N \varphi_n^* = 0$ regardless of $\epsilon^*_{\boldsymbol{\gamma}}$. Then, to satisfy Eq. (7.19), we get $\epsilon^*_{\boldsymbol{\gamma}} = 0$. If there is some n such that $\gamma_n > 1/(\beta_n \psi_n^*)$, $\epsilon^*_{\boldsymbol{\gamma}} \to 0$ leads to $\varphi_n^* \to \infty$, which is not feasible. Thus, we must have $\epsilon^*_{\boldsymbol{\gamma}} > 0$, and obtain Eq. (7.14) by using Eq. (7.19). Similarly, $\psi_n^* = 0$ results in $(\varphi_n^*, \nu_n^*) = (0, 0)$ for all $\boldsymbol{\gamma}$, which violates Eq. (7.9). Therefore, $\psi_n^* > 0$ holds for all n and we get Eq. (7.15) by

solving Eq. (7.20), which completes the proof of Theorem 11. □

Equation (7.15) in Theorem 11 shows that under the adaptation policy $(\boldsymbol{\varphi}^*, \boldsymbol{\nu}^*)$, we get $\mathscr{C}_n(\theta_n) = \rho_n \mathscr{C}_0$, $\forall\, n$, i.e., the proportional-effective-capacity constraint is satisfied. Moreover, we have the following Corollary 1.

Corollary 1. *The average power $\overline{P}^*(\mathscr{C}_0)$ consumed under $(\boldsymbol{\varphi}^*, \boldsymbol{\nu}^*)$ is a monotonically increasing function of $\mathscr{C}_0$.*

Proof. Assume that there exist $\mathscr{C}_0'$ and $\mathscr{C}_0''$, where $\mathscr{C}_0' < \mathscr{C}_0''$, such that $\overline{P}^*(\mathscr{C}_0') > \overline{P}^*(\mathscr{C}_0'')$ holds. Since $2^{-\beta_n \rho_n \mathscr{C}_0''} < 2^{-\beta_n \rho_n \mathscr{C}_0'}$, the policy generated by Theorem 11 with $\mathscr{C}_0 = \mathscr{C}_0''$ is also feasible to SP I with $\mathscr{C}_0 = \mathscr{C}_0'$. Based on Theorem 11 we then get $\overline{P}^*(\mathscr{C}_0') \leq \overline{P}^*(\mathscr{C}_0'')$, however, which contradicts the assumption $\overline{P}^*(\mathscr{C}_0') > \overline{P}^*(\mathscr{C}_0'')$. Thus, Corollary 1 follows by contradiction. □

To implement the policy generated by Theorem 11, we need to solve for $\boldsymbol{\psi}^* \triangleq (\psi_1^*, \psi_2^*, \ldots, \psi_N^*)$ and $\epsilon_{\boldsymbol{\gamma}}^*$. Unfortunately, the general analytical expressions for these solutions are usually intractable. Given $\boldsymbol{\psi}^*$, Theorem 11 shows that $\epsilon_{\boldsymbol{\gamma}}^*$ can be obtained in each fading state by solving $\sum_{n=1}^{N} \varphi_n^* = 1$. Moreover, Eq. (7.13) shows that $\sum_{n=1}^{N} \varphi_n^*$ is a decreasing function of $\epsilon_{\boldsymbol{\gamma}}^*$, and thus it is easy to determine $\epsilon_{\boldsymbol{\gamma}}^*$ by using the numerical searching techniques.

Based on the dual convex optimization theory [77, Chapter 5], $\boldsymbol{\psi}^*$ is also the maximizer of the *Lagrangian dual function* [77], which is concave over $\boldsymbol{\psi}^*$. We then can apply the widely used iterative subgradient optimization method [90] to optimize the Lagrangian dual function and track the maximizer $\boldsymbol{\psi}^*$.

3. The Optimal Solution to Problem **VII-*A***

Since $\overline{P}^*(\mathscr{C}_0)$ is an increasing function of $\mathscr{C}_0$, it is not difficult to determine $\mathscr{C}_0^*$ in Eq. (7.12) by using numerical searching techniques. Moreover, based on the mono-

tonic property of $\overline{P}^*(\mathscr{C}_0)$, we can show that the decomposition method developed in Section D-1 yields the optimal adaption policy to Problem **VII-*A***, which is summarized in Theorem 12 as follows.

Theorem 12. *The optimal adaptation policy to Problem* **VII-*A*** *is given by*

$$(\boldsymbol{\varphi}^{\text{opt}}, \boldsymbol{\nu}^{\text{opt}}) = (\boldsymbol{\varphi}^*, \boldsymbol{\nu}^*)|_{\mathscr{C}_0=\mathscr{C}_0^*}, \tag{7.23}$$

where $(\boldsymbol{\varphi}^*, \boldsymbol{\nu}^*)$ *is determined by Theorem 11 and* $\mathscr{C}_0^*$ *is the solution to Eq. (7.12).*

Proof. Assume there is a policy $(\boldsymbol{\varphi}', \boldsymbol{\nu}') \neq (\boldsymbol{\varphi}^{\text{opt}}, \boldsymbol{\nu}^{\text{opt}})$, which is feasible to Problem **VII-*A*** with $\mathscr{C}_n(\beta_n)/\rho_n = \mathscr{C}_0'$ and the average power equal to $\overline{P}'$. To prove the optimality of $(\boldsymbol{\varphi}^{\text{opt}}, \boldsymbol{\nu}^{\text{opt}})$, we only need to show $\mathscr{C}_0^* \geq \mathscr{C}_0'$. We then derive

$$\overline{P}^*(\mathscr{C}_0^*) \overset{(a)}{=} \mathcal{P} \overset{(b)}{\geq} \overline{P}' \overset{(c)}{\geq} \overline{P}^*(\mathscr{C}_0'), \tag{7.24}$$

where (a) follows by using Eq. (7.12), (b) holds because $(\boldsymbol{\varphi}', \boldsymbol{\nu}')$ is feasible to Problem **VII-*A***, and (c) is due to Theorem 11. Finally, applying Corollary 1 to Eq. (7.24), we get $\mathscr{C}_0^* \geq \mathscr{C}_0'$ and complete the proof of Theorem 12. □

It is worth noting that when $N = 1$, the one-to-many communication network becomes the point-to-point wireless communication link and we only need to regulate the transmit power. Correspondingly, we can show that the optimal power allocation given in Eq. (7.13) will then reduce to the same form as the optimal power-adaptation scheme derived in [19], where the effective capacity is maximized for a single wireless link with the average power constraint.

4. The Suboptimal Equal-Length TD Policy

We also develop a suboptimal but simpler scheme for resource allocation. In particular, we fix $\varphi_n = 1/N$ in each fading state and only regulate transmit power vector $\boldsymbol{\nu}$

to optimize $\mathscr{C}_{\text{sum}}(\boldsymbol{\beta})$ under the proportional-effective-capacity constraint in Eq. (7.5). Because we equally allocate time-slot length among mobile users, we call this suboptimal strategy the *equal-length TD* policy. This problem can be solved by applying the similar decomposition method used in Section D-1. We first specify a $\mathscr{C}_0$, and then minimize the total average power $\sum_{n=1}^{N} \mathbb{E}_{\boldsymbol{\gamma}}\{\nu_n\}$ required to satisfy $\mathscr{C}_n(\beta_n) = \rho_n \mathscr{C}_0$, $n = 1, 2, \ldots, N$. Since the time-slot length is fixed, minimizing the total average power is equivalent to minimizing $\mathbb{E}\{\nu_n\}$ for N point-to-point links separately (see discussions of Theorem 12 for power allocation over the point-to-point link). Next, we search for $\mathscr{C}_0$ such that the total average power reaches the average power threshold $\mathcal{P}$. Then, we obtain the equal-length TD policy with such a $\mathscr{C}_0$ and the separate power control for N mobile users.

E. Simulation Evaluations

We evaluate our proposed adaptive resource allocation schemes through simulations. In the simulations, we employ Nakagami-m [70] as the typical fading-channel model. Fig. 43(a) plots the normalized effective capacity of mobile user 1 versus that of mobile user 2 for a two mobile-users network, where each plot is obtained by letting ρ_1/ρ_2 vary from 0 to ∞. We can see from Fig. 43(a) that our derived optimal adaptation policy achieves much larger effective capacities as compared to the equal-length TD policy. Thus, our derived optimal policy can use the wireless resources more efficiently to optimize the overall throughput for the entire network. Also, Fig. 43(a) shows that when β_n's for both mobile users get larger, implying more stringent delay QoS requirements, the achievable effective capacities for both mobile users become smaller. Fig. 43(b) illustrates the impact of QoS exponents on resource allocation. In particular, we fix $\beta_2 = 1$ but change β_1. Fig. 43(b) shows that a higher β_1 leads

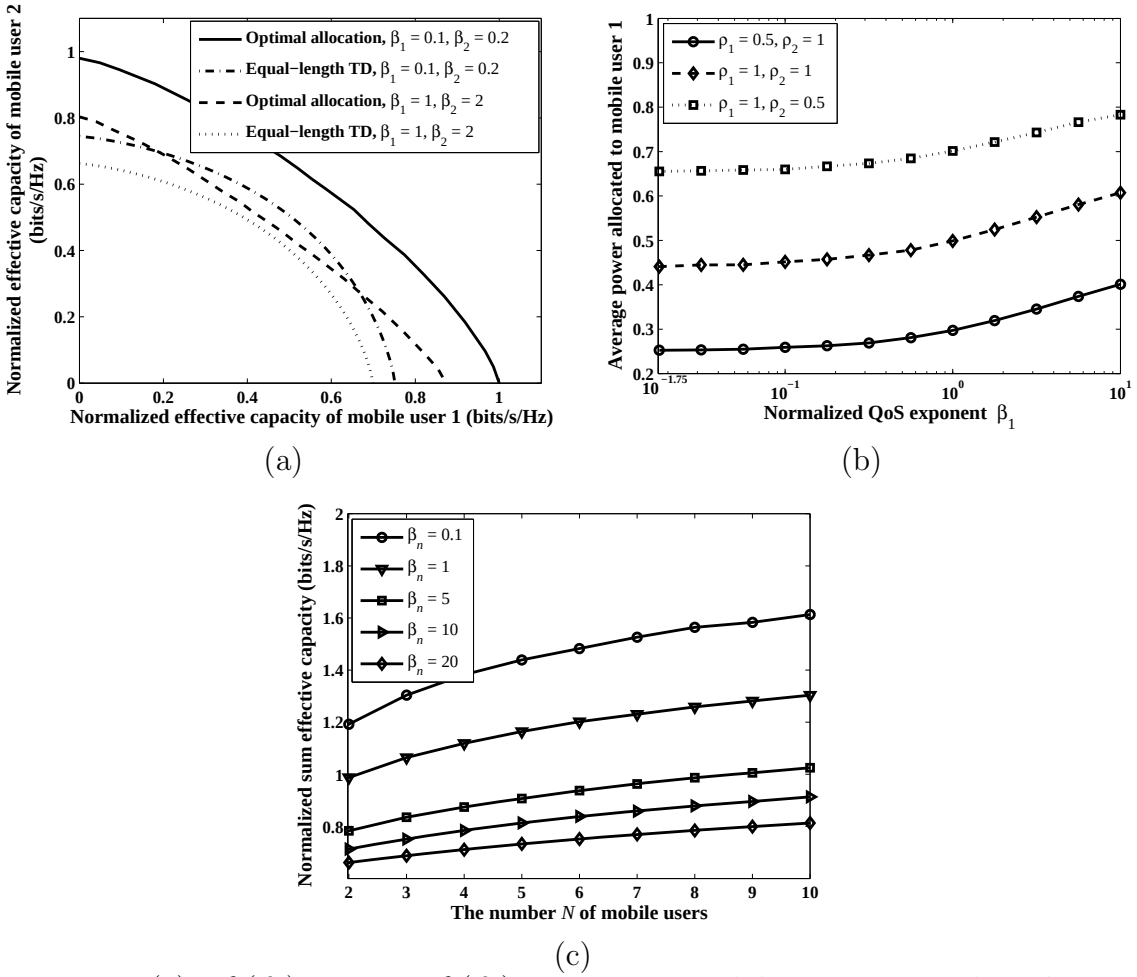


Fig. 43. (a) $\mathscr{C}_1(\beta_1)$ versus $\mathscr{C}_2(\beta_2)$ in a two mobile-users network, where $\overline{\gamma}_1 = \overline{\gamma}_2 = 1$, γ_n's follow independent Nakagami-m fading with $m = 2$, and $\mathcal{P} = 1$. (b) Average power $\mathbb{E}_{\boldsymbol{\gamma}}\{\nu_1\}$ versus the 1st mobile user's normalized QoS exponent β_1, where $\beta_2 = 1$ for mobile user 2 is fixed. γ_n's follow independent Nakagami-m fading with $m = 2$, $\overline{\gamma}_1 = \overline{\gamma}_2 = 1$, and $\mathcal{P} = 1$. (c) Normalized sum effective capacity $\mathscr{C}_{\text{sum}}(\boldsymbol{\beta})$ versus the number N of mobile users, where γ_n's follow independent Nakagami-m fading with $m = 2$, $\overline{\gamma}_1 = \overline{\gamma}_2 = \cdots = \overline{\gamma}_N = 1$, $\rho_1 = \rho_2 = \cdots = \rho_N = 1$, $\beta_1 = \beta_2 = \cdots = \beta_N$, and $\mathcal{P} = 1$.

to more power proportion allocated to user 1. This is expected because in order to support more stringent delay requirement, more power is used to overcome deep channel fading. However, this will cause the degradation of the overall network performance due to the inefficient consumption of the limited resources. Fig. 43(c) studies the effects of the QoS provisionings and the number of mobile users on the network throughput performance. Specifically, Fig. 43(c) plots the normalized sum effective capacity $\mathscr{C}_{\text{sum}}(\boldsymbol{\beta})$ versus the number N of mobile users, where $\beta_1 = \beta_2 = \cdots = \beta_N$. As shown in Fig. 43(c), a larger N yields the higher sum effective capacity, which makes sense because of the following reasons. The existence of multiple mobile users reduces the probability that all mobile users experience deep fading during the same time period. Thus, the base station can take advantage of the better channel qualities among mobile users to improve the overall network throughput. Moreover, we can see from Fig. 43(c) that as the QoS exponent β_n gets larger, increasing N cannot effectively improve the sum effective capacity. This is because given larger QoS exponent, the BS needs to allocate more power to mobile users when their channel qualities are poorer. Then, the cellular system has to sacrifice power efficiency to compensate the deep channel fading, and thus increasing N does not improve the sum effective capacity significantly.

F. Summary

We derived the optimal channel-aware time-slot length and power allocation policy in cellular networks to maximize the sum effective capacity while satisfying the proportional-effective-capacity constraint and guaranteeing the diverse statistical QoS requirements from different mobile users. We also developed a suboptimal but simpler equal-length TD policy. Simulation results demonstrated the impact of QoS

provisionings on the resource allocation across different mobile users and the network performance, and showed that our derived optimal adaptation policy significantly outperforms the equal-length TD policy.

CHAPTER VIII

DELAY-QOS-AWARE BASE-STATION SELECTIONS FOR DISTRIBUTED MIMO LINKS IN BROADBAND WIRELESS NETWORKS

A. Introduction

In order to increase the coverage of broadband wireless networks, distributed multiple-input-multiple-output (MIMO) techniques, where multiple location-independent base stations (BS) cooperatively transmit data to mobile users, have attracted more and more research attentions [40–44]. In particular, the distributed MIMO techniques can effectively organize multiple location-independent BS's to form the distributed MIMO links connecting with mobile users, while not requiring too many multi-antennas, which are expensive, equipped at individual BS's. Like the conventional centralized MIMO system [45–47], the distributed MIMO system can significantly enhance the capability of the broadband wireless networks in terms of the quality-of-service (QoS) provisioning for wireless transmissions as compared to the single antenna system. However, the distributed nature for cooperative multi-BS transmissions also imposes many new challenges in wide-band wireless communications, which are not encountered in the centralized MIMO systems.

First, the cooperative distributed transmissions cause the severe difficulty for synchronization among multiple location-independent BS transmitters. Second, as the number of cooperative BS's increases, the computational complexity for MIMO signal processing and coding also grow rapidly. Third, because the coordinated BS's are located at different geographical positions, the cooperative communications in fact enlarges the interfering areas for the used spectrum, thus drastically degrading the frequency-reuse efficiency in the spatial domain. Finally, many wide-band trans-

missions are sensitive to the delay, and thus we need to design QoS-aware distributed MIMO techniques, such that the scarce wireless resources can be more efficiently utilized.

Towards the above issues, many research works on distributed MIMO transmissions have been proposed recently. The feasibility of transmit beamforming with efficient synchronization techniques over distributed MIMO link has been demonstrated through experimental tests and theoretical analyses [41, 44], suggesting that complicated MIMO signal processing techniques are promising to implement in realistic systems. While the antenna selection [46,47] is an effective approach to reduce the complexity for centralized MIMO systems, which can be also extended to distributed MIMO systems for the BS selection. It is clear that the BS-selection techniques can significantly decrease the processing complexity, while still achieving high throughput gain over the single BS transmission. Also, it is desirable to minimize the number of selected BS's through BS-selection techniques, which can effectively decrease the interfering range and thus improve the frequency-reuse efficiency of the entire wireless network. Most previous research works for BS selections mainly focused on the scenarios of selecting a subset of BS's/antennas with the fixed cardinality [42, 43]. However, it is evident that based on the wireless-channel status, BS-subset selections with dynamically adjusted cardinality can further decrease the BS usage. More importantly, how to efficiently support diverse delay-QoS requirements through BS-selection in distributed MIMO systems sill remains a widely cited open problem.

To overcome the aforementioned problems, we propose the QoS-aware BS-selection schemes for the distributed wireless MIMO links, which aim at minimizing the BS usages and reducing the interfering range, while satisfying diverse statistical delay-QoS constraints. In particular, based on the channel state information (CSI) and QoS requirements, the subset of BS with variable cardinality for the distributed

MIMO transmission is dynamically selected, where the selections are controlled by a central server. For the *single-user* scenario, we consider the optimization framework which uses the incremental BS-selection and time-sharing (IBS-TS) strategies, and study another framework which employs the ordered-gain based BS-selection and probabilistic-transmissions (OGBS-PT) techniques. For the *multi-user* scenario, we propose the optimization framework applying the priority BS-selection, block-diagonalization multiple-access, and probabilistic transmission (PBS-BD-PT) techniques. We derive the optimal transmission schemes for the above frameworks, respectively, and conduct comparative analyses with the baseline schemes through simulations.

The rest of this chapter is organized as follows. Section B describes the system model for distributed MIMO transmissions. Section D proposes the optimization framework for QoS-aware BS sections of the single-user case and develops its corresponding optimal solution. Section E develops the optimization framework for multi-user case and derives its optimal solution. Section F simulates our proposed schemes. The chapter concludes with Section G.

B. System Model

1. System Architecture

We concentrate on the wireless *distributed MIMO* system for downlink transmissions depicted in Fig. 44, which consists of K_{bs} distributed BS', K_{mu} mobile users, and one central server. The mth BS has M_m transmit antennas for $m = 1, 2, \ldots, K_{\text{bs}}$ and the nth mobile user has N_n receive antennas for $n = 1, 2, \ldots, K_{\text{mu}}$. All distributed BS's are connected to the central server through high-speed optical connections. The data to be delivered to the nth mobile user, $n = 1, 2, \ldots, K_{\text{mu}}$, arrives at the central server

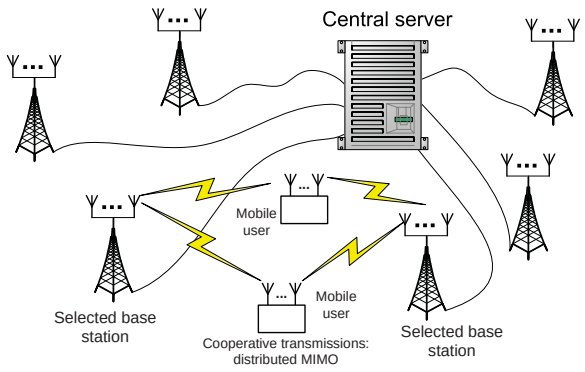


Fig. 44. System model of a wireless distributed MIMO system for downlink transmissions.

with a constant rate denoted by $\overline{C}_n$. Then, the central server dynamically controls these distributed BS's to cooperatively transmit data to the corresponding mobile users under the specified delay-QoS requirements.

For the case of $K_{\rm mu} = 1$, the distributed BS's and the mobile user form a single wireless MIMO link; when $K_{\rm mu} \geq 2$, the distributed BS's and the mobile users form the broadcast MIMO link for data transmissions. The wireless fading channels between the mth BS and the nth mobile user is modeled by an $N_n \times M_m$ matrix $\mathbf{H}_{n,m}$. The element at the ith row and jth column of $\mathbf{H}_{n,m}$, denoted by $(\mathbf{H}_{n,m})_{i,j}$, is the complex channel gain between the ith receive antenna of nth mobile user and the jth transmit antenna of the mth BS. All elements of $\mathbf{H}_{n,m}$ are independent and circularly symmetric complex Gaussian random variables with zero mean and the variance equal to $\overline{h}_{n,m}$. Also, the instantaneous aggregate power gain of the MIMO link between the nth mobile user and the mth BS, denoted by $\gamma_{n,m}$, is defined by

$$\gamma_{n,m} \triangleq \frac{1}{M_m} \sum_{i=1}^{N_n} \sum_{j=1}^{M_m} \left| (\mathbf{H}_{n,m})_{i,j} \right|^2 . \tag{8.1}$$

We define $\mathbf{H}_n \triangleq [\mathbf{H}_{n,1}\ \mathbf{H}_{n,2}\ \cdots\ \mathbf{H}_{n,K_{\rm bs}}]$ as the CSI for the nth mobile user for $n = 1, 2, \ldots, K_{\rm mu}$. The matrix $\mathbf{H}_n$ follows the independent block-fading model, where

$\mathbf{H}_n$ does not change within a time period with the fixed length T, called a time frame, but varies independently from one frame to the other frame. Furthermore, we define $\mathbf{H} \triangleq [\mathbf{H}_1^\tau \ \mathbf{H}_2^\tau \ \cdots \ \mathbf{H}_{K_{\rm bs}}^\tau]^\tau$, representing a fading state of the entire distributed MIMO system, where the superscript τ denotes the transpose operation on a matrix or a vector.

In order to decrease the complexity and suppress the interfering range of the distributed MIMO transmission, the cental server will dynamically select a subset of BS's to construct the distributed MIMO link. Accordingly, our design target is to minimize the average number of needed BS's subject to the specified QoS constraints. We suppose that each mobile user can perfectly estimate its CSI at the beginning of every time frame and reliably fed CSI back to the central server through dedicated control channels. Based on CSI $\mathbf{H}$ and QoS requirements, the central server then adaptively selects the subset of BS's and organizes them to transmit data to mobile users through the distributed MIMO links.

2. The Delay QoS Requirements

The central data server maintains a queue for the incoming traffic to each mobile user. We mainly focus on the queueing delay in this chapter because the wireless channel is the major bottleneck for high-rate wireless transmissions. Since it is usually unrealistic to guarantee the hard delay bound over the highly time-varying wireless channels, we employ the statistical metric, namely, the *delay-bound violation probability*, to characterize the diverse delay QoS requirements. Specifically, for the nth mobile user, the probability of violating a specified delay bound, denoted by $D_{\rm th}^{(n)}$, cannot exceed a given threshold ξ_n. That is, the inequality

$$\Pr\left\{D_n > D_{\rm th}^{(n)}\right\} \leq \xi_n \tag{8.2}$$

needs to hold for all $n = 1, 2, \ldots, N_{\text{mu}}$, where D_n denotes the queueing delay in the queue of nth mobile user's queueing system.

3. Performance Metrics and Design Objective

We denote by L the cardinality of the selected BS subset (the number of selected BS') for the distributed MIMO transmission in a fading state. Then, we denote the expectation of L by $\overline{L}$ and call it the *average BS usage*. As mentioned in Section B-1, our major objective is to minimize $\overline{L}$ through dynamic BS selection while guaranteeing the delay QoS constraint specified by Eq. (8.2). Besides the average BS usage, we also need to evaluate the *average interfering range* caused by the distributed MIMO transmission. The instantaneous interfering range, denoted by A, is defined as the area of the region where the average received power under the current MIMO transmission is larger than a certain threshold denoted by σ^2_{th}. The average interfering area is then defined as the expectation $\mathbb{E}\{A\}$ over all fading states. Clearly, minimizing $\overline{L}$ can not only reduce implementation complexity, but also decrease the average interfering range caused by the transmit power.

4. The Power Control Strategy

The transmit power of our distributed MIMO system varies with the number of selected BS'. In particular, given the number L of selected BS', the total instantaneous transmitted power used for distributed MIMO transmissions is set as a constant equal to $\mathcal{P}_L$. Furthermore, $\mathcal{P}_L$ linearly increases with L by using the strategy as follows:

$$\mathcal{P}_L = \mathcal{P}_{\text{ref}} + \kappa(L-1), \tag{8.3}$$

where $\mathcal{P}_{\text{ref}} > 0$ is called the *reference power* and $\kappa \geq 0$ describes the power increasing rate against L. Also, we define $\mathcal{P}_L \triangleq 0$ for $L = 0$. The above power adaptation

strategy is simple to implement, while the average transmit power can be effectively decreased through minimizing the average number of used BS'. In addition, Eq. (8.3) can upper-bound the instantaneous interferences and the interfering range over the entire network. Further note that when $\kappa = 0$, we get the scenario using the constant transmit power; when $\kappa = \mathcal{P}_{\text{ref}}$, the average transmit power is linearly proportional to the average BS usage. Thus, our framework of minimizing the average BS usage also leads to minimizing the average transmit power for the distributed MIMO system.

C. Statistical Delay-QoS Requirements and Guarantees

In this chapter, we apply the effective capacity approach [8, 19] to integrate the constraint on delay-bound violation probability given by Eq. (8.2) into our BS selection design.

As addressed in Chapter II, by using the QoS exponent θ, the delay-bound violation probability can be approximated [3, 8] by

$$\Pr\{D > D_{\text{th}}\} \approx e^{-\theta\varphi D_{\text{th}}}, \tag{8.4}$$

where D and D_{th} denote the queueing delay and delay bound, respectively, and φ is a constant determined by the arrival and departure processes. When the arrival rate is a constant equal to $\overline{C}$ and the departure rate is time-varying, Eq. (8.4) can be rewritten by

$$\Pr\{D > D_{\text{th}}\} \approx e^{-\theta\overline{C}D_{\text{th}}}. \tag{8.5}$$

Then, to upper-bound $\Pr\{D > D_{\text{th}}\}$ with a threshold ξ, using Eq. (8.5), we get the minimum required QoS exponent θ as follows:

$$\theta = -\frac{\log(\xi)}{\overline{C}D_{\text{th}}}. \tag{8.6}$$

Consider a discrete-time arrival process with constant rate $\overline{C}$ and a discrete-time time-varying departure process, denoted by $R[k]$, where k is the time index. In order to guarantee the desired θ determined by Eq. (8.6), the *effective capacity* $\mathcal{C}(\theta)$ of the service-rate process $R[k]$ needs to satisfy

$$\mathcal{C}(\theta) \geq \overline{C}, \tag{8.7}$$

under the given QoS exponent θ, as discussed in Chapter II.

In our distributed MIMO system, the BS selection result is designed as the function determined by the current CSI. Thus, the corresponding transmission rate (service rate) is time independent under the independent block-fading model (see Section B-1). Then, applying Eqs. (8.6)-(8.7), the delay QoS constraints given by Eq. (8.2) can be equivalently converted to:

$$\mathbb{E}_{\mathbf{H}}\left\{e^{-\theta_n R_n} - e^{-\theta_n \overline{C}_n}\right\} \leq 0 \tag{8.8}$$

for all $n = 1, 2, \ldots, N_{\text{mu}}$, where $\theta_n = -\log(\xi_n)\Big/\left(\overline{C}_n D_{\text{th}}^{(n)}\right)$ and $\mathbb{E}_{\mathbf{H}}\{\cdot\}$ denotes the expectation over all $\mathbf{H}$.

D. QoS-Aware BS Selection for the Single-User Case

We focus on the scenario with a single mobile user in this section, where $K_{\text{mu}} = 1$. For presentation convenience, we use the term *transmission mode* L to denote the case where the cardinality of the selected BS subset is equal to L. Given transmission mode L, we denote by Ω_L the set of indices of selected BS's, where $\Omega_L = \{i_{L,1}, i_{L,2}, \ldots, i_{L,L}\}$ and $i_{L,\ell} \in \{1, 2, \ldots, K_{\text{bs}}\}$ for $\ell = 1, 2, \ldots, L$. Once a BS is selected, we use all its transmit antennas for data transmissions. Then, we characterize the channel matrix

for the selected BS subset by $\mathbf{H}_{\Omega_L}$ and write $\mathbf{H}_{\Omega_L}$ as

$$\mathbf{H}_{\Omega_L} \triangleq \left[\mathbf{H}_{1,i_{L,1}}\ \mathbf{H}_{1,i_{L,2}}\ \cdots\ \mathbf{H}_{1,i_{L,L}}\right],$$

which is an $N_1 \times \mathcal{M}_L$ matrix with $\mathcal{M}_L \triangleq \sum_{\ell=1}^{L} M_{i_{L,\ell}}$. Accordingly, the physical-layer signal transmission is characterized by

$$\mathbf{y} = \mathbf{H}_{\Omega_L}\mathbf{s}_{\Omega_L} + \boldsymbol{\varsigma},$$

where $\mathbf{y}$ is the $N_1 \times 1$ received signal vector and $\boldsymbol{\varsigma}$ denotes the $N_1 \times 1$ additive Gaussian noise vector whose elements are independent with unit power. The variable $\mathbf{s}_{\Omega_L} \triangleq \left[\mathbf{s}_{i_{L,1}}^{\tau}, \mathbf{s}_{i_{L,2}}^{\tau}, \ldots, \mathbf{s}_{i_{L,L}}^{\tau}\right]^{\tau}$ is the input signal vector (transmitted signal vector) for the MIMO channel $\mathbf{H}_{\Omega_L}$, where $\mathbf{s}_{i_{L,\ell}}$ is the $M_{i_{L,\ell}} \times 1$ signal vector transmitted from the $(M_{i_{L,\ell}})$-th BS.

Clearly, for dynamic BS selections of distributed MIMO transmissions, we need to answer the following three questions: (i) For a specified transmission mode L, how do we determine the BS subset Ω_L? (ii) How are the wireless resources shared if applying multiple modes within a time frame? (iii) Which transmission modes will be used and how to quantitatively allocate the wireless resources? We first study associated issues for question (i) in Section D-1. Then, we introduce the time-sharing transmission and probabilistic transmission to share the resources across different modes in Section D-2. Following the discussions in Sections D-1 and D-2, we formulate two mathematical optimization frameworks to answer question (iii). One optimization framework is based on the incremental BS-selection algorithm and the time-sharing strategy; the other framework applies the ordered-gain based BS selection algorithm with probabilistic transmission, which will be detailed in Sections D-3 and D-4, respectively. The BS-selection scheme derived under the former framework can achieve better performance, while the latter one is simpler to implement.

1. BS Selection Strategy Given the Cardinality of the BS Subset

In this chapter, we focus on the spatial multiplexing based MIMO transmissions. Given Ω_L in a fading state, the maximum achievable data rates (Shannon capacity), denoted by $R(\Omega_L)$ (nats/frame), are determined [45] by

$$R(\Omega_L) = \max_{\mathbf{\Xi}:\mathrm{Tr}(\mathbf{\Xi})=\mathcal{P}_L} \left\{ BT \log \left[\det \left(\mathbf{I} + \mathbf{H}_{\Omega_L} \mathbf{\Xi} \mathbf{H}_{\Omega_L}^{\dagger} \right) \right] \right\}, \tag{8.9}$$

where $(\cdot)^{\dagger}$ represents the conjugate transpose, $\det(\cdot)$ generates the determinant of a matrix, $\mathrm{Tr}(\cdot)$ evaluates the trace of a matrix, and $\mathbf{\Xi}$ is the covariance matrix of $\mathbf{s}_{\Omega_L}$.

Before further proceeding, we first summarize the results of MIMO transmission [45] to describe how to attain the maximum rate given by Eq. (8.9). In particular, we apply the singular value decomposition (SVD) [45] on $\mathbf{H}_{\Omega_L}$ and get $\mathbf{H}_{\Omega_L} = \mathbf{U}_L \mathbf{\Lambda}_L \mathbf{V}_L^{\dagger}$, where $\mathbf{U}_L$ and $\mathbf{V}_L$ are unitary matrices, $\mathbf{\Lambda}_L$ is an $N_1 \times \mathcal{M}_L$ rectangular diagonal matrix with only nonnegative elements. More specifically, $\mathbf{U}_L$ and $\mathbf{V}_L$ consist of N_1 left singular vectors and $\mathcal{M}_L$ right singular vectors of $\mathbf{H}_{\Omega_L}$, respectively. The zth diagonal element of $\mathbf{\Lambda}_L$, denoted by $\sqrt{\epsilon_{L,j}}$, is equal to the zth largest singular value of $\mathbf{H}_{\Omega_L}$, where $z = 1, 2, \ldots, \min\{N_1, \mathcal{M}_L\}$.

To achieve the maximum data transmission rate, the transmitted signal needs to be set as $\mathbf{s} = \mathbf{V}_{L,1}\mathbf{x}$, where $\mathbf{V}_{L,1}$ is a precoding matrix consisting of the first Z_L columns of $\mathbf{V}_L$ and $\mathbf{x}_L \triangleq [x_{L,1}, x_{L,2}, \ldots, x_{L,Z_L}]^{\tau}$ is a signal vector with independent elements. Note that $Z_L = \mathrm{rank}(\mathbf{H}_{\Omega_L})$, which is also the number of nonzero singular values of $\mathbf{H}_{\Omega_L}$. Applying the precoding matrix $\mathbf{V}_L$ for the transmitted signals, we can convert the MIMO channel to Z_L parallel Gaussian sub-channels, where the zth sub-channel's SNR is equal to $\varepsilon_{L,j}$. The optimal power used on the zth sub-channel is equal to $\rho_{L,j} = [\mu_L - 1/\varepsilon_{L,j}]^+$, which is known as the water-filling algorithm, where $[\cdot]^+ \triangleq \max\{\cdot, 0\}$ and μ_L is the water level selected such that $\sum_{j=1}^{Z_L} \rho_{L,j} = \mathcal{P}_L$.

Correspondingly, we can get the maximum achievable data rate, which is given by $BT\sum_{j=1}^{Z_L}\left[\log(\mu_L\varepsilon_{L,j})\right]^+$.

Under the aforementioned allocation, the transmit power allocated to the (i_ℓ)-th BS in transmission mode L, denoted by ϱ_{L,i_ℓ}, is determined by the following equations:

$$\begin{cases} \boldsymbol{\rho}_L & \triangleq & (\rho_{L,1},\rho_{L,2},\ldots,\rho_{L,Z_L})^\tau \\ \widehat{\boldsymbol{\rho}}_L & = & \Big(\mathbf{V}_{L,1}\circ \mathrm{conj}\left(\mathbf{V}_{L,1}\right)\Big)\boldsymbol{\rho}_L; \\ \varrho_{L,i_\ell} & = & \sum_{w=W_{L,\ell-1}}^{W_{L,\ell}}\widehat{\boldsymbol{\rho}}_L(w), \end{cases} \tag{8.10}$$

where $(\cdot\circ\cdot)$ denotes the element-wise product between two matrices, $\mathrm{conj}(\cdot)$ represents the element-wise conjugate of a matrix or a vector, and $W_{L,\ell}\triangleq\sum_{j=1}^{\ell}M_{i_{L,j}}$. In Eq. (8.10), $\widehat{\boldsymbol{\rho}}_L$ is the power vector associated with the input signal vector for the MIMO channel $\mathbf{H}_{\Omega_L}$, and $\widehat{\boldsymbol{\rho}}_L(w)$ represents the wth element of $\widehat{\boldsymbol{\rho}}_L$.

After obtaining $R(\Omega_L)$, we have the best selection strategy to optimize the achievable rate as follows:

$$\max_{\Omega_L}\left\{R(\Omega_L)\right\}. \tag{8.11}$$

To derive the optimal solution for this optimization problem, it is clear that we need to examine all $\binom{K_{\mathrm{bs}}}{L}$ possible BS combinations, which leads to the prohibitively high computational complexity as M gets large. Alternatively, we consider two suboptimal approaches with low complexities as follows.

1). Incremental BS-Selection Algorithm: In [46], the authors developed the fast antenna selection algorithm using the incremental-selection strategy. Although this incremental-selection strategy was developed for antenna selection without CSI feedback, it can be readily extended to the scenario for BS selection with CSI feedback to achieve the near optimal data rate. The pseudo codes of the incremental BS-selection algorithm are given in Table VIII. In particular, the idea of this algorithm

Table VIII. The Pseudo Codes to Determine Ω_L by Using the Incremental BS-Selection Algorithm.

00.	Let $\Psi := \{1, 2, \ldots, K_{\rm bs}\}$ and $\overline{\Psi} := \varnothing$, and $Z = \|\Psi\|$, where $\varnothing$ is the empty set and $\|\Psi\|$ denotes the cardinality of the set Ψ; ! Use variables $\overline{\Psi}$ and Ψ to store all selected BS's and all other BS', respectively.
01.	For $i := 1$ to L ! Add one BS to $\overline{\Psi}$ in each step.
02.	For $z := 1$ to Z ! Examine Z BS's in Ψ, respectively
03.	$\Theta_z := \overline{\Psi} \cup \{\psi_z\}$, where ψ_z is the zth element in Ψ; ! Pick a BS in Ψ to form a new subset Θ_z with $\overline{\Psi}$.
04.	$\widetilde{R}_z := R(\Theta_z)$ based on Eq. (8.9) by setting $\Omega_L := \Theta_z$. ! Examine the achievable rate of Θ_z.
05.	End
06.	$z^* := \arg \max_{1 \le z \le Z} \left\{ \widetilde{R}_z \right\}$; ! Select the BS to maximize the achievable rate.
07.	$\overline{\Psi} := \Theta_{z^*}$, $\Psi := \Psi \backslash \{\psi_{z^*}\}$, and $Z := \|\Psi\|$; ! Add the newly selected BS into the BS subset $\overline{\Psi}$.
08.	End
09.	$\Omega_L := \overline{\Psi}$. ! Complete the BS selection and get Ω_L.

is to determine Ω_L through L steps, where in each step one BS is selected, as shown in lines 01-08 of Table VIII. In each step, one selected BS is added to the BS subset denoted by $\overline{\Psi}$, where the selection criterion is to maximize the achievable rate of the updated BS subset $\overline{\Psi}$. Then, after L steps, we have totally added L BS's into $\overline{\Psi}$ and then assign $\Omega_L := \overline{\Psi}$. This algorithm only examines the achievable rates for $L(K_{\rm bs} - (L-1)/2)$ different BS combinations, which requires $L(K_{\rm bs} - (L-1)/2)$ times of SVD, resulting in much less complexity than the optimal approach which examines all $\binom{K_{\rm bs}}{L}$ BS combinations.

2). Ordered-Gain Based BS-Selection Algorithm: The ordered-gain (or ordered-SNR) based BS-selection algorithm selects L BS's with the largest aggregate power gain over all BS's, where the aggregate power gain is defined by Eq. (8.1). Since maximizing the aggregate power gain may not effectively optimize the achievable transmission rate for MIMO links, the incremental BS-selection algorithm usually dominates the ordered-gain based BS-selection algorithm. However, since the ordered-gain based BS-selection algorithm does not need to perform the SVD, its complexity is much lower than that of the incremental BS-selection algorithm.

Figure 45 plots the cumulative distribution functions (CDF) of the achievable

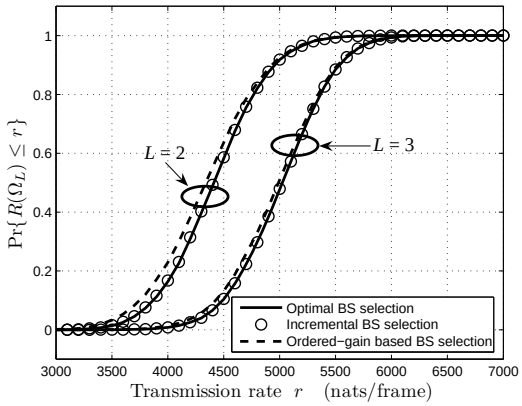


Fig. 45. CDF of the achievable transmission rates $R(\Omega_L)$: $\Pr\{R(\Omega_L) \leq r\}$ versus r, where $K_{\rm bs} = 6$, $M_m = 2$ for all $m = 1, 2, \ldots, K_{\rm bs}$, $N_1 = 2$, $B = 10^5$ Hz, $T = 10$ ms, and $\overline{h}_m = 4$ dB for all $m = 1, 2, \ldots, K_{\rm bs}$.

transmission rate obtained by using the incremental selection algorithm, the optimal selection, and the ordered-gain based BS selection, respectively. As shown in Fig. 45, the incremental selection algorithm and the ordered-gain based BS selection both can achieve the near optimal performance.

2. Time Sharing and Probabilistic Transmissions

To get the more general framework for BS selection, we apply the time sharing and probabilistic transmission strategies, respectively, over different transmission modes, which are described as follows.

1). Time Sharing Transmissions: Each time frame can be divided into $(K_{\rm bs}+1)$ time slots with the lengths equal to $\{T\alpha_m\}_{m=0}^{K_{\rm bs}}$, where α_m is the normalized time-slot length and $\sum_{m=0}^{K_{\rm bs}} \alpha_m = 1$. Within the mth time slot for $m > 1$, the transmission mode L BS will be used; for $m = 0$, no data is transmitted in the corresponding time slot. Then, the total service rate in a time frame is equal to $\sum_{L=0}^{K_{\rm bs}} \alpha_L R(\Omega_L)$, where $R(\Omega_L)$ is given by Eq. (8.9) for $L \neq 0$ and $R(\Omega_L) = 0$ for $L = 0$. Furthermore, the total BS usage is given by $\sum_{L=0}^{K_{\rm bs}} L\alpha_L$. The purpose of applying the time sharing

based transmissions is to increase the system flexibility and to gain the continuous control on the BS usage within each time frame. Accordingly, we need to identify how to optimally adjust $\boldsymbol{\alpha}$ with CSI and QoS constraints, where $\boldsymbol{\alpha} \triangleq (\alpha_0, \alpha_1, \ldots, \alpha_M)$.

2). Probabilistic Transmissions: Under this strategy, within each time frame only one transmission mode will be used for distributed MIMO transmissions. In particular, we will select transmission mode L with probability equal to ϕ_L and define $\boldsymbol{\phi} \triangleq (\phi_0, \phi_1, \phi_2, \ldots, \phi_{K_{\rm bs}})^\tau$. Then, our target is to determine how to dynamically adjust $\boldsymbol{\phi}$ according to the CSI and QoS requirements.

If setting $\boldsymbol{\phi} = \boldsymbol{\alpha}$ and using the same strategy to determine Ω_L over all fading states, we obtain the same BS usage. However, the effective capacities achieved under the time-sharing transmission and the probabilistic transmission, denoted by $\mathcal{C}_{\rm TS}(\boldsymbol{\alpha}, \theta_1)$ and $\mathcal{C}_{\rm PR}(\boldsymbol{\phi}, \theta_1)$, respectively, are different. Specifically, we derive

$$\begin{aligned}\mathcal{C}_{\rm TS}(\boldsymbol{\alpha}, \theta_1) &= -\frac{1}{\theta_1}\log\left(\mathbb{E}_{\mathbf{H}}\left\{e^{-\sum_{L=0}^{K_{\rm bs}}\alpha_L R(\Omega_L)}\right\}\right) \geq -\frac{1}{\theta_1}\log\left(\mathbb{E}_{\mathbf{H}}\left\{\alpha_L e^{-\sum_{L=0}^{K_{\rm bs}} R(\Omega_L)}\right\}\right) \\ &= -\frac{1}{\theta_1}\log\left(\mathbb{E}_{\mathbf{H}}\left\{\phi_L e^{-\sum_{L=0}^{K_{\rm bs}} R(\Omega_L)}\right\}\right) = \mathcal{C}_{\rm PR}(\boldsymbol{\phi}, \theta_1), \end{aligned} \tag{8.12}$$

where the inequality holds because $\mathbb{E}_{\mathbf{H}}\left\{e^{-\sum_{L=0}^{K_{\rm bs}}\alpha_L R(\Omega_L)}\right\}$ is a convex function over $\big(R(\Omega_0), R(\Omega_1), \ldots, R(\Omega_{K_{\rm bs}})\big)$. Equation (8.12) suggests that the time-sharing transmission generally outperforms the probabilistic transmission. However, the probabilistic transmission is more realistic to implement than the time-sharing transmission due to the following reasons. On the one hand, for the optimized time-sharing transmission, the time-slot length $T\alpha_L$ may be quite small and thus very hard to implement. On the other hand, the multiple time slots (for the time-sharing transmission) within a time frame introduces more overhead than the single-slot case (for the probabilistic transmission).

3. Optimization Framework Using Time-Sharing Transmissions with Incremental BS Selection

As discussed in Section B-3, our major objective is to minimize the average BS usage. In this section, we focus on the framework which employs the incremental BS-selection algorithm for each transmission mode and apply the time-sharing transmission for different transmission modes. Then, we develop the efficient BS-selection scheme under the above framework by solving the following optimization problem **VIII-*A*1**, which aims at minimizing the average BS usage while guaranteeing the delay-QoS requirement.

$$\textbf{VIII-}\boldsymbol{A}\textbf{1}: \quad \min_{\boldsymbol{\alpha}} \left\{\overline{L}\right\} = \min_{\boldsymbol{\alpha}} \left\{ \mathbb{E}_{\mathbf{H}} \left\{ \sum_{L=0}^{K_{\rm bs}} \alpha_L L \right\} \right\}$$

$$\text{s.t.:} \quad 1). \ \sum_{L=0}^{K_{\rm bs}} \alpha_L = 1, \quad \forall\, \mathbf{H}; \tag{8.13}$$

$$2). \ \mathbb{E}_{\mathbf{H}} \left\{ e^{-\theta_1 \sum_{L=0}^{K_{\rm bs}} \alpha_L R(\Omega)_L} - e^{-\theta_1 \overline{C}_1} \right\} \leq 0, \tag{8.14}$$

where $\boldsymbol{\alpha}$ is a function of $\mathbf{H}$, Eq. (8.14) is the constraint to guarantee the delay-bound violation probability as derived in Eq. (8.2), we obtain Ω_L through the incremental selection algorithm listed in Table VIII, and determine $R\boldsymbol{\alpha}$ based on Eq. (8.9). We call the solution to **VIII-*A*1** as the *incremental BS-selection and time-sharing based* (IBS-TS) scheme.

Note that the optimization over $\boldsymbol{\alpha}$ is an $(K_{\rm bs}+1)$-dimensional problem. To reduce the dimension of optimization variables, we define

$$\widetilde{R}(\mathcal{L}) \ \triangleq \ \max_{\boldsymbol{\alpha}} \left\{ \sum_{L=0}^{K_{\rm bs}} \alpha_L R(\Omega_L) \right\} \tag{8.15}$$

$$\text{s.t.} \quad \sum_{L=0}^{K_{\rm bs}} \alpha_L = 1; \ \sum_{L=0}^{K_{\rm bs}} \alpha_L L = \mathcal{L}. \tag{8.16}$$

Based on Eqs. (8.15)-(8.16), $\widetilde{R}(\mathcal{L})$ is the maximum achievable rate over all $\boldsymbol{\alpha}$ with the same BS usage. Since $\widetilde{R}(\mathcal{L})$ is the convex combination [79] over $\{R(\Omega_L)\}_{L=1}^{K_{\rm bs}}$, then the definition in Eqs. (8.15)-(8.16) suggests that $\widetilde{R}(\mathcal{L})$ is a piece-wise linear and concave function, which can be written as

$$\widetilde{R}(\mathcal{L}) = \begin{cases} \widetilde{R}\left(m_{j-1}\right) + \nu_j \left(\mathcal{L} - m_{j-1}\right), & \text{if } \mathcal{L} \in (m_{j-1}, m_j],\ j = 1, 2, \ldots, \mathcal{K}, \\ 0, & \text{if } \mathcal{L} = 0; \end{cases} \tag{8.17}$$

for a certain integer $\mathcal{K}$, where $m_0 < m_1 < \cdots < m_{\mathcal{K}}$, $m_0 = 0$, $m_{\mathcal{K}} = K_{\rm bs}$, and $m_i \in \{0, 1, \ldots, K_{\rm bs}\}$. Moreover, using $\left(m_i, \widetilde{R}(m_i)\right)$ to represent the coordinates of a point in the two-dimensional plane, we can identify $\left\{\left(m_i, \widetilde{R}(m_i)\right)\right\}_{i=1}^{\mathcal{K}-1}$ through the following procedures: a). find vertices of the convex hull spanned by two-dimensional points $\left\{\left(L, \widetilde{R}(L)\right)\right\}_{L=0}^{K_{\rm bs}}$; b). $\left\{\left(m_i, \widetilde{R}(m_i)\right)\right\}_{i=1}^{\mathcal{K}-1}$ are located above the line segment with end points $(0, 0)$ and $\left(K_{\rm bs}, \widetilde{R}(K_{\rm bs})\right)$. Accordingly, ν_j is the slope of the line segment starting at the point $\left(m_i, \widetilde{R}(m_i)\right)$ and ending at the point $\left(m_{i-1}, \widetilde{R}(m_{i-1})\right)$, which is determined by

$$\nu_i = \frac{\widetilde{R}(m_i) - \widetilde{R}(m_{i-1})}{m_i - m_{i-1}}, \qquad i = 1, 2, \ldots, \mathcal{K}. \tag{8.18}$$

For presentation convenience, we also define $\nu_0 \triangleq \infty$ and $\nu_{\mathcal{K}+1} \triangleq -\infty$. Furthermore, given $\mathcal{L} \in [m_{j-1}, m_j]$, following the piece-wise linear property, we derive the corresponding $\boldsymbol{\alpha}$ to achieve the service rate $\widetilde{R}(\mathcal{L})$ as follows:

$$\alpha_L = \begin{cases} \frac{m_j - \mathcal{L}}{m_j - m_{j-1}}, & \text{if } L = m_{j-1}; \\ \frac{\mathcal{L} - m_{j-1}}{m_j - m_{j-1}}, & \text{if } L = m_j; \\ 0, & \text{otherwise.} \end{cases} \tag{8.19}$$

Applying Eqs. (8.15), (8.16), and (8.19) into problem **VIII-*A*1**, we can equiva-

lently convert **VIII-$\boldsymbol{A1}$** to the following problem **VIII-$\boldsymbol{A1}'$**:

$$\begin{aligned} \textbf{VIII-}\boldsymbol{A1}'\text{:} \quad & \min_{\mathcal{L}} \left\{ \mathbb{E}_{\mathbf{H}} \left\{ \mathcal{L} \right\} \right\} \\ & \text{s.t.:} \quad \mathbb{E}_{\mathbf{H}} \left\{ e^{-\theta_1 \widetilde{R}(\mathcal{L})} - e^{-\theta_1 \overline{C}_1} \right\} \leq 0, \end{aligned} \tag{8.20}$$

where $\mathcal{L}$ is a function of $\mathbf{H}$ and we can uniquely map $\mathcal{L}$ to $\boldsymbol{\alpha}$ through Eq. (8.19). Since $\widetilde{R}(\mathcal{L})$ is an increasing and concave function, $e^{-\theta_1 \widetilde{R}(\mathcal{L})}$ is convex over $\mathcal{L}$ [90, pp. 84]. Thus, we can see that **VIII-$\boldsymbol{A1}'$** satisfies: a) the objective function is convex; b) the inequality constraint function $\mathbb{E}_{\mathbf{H}} \left\{ e^{-\theta_1 \widetilde{R}(\mathcal{L})} - e^{-\theta_1 \overline{C}_1} \right\}$ is convex. Therefore, **VIII-$\boldsymbol{A1}'$** is a convex problem [90, pp. 137]. Then, using the Lagrangian method, we solve for the optimal solution of **VIII-$\boldsymbol{A1}'$**, as summarized in the following Theorem 13.

Theorem 13. *The optimal solution to* **VIII-$\boldsymbol{A1}'$**, *if existing, is determined by*

$$\mathcal{L}^* = \begin{cases} m_j, & \text{if } \nu_{j+1} \leq \dfrac{e^{\theta_1 \widetilde{R}(m_j)}}{\theta_1 \lambda^*} \leq \nu_j; \\ \dfrac{\log(\theta_1 \lambda^* \nu_j)}{\nu_j \theta_1} - \dfrac{\widetilde{R}(m_{j-1})}{\nu_j} + m_{j-1}, & \text{if } \widetilde{R}(m_{j-1}) < \dfrac{\log(\theta_1 \lambda^* \nu_j)}{\theta_1} < \widetilde{R}(m_j), \end{cases} \tag{8.21}$$

where $\widetilde{R}(\cdot)$ and ν_j are characterized by Eqs. (8.15) through (8.18). In Eq. (8.21), $\lambda^* \geq 0$ is a constant over all fading states, which needs to be selected such that equality in Eq. (8.14) holds.

Proof. The proof of Theorem 13 is provided in Appendix Q. □

Remarks: (i) Having obtained the optimal $\mathcal{L}^*$ for problem **VIII-$\boldsymbol{A1}'$**, the optimal solution to **VIII-$\boldsymbol{A1}$**, denoted by $\boldsymbol{\alpha}^*$, is obtained by setting $\mathcal{L} = \mathcal{L}^*$ in Eq. (8.19). (ii) Under the optimal solution, we do not allocate time slots for all transmission modes. As indicated by Eq. (8.19), within any time frames, we employ at most two transmission modes. (iii) It is clear that by setting $\alpha_{K_{\text{bs}}} = 1$ for all time frames, we use the maximum transmit power and thus obtain the maximum achievable effective

capacity, which is denoted by $\mathcal{C}_{\max}^{(1)}$. If $\mathcal{C}_{\max}^{(1)}$ is still smaller than $\overline{C}_1$, the specified delay-QoS requirement cannot be satisfied since we have used up all power budget. As a result, the optimal solution does not exist for this case. In contrast, if $\mathcal{C}_{\max}^{(1)} \geq \overline{C}_1$, we can always find the optimal solution. (iv) The analytical expressions for λ^* is usually intractable. However, it is easy to verify that $\mathbb{E}_{\mathbf{H}}\left\{e^{-\theta_1 \tilde{R}(\mathcal{L}^*)} - e^{-\theta_1 \overline{C}_1}\right\}$ is an increasing function of λ^*. Thus, we can readily obtain λ^* by using the numerical searching techniques.

4. Optimization Framework Using Probabilistic Transmissions with Ordered-Gain Based BS Selection

We in this section consider the framework using the ordered-gain based BS-selection algorithm and the probabilistic transmission strategy. Specifically, we formulate the optimization problem for this framework as follows:

$$\begin{aligned}
\textbf{VIII-}\boldsymbol{A2}: \quad & \min_{\boldsymbol{\phi}} \left\{\overline{L}\right\} = \min_{\boldsymbol{\phi}} \left\{\mathbb{E}_{\mathbf{H}}\left\{\sum_{L=0}^{K_{\rm bs}} \phi_L L\right\}\right\} \\
& \text{s.t.:} \quad 1). \ \sum_{L=0}^{K_{\rm bs}} \phi_L = 1, \quad \forall\, \mathbf{H}; \qquad (8.22) \\
& \qquad\quad 2). \ \mathbb{E}_{\mathbf{H}}\left\{\left(\sum_{L=0}^{K_{\rm bs}} \phi_L e^{-\theta_1 R(\Omega_L)}\right) - e^{-\theta_1 \overline{C}_1}\right\} \leq 0, \qquad (8.23)
\end{aligned}$$

where $\boldsymbol{\phi}$ is a function of $\mathbf{H}$. We call the optimal solution to the optimization problem **VIII-*A2*** as the ordered-gain and probability transmission based (OGBS-PT) scheme.

Theorem 14. *The optimal solution to problem* **VIII-*A2***, *denoted by* $\boldsymbol{\phi}^*$ *is given by*

$$\phi_L^* = \begin{cases} 1, & \textit{if } L = L^*; \\ 0, & \textit{if } L \neq L^*, \end{cases} \qquad (8.24)$$

for all $\mathbf{H}$*, where*

$$L^* = \arg\min_{L} \left\{ L + \lambda^* e^{-\theta_1 R(\Omega_L)} \right\}. \tag{8.25}$$

In Eq. (8.25), $\lambda^* \geq 0$ *is a constant over all* $\mathbf{H}$*, which is selected such that the equality holds for Eq. (8.23) to guarantee the delay-QoS requirement.*

Proof. The proof of Theorem 14 is provided in Appendix R. □

Remarks: (i) Based on discussions in Sections D-1 and D-2, the OGBS-PT scheme is easier to implement than the IBS-TS scheme. Moreover, although the IBS-TS scheme generally outperforms the OGBS-PT scheme, we will see in Section F that the performance differences between them are little. Therefore, the IBS-TS is more promising to the realistic distributed MIMO systems. (ii) Theorem 14 suggests that under the optimal solution, the probabilistic transmission reduces to a deterministic strategy, where the only transmission mode L^* will be used for data transmission. (iii) Similar to problem **VIII-*A*1**, the parameter λ^* for **VIII-*A*2** also needs to be tracked through numerical searching.

5. Base-Station Selection with Fixed Cardinality

We also study the BS-selection strategy which fixes the cardinality L and the selected BS subsect Ω_L over all fading states. In particular, the mobile user selects L BS's which are closest to itself. Clearly, these L BS' are associated with the L largest average channel gains over all BS's, which can be conveniently measured by the mobile user. This approach is served as a baseline scheme for comparative analyses with our proposed dynamic BS selection schemes. In particular, the fixed BS-subset

cardinality, denoted by L_{fix}, is determined by

$$\begin{aligned} L_{\text{fix}} &= \arg \min_{L=1,2,\ldots,K_{\text{bs}}} \left\{ \mathbb{E}_{\mathbf{H}} \left\{ e^{-\theta_1 R(\Omega_L)} \right\} \right\} \\ &\quad \text{s.t.: } \mathbb{E}_{\mathbf{H}} \left\{ e^{-\theta_1 R(\Omega_L)} \right\} \leq e^{-\theta_1 \overline{C}_1}, \end{aligned} \tag{8.26}$$

where Ω_L consists of BS's with L largest average channel gain over all K_{bs} BS's. We call the above described strategy the *fixed* BS-selection scheme.

E. QoS-Aware BS Selection for the Multi-User Case

We next consider the distributed MIMO transmissions for the case with *multiple mobile users*. Clearly, it is more challenging for BS selection in the multi-user case as compared to the single-user case. On one hand, different users[1] may select the same BS's and thus we need to apply the multiple access techniques to avoid the interference and to develop the resource allocation strategies across multiple links. On the other hand, even different users select different BS's, the cross interferences among coexisting links may significantly degrade the throughput and complicate the BS selection algorithms.

For efficient BS selection and distributed MIMO transmissions, the central sever controls the selected distributed BS's and the mobile users to constitute the broadcast MIMO link, as mentioned in Section B. Specifically, given the transmission mode[2] L and BS-index subset $\Omega_L = \{i_{L,1}, i_{L,2}, \ldots, i_{L,L}\}$ of the selected BS's, the channel matrix of the nth mobile user, modeled by $\mathbf{H}_{\Omega_L}^{(n)}$, is determined by

$$\mathbf{H}_{\Omega_L}^{(n)} \triangleq \left[\mathbf{H}_{n,i_{L,1}} \; \mathbf{H}_{n,i_{L,2}} \; \cdots \; \mathbf{H}_{n,i_{L,L}} \right], \quad n = 1, 2, \ldots, K_{\text{mu}}$$

[1]We use the terms of "mobile user" and "user" exchangeably in the rest of this chapter.

[2]For the multi-user case, we also use the term of transmission mode L to denote the case where the BS-subset's cardinality is L.

where $\mathbf{H}_{\Omega_L}^{(n)}$ is an $N_n \times \left(\sum_{\ell=1}^{L} M_{i_{L,\ell}}\right)$ matrix. Then, the physical-layer signal transmissions can be characterized by

$$\mathbf{y}_{\Omega_L}^{(n)} = \mathbf{H}_{\Omega_L}^{(n)} \sum_{i=1}^{K_{\text{mu}}} \mathbf{s}_{\Omega_L}^{(i)} + \boldsymbol{\varsigma}^{(n)}, \quad n = 1, 2, \ldots, K_{\text{mu}}$$

where $\mathbf{s}_{\Omega_L}^{(i)}$ represents the ith user's input signal vector for the MIMO channel $\mathbf{H}_{\Omega_L}^{(n)}$, $\mathbf{y}_{\Omega_L}^{(n)}$ is the signal vector received at the nth user's receive antennas, and $\boldsymbol{\varsigma}^{(n)}$ is the complex additive white Gaussian noise (AWGN) vector with unit power for each element of this vector.

It is well-known that the optimal capacity of the broadcast MIMO link can be achieved through the dirty-paper coding techniques [91], which is, however, hard to implement due to its high complexity [92]. Alternatively, we apply the block-diagonalization precoding techniques [92] for distributed MIMO transmissions and then concentrate on developing efficient QoS-aware BS-selection scheme. In particular, we introduce the block-diagonalization precoding in Section E-1. In Section E-2, we develop the priority BS-selection scheme under the specified cardinality of the BS subset. Applying the above techniques, in Section E-3 we formulate the QoS-aware optimization problem to develop the joint BS-selection and resource allocation scheme, which aims at minimizing the average BS usage. We also discuss the time-division-multiple-access (TDMA) based BS selections in Section E-4 as the baseline scheme for comparative analyses.

1. The Block Diagonalization Technique for Distributed MIMO Transmissions

The idea of block diagonalization [92] is to use a precoding matrix, denoted by $\mathbf{\Gamma}_{\Omega_L}^{(n)}$, for the nth user's transmitted signal vector, such that $\mathbf{H}_{\Omega_L}^{(i)} \mathbf{\Gamma}_{\Omega_L}^{(n)} = \mathbf{0}$ for all $i \neq n$. By setting $\mathbf{s}_{\Omega_L}^{(n)} = \mathbf{\Gamma}_{\Omega_L}^{(n)} \widehat{\mathbf{s}}_{\Omega_L}^{(n)}$, where $\widehat{\mathbf{s}}_{\Omega_L}^{(n)}$ is the nth user's data vector to be precoded by $\mathbf{\Gamma}_{\Omega_L}^{(n)}$,

we can rewrite the received signal $\mathbf{y}_{\Omega_L}^{(n)}$ as

$$\mathbf{y}^{(n)} \;=\; \mathbf{H}_{\Omega_L}^{(n)} \sum_{i=1}^{K_{\mathrm{mu}}} \mathbf{\Gamma}_{\Omega_L}^{(i)} \widehat{\mathbf{s}}_{\Omega_L}^{(i)} + \boldsymbol{\varsigma}^{(n)} = \mathbf{H}_{\Omega_L}^{(n)} \mathbf{\Gamma}_{\Omega_L}^{(n)} \widehat{\mathbf{s}}_{\Omega_L}^{(n)} + \boldsymbol{\varsigma}^{(n)} \;=\; \widehat{\mathbf{\Gamma}}_{\Omega_L}^{(n)} \widehat{\mathbf{s}}_{\Omega_L}^{(n)} + \boldsymbol{\varsigma}^{(n)}, \quad (8.27)$$

where $\widehat{\mathbf{\Gamma}}_{\Omega_L}^{(n)} \triangleq \mathbf{H}_{\Omega_L}^{(n)} \mathbf{\Gamma}_{\Omega_L}^{(n)}$. Under this strategy, the nth user's signal will not cause interferences to other users. Accordingly, the MIMO broadcast transmissions are virtually converted to K_{mu} orthogonal MIMO channels with channel matrices $\left\{\widehat{\mathbf{\Gamma}}_{\Omega_L}^{(n)}\right\}_{n=1}^{K_{\mathrm{mu}}}$. Then, we can get the nth user's maximum achievable rate in a fading state, denoted by $R^{(n)}\big(\Omega_L, \mathcal{P}_L^{(n)}\big)$, as follows:

$$R^{(n)}\big(\Omega_L, \mathcal{P}_L^{(n)}\big) \triangleq \max_{\mathbf{\Xi}^{(n)}:\mathrm{Tr}\left(\mathbf{\Xi}^{(n)}\right)=\mathcal{P}_L^{(n)}} \left\{ BT \log \left[\det \left(\mathbf{I} + \widehat{\mathbf{\Gamma}}_{\Omega_L}^{(n)} \mathbf{\Xi}^{(n)} \left(\widehat{\mathbf{\Gamma}}_{\Omega_L}^{(n)} \right)^{\dagger} \right) \right] \right\}, \quad (8.28)$$

where $\mathbf{\Xi}^{(n)}$ is the covariance matrix of $\widehat{\mathbf{s}}_{\Omega_L}^{(n)}$ and $\mathcal{P}_L^{(n)}$ is the power allocated for the nth user.

The precoding matrix $\mathbf{\Gamma}_{\Omega_L}^{(n)}$ for $n = 1, 2, \ldots, K_{\mathrm{mu}}$, can be determined by the following procedures [92]. Let us first define the following matrix:

$$\widehat{\mathbf{H}}_{\Omega_L}^{(n)} \triangleq \left[\left(\mathbf{H}_{\Omega_L}^{(1)} \right)^{\tau} \;\cdots\; \left(\mathbf{H}_{\Omega_L}^{(n-1)} \right)^{\tau} \; \left(\mathbf{H}_{\Omega_L}^{(n+1)} \right)^{\tau} \;\cdots\; \left(\mathbf{H}_{\Omega_L}^{(K_{\mathrm{mu}})} \right)^{\tau} \right]^{\tau}, \quad (8.29)$$

which is a $\left(\sum_{n=1}^{K_{\mathrm{mu}}} N_n\right) \times \left(\sum_{\ell=1}^{L} M_{i_{L,\ell}}\right)$ matrix, representing the CSI from all BS's except for the nth BS. Then, performing SVD on $\widehat{\mathbf{H}}_{\Omega_L}^{(n)}$, we get

$$\widehat{\mathbf{H}}_{\Omega_L}^{(n)} = \widehat{\mathbf{U}}_{\Omega_L}^{(n)} \mathbf{\Upsilon}_{\Omega_L}^{(n)} \left(\widehat{\mathbf{V}}_{\Omega_L}^{(n)} \right)^{\dagger}. \quad (8.30)$$

Letting $\widehat{L} \triangleq \mathrm{rank}\left(\widehat{\mathbf{H}}_{\Omega_L}^{(n)}\right)$, we define $\widehat{\mathbf{V}}_{\Omega_L,1}^{(n)}$ as the $\left(\sum_{n=1}^{K_{\mathrm{mu}}} N_n\right) \times \widehat{L}$ matrix consisting of the first $\widehat{L}$ singular vectors of $\widehat{\mathbf{H}}_{\Omega_L}^{(n)}$, and also define $\widehat{\mathbf{V}}_{\Omega_L,0}^{(n)}$ as the $\left(\sum_{n=1}^{K_{\mathrm{mu}}} N_n\right) \times \left(\sum_{\ell=1}^{L} M_{i_\ell} - \widehat{L}\right)$ matrix consisting of the rest singular vectors of $\widehat{\mathbf{H}}_{\Omega_L}^{(n)}$. Since the column vectors of $\widehat{\mathbf{V}}_{\Omega_L,0}^{(n)}$ span the null space [93] of $\widehat{\mathbf{H}}_{\Omega_L}^{(n)}$ [92], we can set the precoding

matrix $\mathbf{\Gamma}_{\Omega_L}^{(n)}$ as follows:

$$\mathbf{\Gamma}_{\Omega_L}^{(n)} \triangleq \begin{cases} \widehat{\mathbf{V}}_{\Omega_L,0}^{(n)}, & \text{if } \widehat{L} < \sum_{\ell=1}^{L} M_{i_{L,\ell}}; \\ 0, & \text{if } \widehat{L} = \sum_{\ell=1}^{L} M_{i_{L,\ell}}, \end{cases} \tag{8.31}$$

Note that if $\widehat{L} = \sum_{\ell=1}^{L} M_{i_{L,\ell}}$, implying that $\widehat{\mathbf{H}}_{\Omega_L}^{(n)}$ has full row rank, there does not exist such a $\mathbf{\Gamma}_{\Omega_L}^{(n)}$ satisfying $\mathbf{H}_{\Omega_L}^{(i)}\mathbf{\Gamma}_{\Omega_L}^{(n)} = \mathbf{0}$ for all $i \neq n$. For this case, the nth user does not transmit to avoid interferences to other users.

2. Priority BS-Selection Strategy Given the BS Subset Cardinality

When the transmission mode is specified, i.e., the cardinality of the BS subset is given, every user expects to select the BS's to maximize its own transmission rate. However, it is clear that this objective cannot be obtained for all users in the multi-user case. Moreover, the derivation of global optimal selection strategy in terms of minimizing the average BS usage is intractable due to the too high complexity, where we need to examine all $\binom{K_{\text{bs}}}{L}$ possible BS combinations. Therefore, we propose a simple yet efficient BS-selection algorithm, called priority BS-selection, which is detailed as follows.

For the nth user, the global maximum achievable transmission rate is attained when all BS's are used and all the other users do not transmit. Thus, the maximum achievable rate is given by

$$R^{(n)}\left(\Omega_{K_{\text{bs}}}, \mathcal{P}_{K_{\text{bs}}}\right) = \max_{\mathbf{\Xi}^{(n)}:\text{Tr}\left(\mathbf{\Xi}^{(n)}\right)=\mathcal{P}_{K_{\text{bs}}}} \left\{ BT \log\left[\det\left(\mathbf{I} + \mathbf{H}_n \mathbf{\Xi}^{(n)} \mathbf{H}_n^{\dagger}\right)\right]\right\}. \tag{8.32}$$

Correspondingly, we get the maximum achievable effective capacity of the nth user, denoted by $\mathcal{C}_{\max}^{(n)}$, as follows:

$$\mathcal{C}_{\max}^{(n)} = -\frac{1}{\theta_n} \log\left(\mathbb{E}_{\mathbf{H}}\left\{e^{-\theta_n R^{(n)}\left(\Omega_{K_{\text{bs}}}, \mathcal{P}_{K_{\text{bs}}}\right)}\right\}\right), \qquad n = 1, 2, \ldots, K_{\text{mu}}. \tag{8.33}$$

Table IX. The Pseudo Codes to Determine Ω_L in Each Fading State by Using the Priority BS-Selection Algorithm for the Multi-User Case.

```
01.  Let Ψ := {1, 2, ..., K_bs} and Ψ̄ := ∅, and ℓ = |Ψ̄|;
                          ! Use variables Ψ̄ and Ψ to store selected BS's and all other BS', respectively.
02.  j := 1.              ! User π(j) is selecting BS.
03.  While (ℓ < L)        ! Iterative selections until L BS's are selected.
04.    m* = arg min_{m∈Ψ}{γ_{π(j),m}}.   ! γ_{π(j),m} is the aggregate power gain associated with user π(j).
                          ! Select the BS with the maximum aggregate power gain for user π(j).
05.    Ψ̄ := Ψ̄ ∪ {m*}, Ψ := Ψ\{m*}, and ℓ := ℓ + 1.   ! Update Ψ̄, Ψ, and ℓ.
06.    If j = K_mu, then j := 1; else j := j + 1.    ! Let next user with lower priority to select BS.
07.  End
08.  Ω_L := Ψ̄.                                      ! Complete the BS selection and get Ω_L.
```

We further define the effective-capacity fraction for the nth user as the ratio between the traffic load and the maximum achievable effective capacity. Denoting the effective-capacity fraction by $\widehat{C}_n$, we have $\widehat{C}_n \triangleq \overline{C}_n/\mathcal{C}_{\max}^{(n)}$. Clearly, the higher $\widehat{C}_n$ is, the more wireless resources the nth user requires to meet its QoS requirements. Thus, in order to satisfy the QoS requirements for all users, we assign higher BS-selection priority to the user with larger $\widehat{C}_n$. Following this principle, we design the priority BS-selection algorithm to determine Ω_L in each fading state and provide the pseudo code in Table IX. For presentation convenience, we permute $\{\widehat{C}_n\}_{n=1}^{K_{\rm mu}}$ in the decreasing order and denote the permuted version by $\{\widehat{C}_{\pi(j)}\}_{h=1}^{K_{\rm mu}}$, where $\widehat{C}_{\pi(1)} \geq \widehat{C}_{\pi(2)} \geq \cdots \geq \widehat{C}_{\pi(K_{\rm mu})}$ indicates the order from the higher priority to the lower priority. In the rest of this chapter, we use the term of user $\pi(i)$ to denote the user associated with the ith largest effective-capacity fraction.

As shown in Table IX, in each fading state the BS-selection procedure starts with the selection for user $\pi(1)$, who has the highest priority. After picking one BS for user $\pi(1)$, we select one different BS for user $\pi(2)$. More generally, after selecting for user $\pi(j)$, we choose one BS for user $\pi(j+1)$ from the BS-subset Ψ, which consists of the BS's that have not been selected. This procedure repeats until L BS's are selected. For user-$\pi(j)$'s selection, we choose the BS with the maximum aggregate power gain

(see Eq. (8.1) for its definition) over the subset Ψ. In addition, after user-$\pi(K_{\rm mu})$'s selection, if the number of selected BS's is still smaller than L, we continue selecting one more BS for user $\pi(1)$, as shown in line 06 in Table IX, and repeat this iterative selection procedure until having selected L BS's.

3. The Optimization Framework for BS-Selection and Resource Allocation

a. Problem formulation for average BS-usage minimization

We next study how to determine which transmission modes will be used, and how to derive the corresponding resource allocation strategy by integrating the block diagonalization and the priority BS selection. Similar to the OGBS-PT scheme for the single-user case, we also apply the probabilistic transmission for the multi-user case, where transmission mode L is used with a probability denoted by ϕ_L, $L = 1, 2, \ldots, K_{\rm bs}$. Note that for any transmission mode, there are $K_{\rm mu}$ coexisting links towards $K_{\rm mu}$ mobile users. Consequently, we also need to determine how to allocate the total power $\mathcal{P}_L$ to these $K_{\rm mu}$ coexisting links. In particular, we describe the power allocation strategy in a fading state by

$$\begin{cases} \boldsymbol{\mathcal{P}} & \triangleq \ (\boldsymbol{\mathcal{P}}_1, \boldsymbol{\mathcal{P}}_2, \ldots, \boldsymbol{\mathcal{P}}_{K_{\rm bs}}) \, ; \\ \boldsymbol{\mathcal{P}}_L & \triangleq \ \left(\mathcal{P}_L^{(1)}, \mathcal{P}_L^{(2)}, \ldots, \mathcal{P}_L^{(K_{\rm bs})} \right), \quad L = 1, 2, \ldots, K_{\rm bs}; \end{cases} \tag{8.34}$$

where $\boldsymbol{\mathcal{P}}$ denotes the power-allocation policy for the entire system and the vector $\boldsymbol{\mathcal{P}}_L$ represents the power-allocation policy for transmission mode L. Then, we formulate the following optimization problem **VIII-*A*3** to derive the optimal QoS-aware probability-vector $\boldsymbol{\phi}^* \triangleq (\phi_1^*, \phi_2^*, \ldots, \phi_{K_{\rm mu}}^*)$ and its corresponding power-allocation pol-

icy $\boldsymbol{\mathcal{P}}^*$.

$$\textbf{VIII-}\boldsymbol{A3}: \quad \min_{(\boldsymbol{\phi},\boldsymbol{\mathcal{P}})} \left\{\overline{L}\right\} = \min_{(\boldsymbol{\phi},\boldsymbol{\mathcal{P}})} \left\{ \mathbb{E}_{\mathbf{H}} \left\{ \sum_{L=0}^{K_{\rm bs}} \phi_L L \right\} \right\}$$

$$\text{s.t.: } 1). \sum_{L=0}^{K_{\rm bs}} \phi_L = 1, \qquad \forall\, \mathbf{H} \tag{8.35}$$

$$2). \sum_{n=1}^{K_{\rm mu}} \mathcal{P}_L^{(n)} = \mathcal{P}_L, \qquad \forall\, L,\, \mathbf{H}; \tag{8.36}$$

$$3). \mathbb{E}_{\mathbf{H}} \left\{ \left(\sum_{L=0}^{K_{\rm bs}} \phi_L e^{-\theta_n R^{(n)}\left(\Omega_L, \mathcal{P}_L^{(n)}\right)} \right) - e^{-\theta_n \overline{C}_n} \right\} \leq 0, \quad \forall\, n \tag{8.37}$$

where $R^{(n)}\left(\Omega_L, \mathcal{P}_L^{(n)}\right)$ is determined through Eq. (8.28). We call the optimal solution to **VIII-*A*3** as the PBS-BD-PT scheme.

b. The properties of $R^{(n)}\left(\Omega_L, \mathcal{P}_L^{(n)}\right)$

Before solving **VIII-*A*3**, we need to study the properties of $R^{(n)}\left(\Omega_L, \mathcal{P}_L^{(n)}\right)$. Let us consider the nth user with $\boldsymbol{\Gamma}_{\Omega_L}^{(n)}$ not equal to zero. Similar to the results summarized in Section D-1, the nth user's MIMO channel $\widehat{\boldsymbol{\Gamma}}_{\Omega_L}^{(n)}$ (after the block diagonalization) can be converted to $Z_L^{(n)}$ parallel Gaussian sub-channels, where $Z_L^{(n)}$ is the rank of $\widehat{\boldsymbol{\Gamma}}_{\Omega_L}^{(n)}$, the zth sub-channel's SNR is equal to $\varepsilon_{L,z}^{(n)}$, and $\sqrt{\varepsilon_{L,z}^{(n)}}$ is the zth largest nonzero singular value of $\widehat{\boldsymbol{\Gamma}}_{\Omega_L}^{(n)}$. The optimal power $\rho_{L,z}^{(n)}$ allocated to the zth sub-channel follows the water-filling allocation, which is equal to $\rho_{L,z}^{(n)} = \left[\mu_L^{(n)} - 1/\varepsilon_{L,z}^{(n)}\right]^+$, where $\mu_L^{(n)}$ is selected such that $\sum_{z=1}^{Z_L^{(n)}} \rho_{L,z}^{(n)} = \mathcal{P}_L^{(n)}$. Since $\widehat{\boldsymbol{\Gamma}}_{\Omega_L}^{(n)}$ has only $Z_L^{(n)}$ non-zero singular values, for presentation convenience, we define $1/\varepsilon_{L,i}^{(n)} \triangleq \infty$ for $i = Z_L^{(n)} + 1$. Accordingly, we can show that

$$\frac{dR^{(n)}\left(\Omega_L, \mathcal{P}_L^{(n)}\right)}{d\mathcal{P}_L^{(n)}} = \frac{BT}{\mu_L^{(n)}} \tag{8.38}$$

holds and that $R^{(n)}\big(\Omega_L, \mathcal{P}_L^{(n)}\big)$ is strictly concave over $\mathcal{P}_L^{(n)}$. Moreover, if $\mu_L^{(n)} \in \Big[1/\varepsilon_{L,i}^{(n)}, 1/\varepsilon_{L,i+1}^{(n)}\Big)$ for $i = 0, 1, 2, \ldots, Z_L^{(n)}$, we get:

$$\begin{cases} \text{(a).} \quad \mathcal{P}_L^{(n)} & = & i\mu_L^{(n)} - \sum_{j=1}^{i} \frac{1}{\varepsilon_{L,j}^{(n)}} \\ \text{(b).} \quad R^{(n)}\big(\Omega_L, \mathcal{P}_L^{(n)}\big) & = & BT \log\left(\prod_{j=1}^{i} \varepsilon_{L,j}^{(n)}\right) + BTi \log \mu_L^{(n)}. \end{cases} \tag{8.39}$$

Under the above transmission setups, we obtain the nth user's actual transmit signal as follows:

$$\mathbf{s}_{\Omega_L}^{(n)} = \mathbf{\Gamma}_{\Omega_L}^{(n)} \widehat{\mathbf{s}}_{\Omega_L}^{(n)} = \mathbf{\Gamma}_{\Omega_L}^{(n)} \mathbf{V}_{L,1}^{(n)} \mathbf{x}_L^{(n)},$$

where $\mathbf{x}_L^{(n)} = \left(x_{L,1}^{(n)}, x_{L,1}^{(n)}, \ldots, x_{L,Z_L^{(n)}}^{(n)}\right)^{\tau}$ is the signal vector for $Z_L^{(n)}$ parallel Gaussian sub-channels, the power of $x_{L,z}^{(n)}$ is equal to $\rho_{L,z}^{(n)}$, and $\mathbf{V}_{L,1}^{(n)}$ is composed by the first $Z_L^{(n)}$ right singular vectors generated from SVD on $\widehat{\mathbf{\Gamma}}_{\Omega_L}^{(n)}$. Then, using the similar expressions as given in Eq. (8.10), the transmit power used for the nth user on the i_ℓth BS, denoted by $\varrho_{L,i_\ell}^{(n)}$, is given by the following equations:

$$\begin{cases} \boldsymbol{\rho}_L^{(n)} & = & \left(\rho_{L,1}^{(n)}, \rho_{L,2}^{(n)}, \ldots, \rho_{L,Z_L^{(n)}}^{(n)}\right)^{\tau}; \\ \widehat{\boldsymbol{\rho}}_L^{(n)} & = & \left(\left(\mathbf{\Gamma}_{\Omega_L}^{(n)} \mathbf{V}_{L,1}^{(n)}\right) \circ \text{conj}\left(\mathbf{\Gamma}_{\Omega_L}^{(n)} \mathbf{V}_{L,1}^{(n)}\right)\right) \boldsymbol{\rho}_L^{(n)}; \\ \varrho_{L,i_{L,\ell}}^{(n)} & = & \sum_{w=W_{L,\ell-1}}^{W_{L,\ell}} \widehat{\boldsymbol{\rho}}_L^{(n)}(w). \end{cases} \tag{8.40}$$

where $W_{L,\ell} \triangleq \sum_{j=1}^{\ell} M_{i_{L,j}}$ and $(\cdot \circ \cdot)$ denotes the element-wise product between two matrices.

c. The optimal solution to **VIII-*A*3**

Theorem 15. *The optimal power-allocation policy $\boldsymbol{\mathcal{P}}^*$ for optimization problem* **VIII-*A*3**, *if existing, is given as follows:*

$$\left(\mathcal{P}_L^{(n)}\right)^* = i^* \left(\prod_{j=1}^{i^*} \varepsilon_{L,j}^{(n)}\right)^{-\frac{BT\theta_n}{1+i^*BT\theta_n}} \left(\frac{\zeta_{\mathbf{H},L}^*}{BT\theta_n\lambda_n^*}\right)^{-\frac{1}{1+i^*BT\theta_n}} - \sum_{j=1}^{i^*} \frac{1}{\varepsilon_{L,j}^{(n)}}, \quad (8.41)$$

for all n, L, and $\mathbf{H}$, where $\varepsilon_{L,j}^{(n)}$ is the square of $\widehat{\mathbf{\Gamma}}_{\Omega_L}^{(n)}$'s jth largest singular value, and i^* is the unique solution satisfying the following condition:

$$\mu_L^{(n)} \in \left[\frac{1}{\varepsilon_{L,i^*}^{(n)}}, \frac{1}{\varepsilon_{L,i^*+1}^{(n)}}\right), \quad \forall\, n,\, L,\, \mathbf{H}, \quad (8.42)$$

where

$$\mu_L^{(n)} = \max\left\{\frac{1}{\varepsilon_{L,1}^{(n)}}, \quad \left(\prod_{j=1}^{i^*} \varepsilon_{L,j}^{(n)}\right)^{-\frac{BT\theta_n}{1+i^*BT\theta_n}} \left(\frac{\zeta_{\mathbf{H},L}^*}{BT\theta_n\lambda_n^*}\right)^{-\frac{1}{1+i^*BT\theta_n}}\right\}. \quad (8.43)$$

The corresponding optimal probability-transmission policy is determined by

$$\phi_L^* = \begin{cases} 1, & \text{if } L = L^*; \\ 0, & \text{otherwise} \end{cases} \quad (8.44)$$

where

$$L^* = \arg\min_{0\le L\le K_{\rm bs}} \left\{L + \sum_{n=1}^{K_{\rm mu}} \lambda_n^* e^{-\theta_n BT\left(\log\left(\prod_{j=1}^{i^*}\varepsilon_{L,j}^{(n)}\right) + i^*\log\mu_L^{(n)}\right)}\right\}, \quad \forall\, \mathbf{H}, \quad (8.45)$$

where in Eqs. (8.41)-(8.43), for the given $\{\lambda_n^*\}_{n=1}^{K_{\rm mu}}$ the optimal $\zeta_{\mathbf{H},L}^*$ is selected to satisfy the equation $\sum_{n=1}^{K_{\rm mu}}\left(\mathcal{P}_L^{(n)}\right)^* = \mathcal{P}_L$ for all L and $\mathbf{H}$; $\{\lambda_n^*\}_{n=1}^{K_{\rm mu}}$ are constants over all $\mathbf{H}$, which are selected such that the equality of Eq. (8.37) holds.

Proof. Note that the optimization problem **VIII-*A*3** is not convex, because the constraint function on the left-hand side of Eq. (8.37) is not convex with respect to (w.r.t.) $\boldsymbol{\phi}$ and $\boldsymbol{\mathcal{P}}$. Then, we need to apply the Lagrangian duality theory [79] to solve

for the optimal solution. We construct **VIII-*A3***'s Lagrangian function, denoted by $\mathcal{J}_{A3}(\boldsymbol{\phi}, \boldsymbol{\mathcal{P}}; \boldsymbol{\lambda}, \boldsymbol{\zeta}_{\mathbf{H}})$, as

$$\mathcal{J}_{A3}(\boldsymbol{\phi}, \boldsymbol{\mathcal{P}}; \boldsymbol{\lambda}, \boldsymbol{\zeta}_{\mathbf{H}}) = \mathbb{E}_{\mathbf{H}} \left\{ J_{A3}(\boldsymbol{\phi}, \boldsymbol{\mathcal{P}}; \boldsymbol{\lambda}, \boldsymbol{\zeta}_{\mathbf{H}}) \right\} \tag{8.46}$$

with

$$\begin{aligned} J_{A3}(\boldsymbol{\phi}, \boldsymbol{\mathcal{P}}; \boldsymbol{\lambda}, \boldsymbol{\zeta}_{\mathbf{H}}) \triangleq & \sum_{L=0}^{K_{\mathrm{bs}}} \phi_L L + \sum_{n=1}^{K_{\mathrm{mu}}} \lambda_n \left[-e^{-\theta_n \overline{C}_n} + \sum_{L=0}^{K_{\mathrm{bs}}} \phi_L e^{-\theta_n R^{(n)}\left(\Omega_L, \mathcal{P}_L^{(n)}\right)} \right] \\ & + \sum_{L=1}^{K_{\mathrm{bs}}} \zeta_{\mathbf{H},L} \left(-\mathcal{P}_L + \sum_{n=1}^{K_{\mathrm{mu}}} \mathcal{P}_L^{(n)} \right) \end{aligned} \tag{8.47}$$

for $\sum_{L=1}^{K_{\mathrm{bs}}} \phi_L = 1$, where $\lambda_n \geq 0$ for $n = 1, 2, \ldots, K_{\mathrm{mu}}$ are the Lagrangian multipliers associated with Eq. (8.37), which are constants over all fading states, and $\boldsymbol{\lambda} \triangleq (\lambda_1, \lambda_2, \ldots, \lambda_{K_{\mathrm{mu}}})$; $\{\zeta_{\mathbf{H},L}\}_{L=1}^{K_{\mathrm{bs}}}$ are the Lagrangian multipliers associated with Eq. (8.36) for L transmission modes in each fading state, which are functions of $\mathbf{H}$ and L, and $\boldsymbol{\zeta}_{\mathbf{H},L} \triangleq (\zeta_{\mathbf{H},1}, \zeta_{\mathbf{H},2}, \ldots, \zeta_{\mathbf{H},K_{\mathrm{bs}}})$.

The optimization problem **VIII-*A3***'s Lagrangian dual function [77, 79], denoted by $\widetilde{\mathcal{J}}_{A3}(\boldsymbol{\lambda}, \boldsymbol{\zeta}_{\mathbf{H}})$, is determined by

$$\widetilde{\mathcal{J}}_{A3}(\boldsymbol{\lambda}, \boldsymbol{\zeta}_{\mathbf{H}}) \triangleq \min_{\boldsymbol{\phi}, \boldsymbol{\mathcal{P}}} \left\{ \mathcal{J}_{A3}(\boldsymbol{\phi}, \boldsymbol{\mathcal{P}}; \boldsymbol{\lambda}, \boldsymbol{\zeta}_{\mathbf{H}}) \right\} = \mathbb{E}_{\mathbf{H}} \left\{ \min_{\boldsymbol{\phi}, \boldsymbol{\mathcal{P}}} \left\{ J_{A3}(\boldsymbol{\phi}, \boldsymbol{\mathcal{P}}; \boldsymbol{\lambda}, \boldsymbol{\zeta}_{\mathbf{H}}) \right\} \right\} \tag{8.48}$$

We denote the minimizer pair in Eq. (8.48) by $(\widetilde{\boldsymbol{\phi}}, \widetilde{\boldsymbol{\mathcal{P}}})$. Then, we can derive

$$\widetilde{\boldsymbol{\phi}} = \arg \min_{\boldsymbol{\phi}: \sum_{L=1}^{K_{\mathrm{bs}}} \phi_L = 1} \left\{ J_{A3}(\boldsymbol{\phi}, \widetilde{\boldsymbol{\mathcal{P}}}; \boldsymbol{\lambda}, \boldsymbol{\zeta}_{\mathbf{H}}) \right\} \tag{8.49}$$

$$\overset{(a)}{=} \arg \min_{\boldsymbol{\phi}: \sum_{L=1}^{K_{\mathrm{bs}}} \phi_L = 1} \left\{ \sum_{L=1}^{K_{\mathrm{bs}}} \phi_L \left(L + \sum_{n=1}^{K_{\mathrm{mu}}} \lambda_n e^{-\theta_n R^{(n)}\left(\Omega_L, \widetilde{\mathcal{P}}_L^{(n)}\right)} \right) \right\}, \quad \forall\, \mathbf{H}, \tag{8.50}$$

where equation (a) holds by plugging Eq. (8.47) into Eq. (8.49) and removing the

terms independent of $\boldsymbol{\phi}$. Solving Eq. (8.50), we obtain

$$\widetilde{\phi}_L = \begin{cases} 1, & \text{if } L = \arg \min\limits_{\ell=1,2,\ldots,K_{\rm bs}} \left\{ \ell + \sum_{n=1}^{K_{\rm mu}} \lambda_n e^{-\theta_n R^{(n)}\left(\Omega_\ell, \widetilde{\mathcal{P}}_\ell^{(n)}\right)} \right\}; \\ 0, & \text{otherwise.} \end{cases} \tag{8.51}$$

Based on Eq. (8.51), the opportunity of transmitting the data in a fading state will be given to only one transmission mode. Moreover, when $\widetilde{\phi}_L = 1$, the values of $\boldsymbol{\mathcal{P}}_j$ for $j \neq L$ do not affect the Lagrangian function $\mathcal{J}_{A3}(\boldsymbol{\phi}, \boldsymbol{\mathcal{P}}; \boldsymbol{\lambda}, \boldsymbol{\zeta}_{\mathbf{H}})$. Therefore, $\widetilde{\boldsymbol{\mathcal{P}}}_L$ needs to minimize $J_{A3}(\boldsymbol{\phi}, \boldsymbol{\mathcal{P}}; \boldsymbol{\lambda}, \boldsymbol{\zeta}_{\mathbf{H}})$ given $\phi_L = 1$ and $\phi_j = 0$ for all $j \neq L$.

Following the above derivations, we define a set of functions $J_{A3,L}(\boldsymbol{\mathcal{P}}; \boldsymbol{\lambda}, \boldsymbol{\zeta}_{\mathbf{H},L})$ for $L = 1, 2, \ldots, K_{\rm bs}$, where $J_{A3,L}(\boldsymbol{\mathcal{P}}; \boldsymbol{\lambda}, \boldsymbol{\zeta}_{\mathbf{H},L}) \triangleq J_{A3}(\boldsymbol{\phi}, \boldsymbol{\mathcal{P}}; \boldsymbol{\lambda}, \boldsymbol{\zeta}_{\mathbf{H},L})|_{\phi_L=1;\phi_j=0,j\neq L}$. Taking the derivative of $J_{A3,L}(\boldsymbol{\mathcal{P}}; \boldsymbol{\lambda}, \boldsymbol{\zeta}_{\mathbf{H},L})$ w.r.t. $\mathcal{P}_L^{(n)}$ and letting the derivative equal to zero, we get

$$\zeta_{\mathbf{H},L} - BT\lambda_n\theta_n\mu_L^{(n)} e^{-\theta_n R^{(n)}\left(\Omega_L, \mathcal{P}_L^{(n)}\right)} = 0, \quad \forall n,\ L,\ \mathbf{H} \tag{8.52}$$

where $\mu_L^{(n)} = dR^{(n)}\left(\Omega_L, \mathcal{P}_L^{(n)}\right)/d\mathcal{P}_L^{(n)}$ (see Eq. (8.38)). Plugging Eq. (8.39)-(b) into Eq. (8.52) and solving for the optimal $\mu_L^{(n)}$ under the boundary condition of $\mu_L^{(n)} \geq 1/\varepsilon_{L,1}^{(n)}$, we obtain Eq. (8.43). Since Eq. (8.39) is obtained under the condition of $\mu_L^{(n)} \in \left[1/\varepsilon_{L,i}^{(n)}, 1/\varepsilon_{L,i+1}^{(n)}\right)$, the variable i^* in Eq. (8.43) must satisfy the condition of $\mu_L^{(n)} \in \left[1/\varepsilon_{L,i^*}^{(n)}, 1/\varepsilon_{L,i^*+1}^{(n)}\right)$, as shown in Eq. (8.42). Moreover, we can show that $J_{A3,L}(\boldsymbol{\mathcal{P}}; \boldsymbol{\lambda}, \boldsymbol{\zeta}_{\mathbf{H},L})$ is a strictly convex function, and thus i^* for Eq. (8.43) is unique. Then, we can obtain $\boldsymbol{\mathcal{P}}_L$ by using Eq. (8.39)-(a).

The Lagrangian duality principle [79] shows that $\widetilde{\mathcal{J}}(\boldsymbol{\lambda}, \boldsymbol{\zeta}_{\mathbf{H}}) = \mathcal{J}(\widetilde{\boldsymbol{\phi}}, \widetilde{\boldsymbol{\mathcal{P}}}; \boldsymbol{\lambda}, \boldsymbol{\zeta}_{\mathbf{H}})$ is concave over $\boldsymbol{\lambda}$ and $\boldsymbol{\zeta}_{\mathbf{H}}$. Moreover, the original optimization problem (also called the primal problem) **VIII-*A3***'s dual problem is defined by

$$\max_{\boldsymbol{\lambda}, \boldsymbol{\zeta}_{\mathbf{H}}} \left\{ \widetilde{\mathcal{J}}(\boldsymbol{\lambda}, \boldsymbol{\zeta}_{\mathbf{H}}) \right\}. \tag{8.53}$$

Denoting the optimal objective of the primal problem **VIII-*A*3** by $\overline{L}^*$, we always have

$$\overline{L}^* \geq \max_{\boldsymbol{\lambda},\boldsymbol{\zeta}_{\mathbf{H}}} \left\{ \widetilde{\mathcal{J}}(\boldsymbol{\lambda}, \boldsymbol{\zeta}_{\mathbf{H}}) \right\}. \tag{8.54}$$

We can further show that $\widetilde{\mathcal{J}}(\boldsymbol{\lambda}, \boldsymbol{\zeta}_{\mathbf{H}})$ is differentiable w.r.t. $\boldsymbol{\lambda}$ and $\boldsymbol{\zeta}_{\mathbf{H}}$. Then, the Lagrangian duality principle [79] shows that

$$\begin{cases} \frac{\partial \widetilde{\mathcal{J}}(\boldsymbol{\lambda},\boldsymbol{\zeta}_{\mathbf{H}})}{\partial \lambda_n} &= \mathbb{E}_{\mathbf{H}} \left\{ \sum_{L=0}^{K_{\mathrm{bs}}} \widetilde{\phi}_L e^{-\theta_n R^{(n)}\left(\Omega_L, \widetilde{\mathcal{P}}_L^{(n)}\right)} - e^{-\theta_n \overline{C}_n} \right\}, \quad \forall\, n; \\ \frac{\partial \widetilde{\mathcal{J}}(\boldsymbol{\lambda},\boldsymbol{\zeta}_{\mathbf{H}})}{\partial \zeta_{\mathbf{H},L}} &= \left(\sum_{n=1}^{K_{\mathrm{mu}}} \widetilde{\mathcal{P}}_L^{(n)} - \mathcal{P}_L \right) g(\mathbf{H}) d\mathbf{H}, \quad \forall\, L,\, \mathbf{H} \end{cases} \tag{8.55}$$

where $g(\mathbf{H})$ is the probability density function of $\mathbf{H}$ and $d\mathbf{H}$ denotes the integration variable. It is clear that if $\partial\widetilde{\mathcal{J}}(\boldsymbol{\lambda}, \boldsymbol{\zeta}_{\mathbf{H}})/\partial\lambda_n = 0$ (for all n) and $\partial\widetilde{\mathcal{J}}(\boldsymbol{\lambda}, \boldsymbol{\zeta}_{\mathbf{H}})/\partial\zeta_{\mathbf{H},L} = 0$ (for all L and $\mathbf{H}$) hold, $\widetilde{\mathcal{J}}(\boldsymbol{\lambda}, \boldsymbol{\zeta}_{\mathbf{H}})$ attains its maximum. Equations (8.39)-(a) and (8.43) indicate that given any λ_n, $\widetilde{\mathcal{P}}_L^{(n)}$ is a decreasing function of $\zeta_{\mathbf{H},L}$. Moreover, $\zeta_{\mathbf{H},L} \to 0$ and $\zeta_{\mathbf{H},L} \to \infty$ leads to $\widetilde{\mathcal{P}}_L^{(n)} \to \infty$ and $\widetilde{\mathcal{P}}_L^{(n)} \to 0$, respectively. Thus, for any λ_n, we can always find a $\zeta_{\mathbf{H},L} = \zeta_{\mathbf{H}}^*$ such that $\sum_{n=1}^{K_{\mathrm{mu}}} \widetilde{\mathcal{P}}_L^{(n)} - \mathcal{P}_L = 0$, implying $\partial\widetilde{\mathcal{J}}(\boldsymbol{\lambda}, \boldsymbol{\zeta}_{\mathbf{H}}^*)/\partial\zeta_{\mathbf{H},L} = 0$.

Having obtained $\boldsymbol{\zeta}_{\mathbf{H}}^*$, we next focus on the optimal $\boldsymbol{\lambda}^*$ to maximize $\widetilde{\mathcal{J}}(\boldsymbol{\lambda}, \boldsymbol{\zeta}_{\mathbf{H}}^*)$. Due to the concavity of $\widetilde{\mathcal{J}}_{A3}(\boldsymbol{\lambda}, \boldsymbol{\zeta}_{\mathbf{H}})$, $\partial\widetilde{\mathcal{J}}(\boldsymbol{\lambda}, \boldsymbol{\zeta}_{\mathbf{H}}^*)/\partial\lambda_n$ is a decreasing function of λ_n. Also, we can readily show that $\partial\widetilde{\mathcal{J}}(\boldsymbol{\lambda}, \boldsymbol{\zeta}_{\mathbf{H}}^*)/\partial\lambda_n|_{\lambda_n=0} > 0$. Then, if there does not exist $\boldsymbol{\lambda}$ such that $\partial\widetilde{\mathcal{J}}(\boldsymbol{\lambda}, \boldsymbol{\zeta}_{\mathbf{H}}^*)/\partial\lambda_n = 0$ for all n, we have $\lambda_n^* \to \infty$ for some nth user and $\partial\widetilde{\mathcal{J}}(\boldsymbol{\lambda}, \boldsymbol{\zeta}_{\mathbf{H}}^*)/\partial\lambda_n > 0$ always holds. For this case, we get $\overline{L}^* \geq \widetilde{\mathcal{J}}(\boldsymbol{\lambda}^*, \boldsymbol{\zeta}_{\mathbf{H}}^*) \to \infty$, implying that there is no feasible solution for **VIII-*A*3**.

In contrast, if there exists $\boldsymbol{\lambda}^*$ such that $\partial\widetilde{\mathcal{J}}(\boldsymbol{\lambda}^*, \boldsymbol{\zeta}_{\mathbf{H}}^*)/\partial\lambda_n = 0$ for all n, the pair of $(\boldsymbol{\lambda}^*, \boldsymbol{\zeta}_{\mathbf{H}}^*)$ is the optimal solution to the dual problem given by Eq. (8.53). Note that the optimum of Eq. (8.53) is achieved by using $(\widetilde{\boldsymbol{\mathcal{P}}}, \widetilde{\boldsymbol{\phi}})$ given $\boldsymbol{\lambda}^*$ and $\boldsymbol{\zeta}_{\mathbf{H}}^*$. Clearly, this policy is feasible for the primal problem **VIII-*A*3**, implying that the

equality of Eq. (8.54) holds with zero duality gap. As a result, this policy is the optimal solution to **VIII-*A*3**. Then, setting $\boldsymbol{\mathcal{P}}^* = \widetilde{\boldsymbol{\mathcal{P}}}$ and $\boldsymbol{\phi}^* = \widetilde{\boldsymbol{\phi}}$ with $\boldsymbol{\lambda}^*$ and $\boldsymbol{\zeta}^*_{\mathbf{H}}$ in Eq. (8.51), we obtain Eqs. (8.44)-(8.45). Further plugging Eq. (8.42) into Eq. (8.39)-(a), we prove that Eq. (8.41) holds. Finally, comparing $\partial\widetilde{\mathcal{J}}(\boldsymbol{\lambda}^*, \boldsymbol{\zeta}^*_{\mathbf{H}})/\partial\lambda_n = 0$ with Eq. (8.55), we show that the equality of Eq. (8.37) holds, which completes the proof of Theorem 15. □

Note that the general closed-form expressions for the optimal Lagrangian multipliers $\boldsymbol{\zeta}^*_{\mathbf{H},L}$ and $\boldsymbol{\lambda}^*$ are hard to obtain. In contrast, we can use numerical searching techniques to determine the values of $\boldsymbol{\zeta}^*_{\mathbf{H},L}$ and $\boldsymbol{\lambda}^*$. In particular, as shown in the proof of Theorem 15, $\widetilde{\mathcal{P}}^{(n)}_L$ is a monotonically decreasing function of $\zeta_{\mathbf{H},L}$. Then, we define

$$\begin{cases} \zeta'_{\mathbf{H},L,n} & \triangleq & BT\lambda_n\theta_n\mu^{(n)}_L e^{-\theta_n R^{(n)}(\Omega_L, \widetilde{P}^{(n)}_L)}\big|_{\widetilde{P}^{(n)}_L=1}; \\ \zeta''_{\mathbf{H},L,n} & \triangleq & BT\lambda_n\theta_n\mu^{(n)}_L e^{-\theta_n R^{(n)}(\Omega_L, \widetilde{P}^{(n)}_L)}\big|_{\widetilde{P}^{(n)}_L=1/K_{\text{mu}}}, \end{cases}$$

for all $n = 1, 2, \ldots, K_{\text{mu}}$. It is clear that $\zeta^*_{\mathbf{H},L} \in \big[\max_n\{\zeta'_{\mathbf{H},L,n}\}, \max_n\{\zeta''_{\mathbf{H},L,n}\}\big]$. Consequently, we can search obtain $\zeta_{\mathbf{H},L} = \zeta^*_{\mathbf{H},L}$ to satisfy $\sum_{n=1}^{K_{\text{mu}}} \widetilde{\mathcal{P}}^{(n)}_L - \mathcal{P}_L = 0$ through the bisection based numerical searching technique.

Next, we focus on how to derive $\boldsymbol{\lambda}^*$. As discussed in the proof of Theorem 15, $\widetilde{\mathcal{J}}(\boldsymbol{\lambda}, \zeta_{\mathbf{H}})$ is concave over $\boldsymbol{\lambda}$ and $\zeta_{\mathbf{H}}$. Therefore, $\widetilde{\mathcal{J}}(\boldsymbol{\lambda}, \zeta^*_{\mathbf{H}})$ is also concave over $\boldsymbol{\lambda}$ since $\zeta^*_{\mathbf{H}}$ maximizes $\widetilde{\mathcal{J}}(\boldsymbol{\lambda}, \zeta_{\mathbf{H}})$ under the given $\boldsymbol{\lambda}$. Correspondingly, we can apply the gradient descent method to search for the optimal $\boldsymbol{\lambda}^*$, which maximizes $\widetilde{\mathcal{J}}(\boldsymbol{\lambda}, \zeta^*_{\mathbf{H}})$. Specifically, the iterative searching procedures is given by

$$\lambda_n := \lambda_n + \epsilon\lambda_n \frac{\partial\widetilde{\mathcal{J}}(\boldsymbol{\lambda}, \boldsymbol{\zeta}^*_{\mathbf{H}})}{\partial\lambda_n}, \quad n = 1, 2, \ldots, K_{\text{mu}} \tag{8.56}$$

where ϵ is a small positive real number and $\widetilde{\mathcal{J}}(\boldsymbol{\lambda}, \boldsymbol{\zeta}^*_{\mathbf{H}})/\partial\lambda_n$ is given by Eq. (8.55). Moreover, for fast simulations, we can use time average via the first-order AR low-

pass filter to estimate $\widetilde{\mathcal{J}}(\boldsymbol{\lambda}, \boldsymbol{\zeta}_{\mathbf{H}}^*)/\partial\lambda_n$. Denoting the estimate of $\widetilde{\mathcal{J}}(\boldsymbol{\lambda}, \boldsymbol{\zeta}_{\mathbf{H}}^*)/\partial\lambda_n$ at the t-th time by $\widehat{\mathcal{J}}_n[k]$, we can use the following first-order AR low-pass filter to obtain the estimate $\widehat{\mathcal{J}}_n[k]$ as follows:

$$\widehat{\mathcal{J}}_n[k] = \vartheta\widehat{\mathcal{J}}_n[k-1] + (1-\vartheta)\left(\sum_{L=0}^{K_{\text{bs}}} \widetilde{\phi}_L e^{-\theta_n R^{(n)}\left(\Omega_L[t], \widetilde{\mathcal{P}}_L^{(n)}[t]\right)} - e^{-\theta_n \overline{C}_n}\right), \qquad (8.57)$$

where $\vartheta \in (0,1)$ is a real number close to 1. If the optimal solution exists, the above searching algorithm can effectively converge with the appropriately selected ϵ and ϑ.

4. The TDMA Based BS-Selection Scheme

We next study the TDMA based BS-selection scheme, which servers as the baseline scheme for comparative analyses with our proposed scheme. The TDMA based multi-user transmissions in MIMO link typically lead to lower throughput than the block-diagonalization based scheme. However, the previous research works did not study and compare the TDMA based and block-diagonalization based approaches in terms of BS-selection and QoS provisioning for distributed MIMO systems.

In the TDMA based BS-selection, we also apply the priority BS-selection algorithm given by Table IX when transmission mode L is specified. For transmission mode L, we further divide each time frame into K_{mu} time slots[3] for data transmissions to K_{mu} users, respectively. The nth user's time-slot length is set equal to $T \times t_{L,n}$ for $n = 1, 2, \ldots, K_{\text{mu}}$, where $t_{L,n}$ is the normalized time-slot length. Moreover, we still use the probabilistic transmission strategy across different transmission modes, where the probability of using transmission mode L to transmit data is equal to ϕ_L.

[3]In this chapter, the TDMA based scheme described in Section E-4 and the time-sharing based scheme described in Section D-3 both partition each time frame into a number of time slots. However, note that we use the term of TDMA for multiple access across multiple users, and use the term of time-sharing for time-division transmission across different transmission modes.

Then, we derive the TDMA and probabilistic transmission policies through solving the following optimization problem **VIII-*A*4**.

$$\textbf{VIII-}\boldsymbol{A}\textbf{4}: \quad \min_{(\boldsymbol{t},\boldsymbol{\phi})} \left\{\overline{L}\right\} = \min_{(\boldsymbol{t},\boldsymbol{\phi})} \left\{ \mathbb{E}_{\mathbf{H}} \left\{ \sum_{L=0}^{K_{\rm bs}} L\phi_L \right\} \right\}$$

$$\text{s.t.: } 1). \sum_{L=0}^{K_{\rm bs}} \phi_L = 1, \qquad \forall\, \mathbf{H}, \tag{8.58}$$

$$2). \sum_{n=1}^{K_{\rm mu}} t_{L,n} = 1, \qquad \forall\, \mathbf{H},\ L = 1, 2, \ldots, K_{\rm bs}, \tag{8.59}$$

$$3). \ \mathbb{E}_{\mathbf{H}} \left\{ \left(\sum_{L=0}^{K_{\rm bs}} \phi_L e^{-\theta_n t_n R^{(n)}(\Omega_L, \mathcal{P}_L)} \right) - e^{-\theta_n \overline{C}_n} \right\} \leq 0, \quad \forall\, n, \tag{8.60}$$

where $\boldsymbol{\phi}$ and $\boldsymbol{t}$ are functions of $\mathbf{H}$, $\boldsymbol{\phi} \triangleq (\phi_0, \phi_1, \phi_2, \ldots, \phi_{K_{\rm mu}})$, $\boldsymbol{t} \triangleq (\boldsymbol{t}_1, \boldsymbol{t}_2, \ldots, \boldsymbol{t}_{K_{\rm bs}})$, and $\boldsymbol{t}_L \triangleq (t_{L,1}, t_{L,2}, \ldots, t_{L,K_{\rm bs}})$.

Theorem 16. *Problem* **VIII-*A*4***'s optimal solution pair* $(\boldsymbol{t}^*, \boldsymbol{\phi}^*)$*, if existing, is determined by*

$$t^*_{L,n} = \left[\frac{1}{\theta_n R^{(n)}\left(\Omega_L, \mathcal{P}_L\right)} \log \left(\frac{\lambda^*_n \theta_n R^{(n)}\left(\Omega_L, \mathcal{P}_L\right)}{\delta^*_{\mathbf{H},L}} \right) \right]^+, \quad \forall L,\ n,\ \mathbf{H} \tag{8.61}$$

and

$$\phi^*_L = \begin{cases} 1, & \text{if } L = \arg \min\limits_{\ell=0,1,\ldots,K_{\rm bs}} \left\{ \ell + \sum_{n=1}^{K_{\rm mu}} \lambda^*_n e^{-\theta_n t^*_{L,n} R^{(n)}(\Omega_\ell, \mathcal{P}_\ell)} \right\}; \\ 0, & \text{otherwise} \end{cases} \tag{8.62}$$

for all L *and* $\mathbf{H}$*, where* $\delta^*_{\mathbf{H},L}$ *under given* $\{\lambda^*_n\}_{n=1}^{K_{\rm mu}}$ *is determined by satisfying* $\sum_{n=1}^{K_{\rm mu}} t^*_{L,n} = 1$*, and* $\{\lambda^*_n\}_{n=1}^{K_{\rm mu}}$ *needs to be selected such that the equality of Eq. (8.60) holds.*

Proof. We construct **VIII-*A*4**'s Lagrangian function, denoted by $\mathcal{J}_{A4}(\boldsymbol{t}, \boldsymbol{\phi}; \boldsymbol{\lambda}, \boldsymbol{\delta}_{\mathbf{H}})$, as follows:

$$\mathcal{J}_{A4}(\boldsymbol{t}, \boldsymbol{\phi}; \boldsymbol{\lambda}, \boldsymbol{\delta}_{\mathbf{H}}) = \mathbb{E}_{\mathbf{H}} \left\{ J_{A4}(\boldsymbol{t}, \boldsymbol{\phi}; \boldsymbol{\lambda}, \boldsymbol{\delta}_{\mathbf{H}}) \right\} \tag{8.63}$$

with

$$J_{A4}(\boldsymbol{t},\boldsymbol{\phi};\boldsymbol{\lambda},\boldsymbol{\delta}_{\mathbf{H}}) \triangleq \sum_{L=0}^{K_{\rm bs}} L\phi_L + \sum_{L=1}^{K_{\rm bs}} \delta_{\mathbf{H},L}\left(\sum_{n=1}^{K_{\rm mu}} t_{L,n} - 1\right) + \sum_{n=1}^{K_{\rm mu}} \lambda_n \left(\sum_{L=0}^{K_{\rm bs}} \phi_L e^{-\theta_n t_n R^{(n)}(\Omega_L,\mathcal{P}_L)} - e^{-\theta_n \overline{C}_n}\right) \quad (8.64)$$

subject to $\sum_{L=0}^{K_{\rm bs}} \phi_L = 1$ for all $\mathbf{H}$, where $\delta_{\mathbf{H}} \triangleq (\delta_{\mathbf{H},1}, \delta_{\mathbf{H},2}, \ldots, \delta_{\mathbf{H},K_{\rm bs}})$ represents the Lagrangian multipliers associated with Eq. (8.59), $\lambda_n \geq 0$ for $n = 1, 2, \ldots, K_{\rm mu}$ are the Lagrangian multipliers associated with Eq. (8.60), and $\boldsymbol{\lambda} \triangleq (\lambda_1, \lambda_2, \ldots, \lambda_{K_{\rm mu}})$. The optimization problem **VIII-*A4***'s Lagrangian dual function, denoted by $\widetilde{\mathcal{J}}_{A4}(\boldsymbol{\lambda},\boldsymbol{\delta}_{\mathbf{H}})$, is determined by

$$\begin{aligned}\widetilde{\mathcal{J}}_{A4}(\boldsymbol{\lambda},\boldsymbol{\delta}_{\mathbf{H}}) &\triangleq \min_{(\boldsymbol{t},\boldsymbol{\phi})}\left\{\mathcal{J}_{A4}(\boldsymbol{t},\boldsymbol{\phi};\boldsymbol{\lambda},\boldsymbol{\delta}_{\mathbf{H}})\right\} = \mathbb{E}_{\mathbf{H}}\left\{\min_{(\boldsymbol{t},\boldsymbol{\phi})}\left\{J_{A4}(\boldsymbol{t},\boldsymbol{\phi};\boldsymbol{\lambda},\boldsymbol{\delta}_{\mathbf{H}})\right\}\right\} \\ &= \mathbb{E}_{\mathbf{H}}\left\{\min_{(\boldsymbol{t},\boldsymbol{\phi})}\left\{\sum_{L=0}^{K_{\rm bs}}\phi_L\left(L + \sum_{n=1}^{K_{\rm mu}}\lambda_n e^{-\theta_n t_n R^{(n)}(\Omega_L,\mathcal{P}_L)}\right) + \sum_{L=1}^{K_{\rm bs}}\delta_{\mathbf{H},L}\left(\sum_{n=1}^{K_{\rm mu}} t_{L,n} - 1\right) - \sum_{n=1}^{K_{\rm mu}}\lambda_n e^{-\theta_n \overline{C}_n}\right\}\right\}. \quad (8.65)\end{aligned}$$

We denote the minimizer pair in Eq. (8.65) by $(\boldsymbol{t}^*, \boldsymbol{\phi}^*)$. Lagrangian duality principle [79] suggests that $\widetilde{\mathcal{J}}_{A4}(\boldsymbol{\lambda},\boldsymbol{\delta}_{\mathbf{H}})$ is a concave function and **VIII-*A4***'s dual problem is given by $\max_{\boldsymbol{\lambda},\boldsymbol{\delta}_{\mathbf{H}}}\{\widetilde{\mathcal{J}}_{A4}(\boldsymbol{\lambda},\boldsymbol{\delta}_{\mathbf{H}})\}$. We then denote the maximizer pair for $\widetilde{\mathcal{J}}_{A4}(\boldsymbol{\lambda},\boldsymbol{\delta}_{\mathbf{H}})$ by $\boldsymbol{\lambda}^*$ and $\boldsymbol{\delta}^*_{\mathbf{H}}$. Given $\boldsymbol{\lambda} = \boldsymbol{\lambda}^*$ and $\boldsymbol{\delta}_{\mathbf{H}} = \boldsymbol{\delta}^*_{\mathbf{H}}$ in Eq. (8.65), we solve for $\boldsymbol{t}^*$ and $\boldsymbol{\phi}^*$ as shown in Eqs. (8.61) and (8.62), respectively.

Using the analyses similar to the proof for Theorem 15, we can show that if there does not exist $\boldsymbol{\lambda}^*$ such that the equality in Eq. (8.60) holds for all n, the feasible solution to problem **VIII-*A4*** does not exist. In contrast, applying the similar derivations as used in the proof of Theorem 15, we obtain that if the optimal solution to problem **VIII-*A4*** exists, the duality gap between $\widetilde{\mathcal{J}}_{A4}(\boldsymbol{\lambda}^*,\boldsymbol{\delta}^*_{\mathbf{H}})$ and the optimal

objective value of **VIII-*A*4** is zero. Correspondingly, the optimal solution to **VIII-*A*4** is given by Eqs. (8.61) and (8.62), where $\boldsymbol{\delta}_{\mathbf{H}}^{*}$ needs to satisfy Eq. (8.59) and $\boldsymbol{\lambda}^{*}$ results in the equality in Eq. (8.60). Then, Theorem 16 follows. □

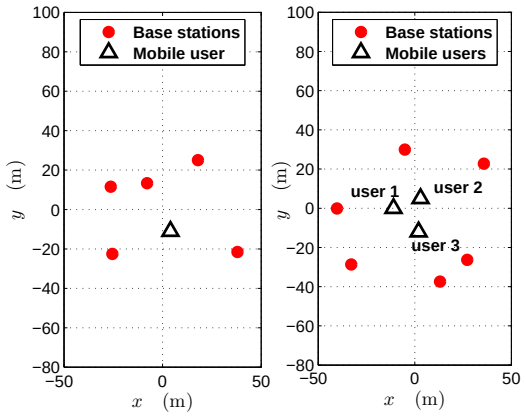
Note that the general closed-form expressions for the optimal Lagrangian multipliers of problem **VIII-*A*4** are also hard to obtain. However, we can apply the gradient descent method similar to problem **VIII-*A*3** (see Section E-3.3) for tracking the optimal Lagrangian multipliers.

F. Simulation Evaluations

We use simulations to evaluate the performances of our proposed BS selection schemes for distributed MIMO links. In particular, we simulate the distributed MIMO transmissions within a 250 m × 250 m square region, where the coordinates of its four vertices are given by $(125, 125)$ m, $(-125, 125)$ m, $(125, -125)$ m , and $(-125, -125)$ m, respectively. For the single-user case, the BS's deployment and the mobile user's position are shown in Fig. 46(a), where $K_{\text{bs}} = 5$. For the multi-user case, the BS's deployment and the mobile users' positions are given by Fig. 46(b), where $K_{\text{bs}} = 6$. Moreover, we set $T = 10$ ms and $B = 10^5$ Hz throughout the simulations.

In the simulations, we employ the following average power degradation propagation model [94]. For the given reference distance d_{ref}, if the transmission distance, denoted by d, is smaller than or equal to d_{ref}, the free-space propagation model is used; if the $d > d_{\text{ref}}$, the power degradation is proportional to $(d/d_{\text{ref}})^{\eta}$, where η is the path loss exponent and typically varies from 2 to 6 indoor environments without LOS [94]. Accordingly, the mean $\overline{h}_{n,m}$ of $H_{n,m}$'s elements can be determined by

$$\overline{h}_{n,m} = \begin{cases} G d_{n,m}^{-2}, & \text{if } d_{n,m} \in (0, d_{\text{ref}}]; \\ G \left(\frac{d_{\text{ref}}}{d_{n,m}}\right)^{\eta}, & \text{if } d_{n,m} \in (d_{\text{ref}}, \infty), \end{cases}$$

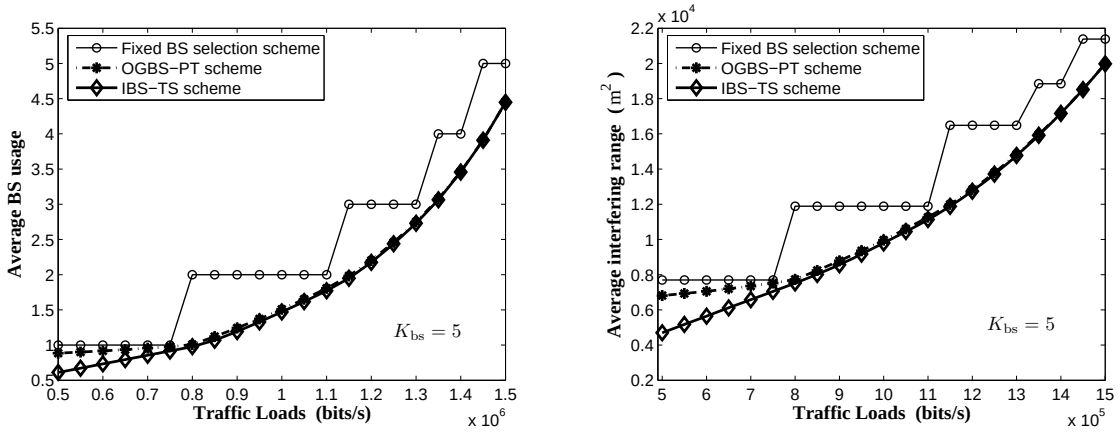


(a) (b)

Fig. 46. The deployment of BS's and the positions of mobile users. (a) Single-user case: $K_{\rm bs} = 5$ BS's, whose coordinates are $(37.96, -21.56)$, $(-7.83, 13.33)$, $(25.50, -22.49)$, $(17.98, 25.00)$, and $(-26.34, 11.62)$; the mobile station's coordinates are $(4, -11)$ (b) Multi-user case: $K_{\rm bs} = 6$, whose coordinates are $(35.77, 22.69)$, $(13.06, -37.45)$, $(27.15, -26.33)$, $(-40.28, -0.14)$, $(-32.86, -28.65)$, and $(-5.10, 29.98)$; $K_{\rm mu} = 3$, whose coordinates are $(-11, 0)$, $(3, 5)$, and $(2, -12)$.

where $d_{n,m}$ is the distance between the mobile user and the mth BS, and G is aggregate power gain generated by the antenna and other factors. In simulations, we set $d_{\rm ref} = 1$ [94] and $\eta = 3$. Furthermore, we set $\mathcal{P}_{\rm ref} = 4$ and further select G such that $\overline{h}_{n,m} = 0$ dB at $d_{n,m} = 50$ m. Also, we set $\sigma^2_{\rm th} = 0$ dB for the evaluation of average interfering range (see Section B-3 for its definition).

Figures 47(a) and 47(b) plot the average BS usage and the average interfering range, respectively, versus the incoming traffic load for the single-user case. As shown in Fig. 47(a), our proposed IBS-TS and OGBS-PT schemes both effectively decrease the average BS usage and the interfering range as compared to the fixed selection scheme. This is expected because our proposed BS-selections can adaptively adjust BS selection in each fading state based on the CSI, traffic load, and the QoS requirements. In contrast, the resulted average BS usage and the interfering range by

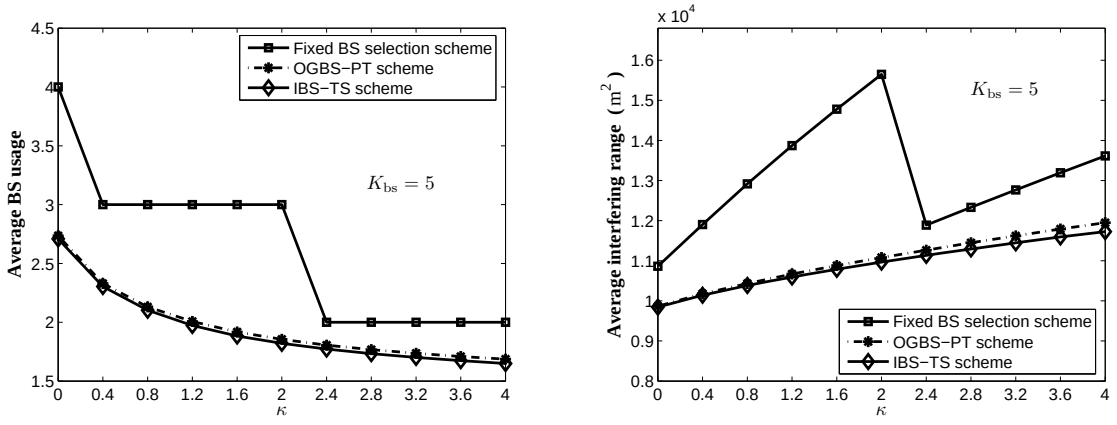


(a) (b)

Fig. 47. Single-user case, where $K_{\rm bs} = 5$, each BS has two transmit antennas, the mobile user has two receive antennas, and $\kappa = 2.4$. The delay bound and its violation probability requirement are given by $D_{\rm th}^{(1)} = 50$ ms and $\xi_1 = 10^{-4}$, respectively. (a) Average BS usage versus traffic load. (b) Average interfering range versus traffic load.

applying the fixed BS-selection scheme cannot smoothly vary with traffic load, which may cause unnecessary BS usage wither high power consumption and thus larger interferences to the entire wireless network. We also observe from Fig. 47 that the IBS-TS scheme needs less BS usage to support the incoming traffic under the specified QoS requirements and therefore generates lower interferences accordingly, verifying our discussions in Sections D-1 and D-2. However, we can see that the performance differences between the IBS-TS and OGBS-PT schemes are little, especially when the incoming traffic load is relatively high. Then, since the OGBS-PT scheme is easier to implement (see Section D-4), the OGBS-PT scheme is more promising for realistic systems than the IBS-TS scheme.

Figure 48 depicts the dynamics of the average BS usage and the interfering range as functions of κ. When κ gets larger, the power budget used for distributed MIMO transmission in each fading state is increased (see Eq. (8.3)). As a result, the average BS usage is reduced correspondingly for all schemes, as illustrated in Fig. 48(a).

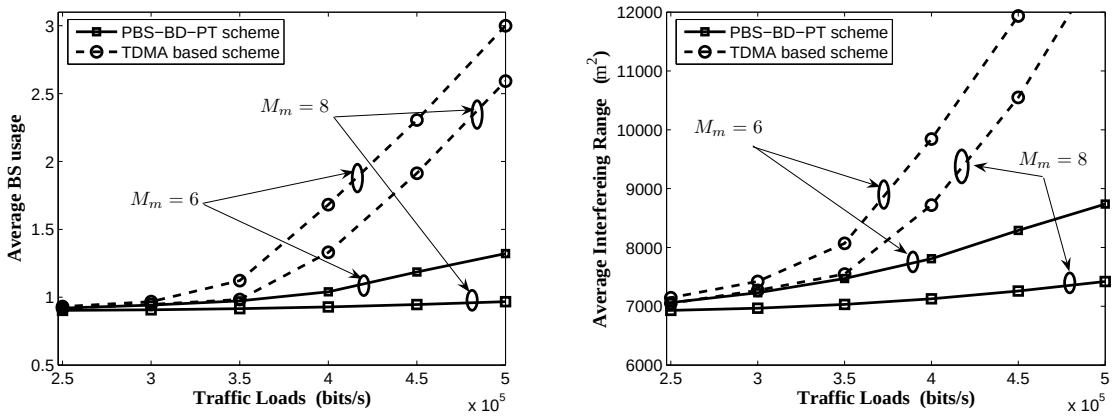


(a) (b)

Fig. 48. Single-user case, where $K_{\rm bs} = 5$, each BS has two transmit antennas, the mobile user has two receive antennas, and $\overline{C}_1 = 1.1$ Mbits/s. The delay bound and its violation probability requirement are given by $D_{\rm th}^{(1)} = 50$ ms and $\xi_1 = 10^{-4}$, respectively. (a) Average BS usage versus κ. (b) Average interfering range versus κ.

Although the BS usage decreases as κ increases, we can observe from Fig. 48(b) that the interfering range is enlarged, which is also expected since higher transmit power is used. Again, Fig. 48 demonstrates the inflexibility of the fixed BS-selection and the significant reduction of BS usage and interferences generated by using our proposed IBS-TS and OGBS-PT schemes.

Figure 49 compares the average BS usage and interfering range between our proposed PBS-BD-PT scheme and the TDMA based scheme for multi-user case under various system parameters. Fig. 49 shows that as the traffic load increases, both scheme's average BS usages and interfering range increase to satisfy the specified QoS constraints for the incoming traffic. However, the TDMA based scheme's BS usage increases much more rapidly than our proposed PBS-BD-PT scheme. This is because block diagonalization for multi-user distributed MIMO communication can effectively take advantage of space multiplexing gain in removing the cross-interferences among all mobile users, and thus can achieve high throughput. In contrast, the TDMA



(a) (b)

Fig. 49. Multi-user case, where $K_{\rm bs} = 6$, $K_{\rm mu} = 3$, $N_n = 2$ for all $n = 1, 2, \ldots, K_{\rm mu}$, $\kappa = 1.2$, and M_m is the same for all users; $\xi_n = 10^{-4}$ for all n, $D_{\rm th}^{(1)} = D_{\rm th}^{(2)} = 50$ ms, and $D_{\rm th}^{(3)} = 40$ ms. (a) Average BS usage versus traffic load. (b) Average interfering range versus traffic load.

based scheme simply assigns orthogonal time slots to mobile users, which severely degrades the sustainable traffic load. Moreover, Fig. 49 shows that as the traffic load becomes smaller, the superiority of our proposed PBS-BD-PT scheme over the TDMA based scheme gradually vanishes, implying that TDMA works well for low traffic load. Fig. 49 also illustrates the impact of the number M_m of transmit antennas at each BS on supporting the traffic load with the specified QoS requirements. As depicted in Fig. 49, higher M_m can significantly decrease the average BS usage and the interfering range, especially for our PBS-BD-PT scheme. Given $M_m = 8$, we can see that the average BS usage and the interfering range for our PBS-BD-PT scheme only need to increase slightly as traffic load gets larger. This is because the block diagonalization typically needs high space multiplexing degree such that multiple user can coexist while all achieving high system throughput. Further comparing Figs. 49(a) and 49(b), we observe that the impact from the traffic load on the interfering range is higher than that on the BS usage.

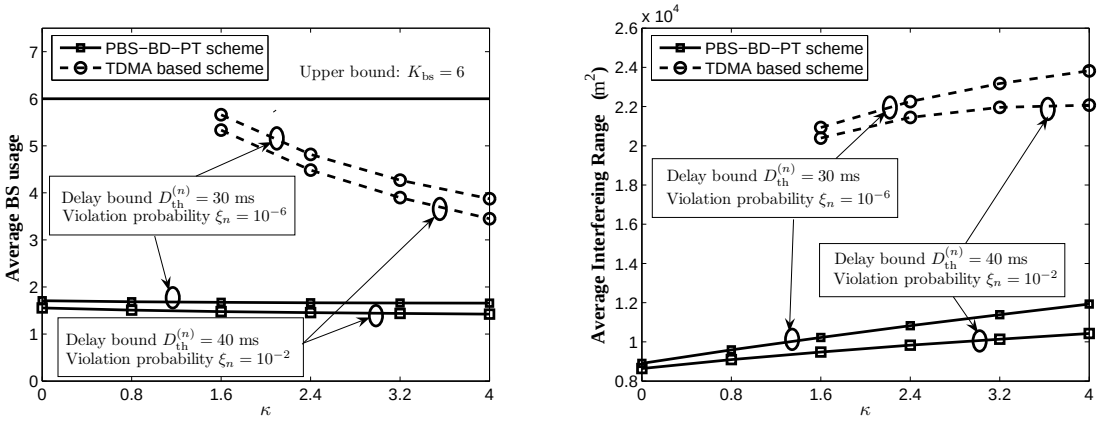


Fig. 50. Multi-user case, where $K_{\rm bs} = 6$, $K_{\rm mu} = 3$, and $\overline{C}_n = 0.6$ Mbits/s, $N_n = 2$ for all $n = 1, 2, \ldots, K_{\rm mu}$ and $M_m = 6$ for all $m = 1, 2, \ldots, K_{\rm bs}$; the delay bounds and their corresponding violation probability requirements are the same for all users. (a) Average BS usage versus κ. (b) Average interfering range versus κ.

Figure 50 illustrates our proposed PBS-BD-PT scheme and TDMA based scheme's performances versus κ under diverse QoS requirements. Similar to the single-user case, the average BS usage deceases but the interfering range increases as κ gets larger. The average BS usage and the interfering range of the PBS-BD-PT scheme are much smaller than the TDMA based scheme. Also, the PBS-BD-PT scheme's BS usage and interfering range is less sensitive to the increasing of power budget as compared to the TDMA based scheme. Figs. 50 also compares the performances under various delay-QoS requirements. As shown in Fig. 50, lower delay bound and smaller violation probability threshold, implying more stringent delay-QoS requirements, cause more BS usage and thus larger interfering range. This is because in order to satisfy more stringent QoS requirements, more BS's need to involve the cooperative downlink transmissions to achieve high system throughput for all mobile users. This also demonstrates that our proposed schemes can effectively adjust the transmission strategy to adapt to the specified QoS requirements. In addition, Figs. 50(a)

and 50(b) shows that the impact of κ is more significant on the average interfering range than the average BS usage.

G. Summary

We proposed the QoS-aware BS-selection schemes for the distributed wireless MIMO downlink to minimize the BS usages and to reduce the interfering range caused by the distributed MIMO system, while satisfying diverse statistical delay-QoS constraints characterized by the delay-bound violation probability and the effective capacity technique. For the single-user scenario, we developed the scheme using the incremental BS selection and time-sharing strategy and proposed the scheme employing the ordered-gain based BS-selection and probabilistic transmission strategy. The former scheme archives better performance, while the latter scheme is easier to implement. For the multi-user scenario, we developed the joint priority BS-selection, block-diagonalization precoding, and probabilistic transmission scheme. We also studied the TDMA based BS selection scheme for multi-user link. Abundant simulation results show that our proposed schemes can effectively support the incoming traffic load under the specified QoS requirements and significantly outperforms baseline schemes in terms of minimizing the average BS usage and the interfering range.

CHAPTER IX

QUEUE-AWARE SPECTRUM SENSING FOR INTERFERENCE CONSTRAINED TRANSMISSIONS IN COGNITIVE RADIO NETWORKS

A. Introduction

Recently, cognitive radio networks have attracted more and more research attentions due to the dramatically increasing demands on wireless services and considerable underutilization of the licensed spectrums [48]. To effectively improve the utilization of wireless spectrum resources, in cognitive radio networks unlicensed users (secondary users) are allowed to use the licensed spectrums when the licensed users (primary users) do not occupy these bandwidths [48–51].

Spectrum sensing is one of the key techniques in cognitive radio networks, through which the secondary users (SUs) can detect the occupancy statuses of the channels licensed to the primary users (PUs). Among various spectrum sensing techniques, energy detection with the threshold-based decision is widely applied [50–52]. The traditional energy-detection based schemes typically compare the received energy with the fixed threshold to decide whether the spectrum is occupied. The threshold is selected such that the probability of causing interference to the PUs is upper-bounded. However, the traditional energy-detection based spectrum-sensing strategy cannot effectively satisfy the statistical quality-of-service (QoS) requirements [3, 8, 53], such as the queue-length-bound violation probability or buffer-overflow probability. This is because the traditional energy-detection threshold is not aware of the queueing status of SUs. In order to effectively decrease the queue-length-bound violation or buffer-overflow probabilities, we need to take into consideration the queue length at the sender of SUs. Specifically, when the queue length is small, the current data

traffic burden is usually not heavy as compared to the channel service capability. For this case, the SUs can use a relatively conservative strategy, e.g., a lower energy threshold for the decision of spectrum occupancy, which causes less interferences to PUs. When the queue length gets larger, SUs apply more aggressive strategies to reduce the chances of buffer overflow or queue-length-bound violation. Furthermore, the adaptation policy need to be carefully designed such that the overall interferences to the PUs do not exceed the acceptable level.

To overcome the above problem, we propose the queue-aware spectrum sensing schemes for interference-constrained opportunistic transmissions of SUs in cognitive radio networks. Specifically, we employ the energy detection to detect the SU's spectrum usage status. The energy threshold to decide the occupancy of the spectrum is dynamically regulated as a function of the queue length at the sender of SUs. We design the dynamic threshold control policies for the scenarios with infinite and finite queue buffer sizes, respectively, which can effectively satisfy the statistical QoS constraints, while upper-bounding the interference probability to the PU's transmissions.

The rest of the chapter is organized as follows. Section B describes the system model. Section C proposes the queue-aware spectrum sensing framework for SUs' transmissions. Section D derives the queue-aware spectrum sensing schemes with infinite and finite queue buffer sizes, respectively. Section E presents the simulation evaluations. The chapter concludes with Section F.

B. System Model

Consider a cognitive radio network consisting of a SU pair (a secondary sender and a secondary receiver) and a PU, as shown in Fig. 51. The bandwidth of the licensed spectrum is equal to B. The secondary sender uses spectrum sensing to detect the

occupancy of the licensed spectrum. Due to the existence of thermal noises, the spectrum sensing result cannot be perfectly accurate and thus interferences to the PU cannot be completely avoided. Then, we use the miss-detection probability (to be detailed in Section C) to characterize the interference constraints imposed to the SUs. Specifically, we require that the miss-detection probability cannot exceed the specified threshold denoted by $\overline{P}_{\mathrm{m}}$, implying that the interference probability to the PU is also upper-bounded by $\overline{P}_{\mathrm{m}}$. The detailed system descriptions are provided in the following sections.

1. The Primary User's Transmission Behaviors

The licensed spectrum's occupance status of the PU is modeled by the discrete-time two-state Markov-Chain Model [49]. In particular, the time axis is divided into consecutive time frames each with the fixed length T. Within the t-th time frame, $t = 1, 2, \ldots$, the channel occupancy by the PU is denoted by a random variable $\mathcal{O}[t] \in \{0, 1\}$. In particular, $\mathcal{O}[t] = 1$ implies that the channel is currently being used by the PU while $\mathcal{O}[t] = 0$ suggests the idle state. The probability transition matrix of the two-state Markov Chain is given by

$$\mathbf{G} = \begin{bmatrix} \beta & 1-\beta \\ 1-\alpha & \alpha \end{bmatrix}, \tag{9.1}$$

where the element $G_{i,j}$ of $\mathbf{G}$ on the ith row ($i = 0, 1$) and jth column ($j = 0, 1$) represents the probability of $\mathcal{O}[t+1] = j$ given $\mathcal{O}[t] = i$.

2. Wireless Channel Model

The PU and the secondary sender both use constant transmit power. Then, the wireless channels for the corresponding wireless links can be characterized by the received

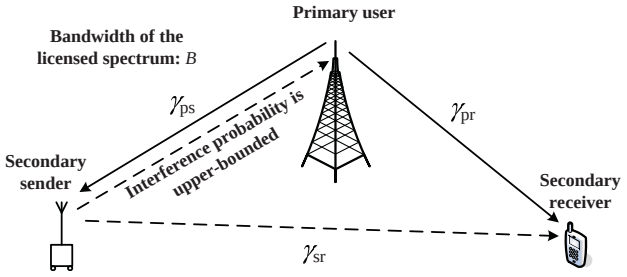


Fig. 51. System model of an interference-constrained cognitive radio network.

SNR. The instantaneous SNR's for the links of primary-user-to-secondary-sender, primary-user-to-secondary-receiver, and secondary-sender-to-secondary-receiver, are denoted by $\gamma_{\rm ps}$, $\gamma_{\rm pr}$, and $\gamma_{\rm sr}$, respectively, which are depicted in Fig. 51. These SNR's are independent and follows Rayleigh fading model. Moreover, $\gamma_{\rm ps}$, $\gamma_{\rm pr}$, and $\gamma_{\rm sr}$ follow the block-fading model, where they remain unchanged within a time frame, but varies independently from frame to frame. In addition, we suppose that the secondary sender knows the distribution of $\gamma_{\rm ps}$.

The PU and the secondary sender may transmit data in the same time frame. Then, the signal-to-interference-plus-noise ratio (SINR) received at the secondary receiver, denoted by $\gamma_{\rm r}$, is determined by

$$\gamma_{\rm r} = \begin{cases} \gamma_{\rm sr}, & \text{if the PU does not transmit;} \\ \frac{\gamma_{\rm sr}}{1+\gamma_{\rm pr}}, & \text{otherwise;} \end{cases} \tag{9.2}$$

We assume that the secondary receiver can perfectly estimate $\gamma_{\rm r}$ and feed $\gamma_{\rm r}$ back to the secondary sender in the beginning of each time frame. However, we suppose that the secondary sender does not use the information of $\gamma_{\rm r}$ for spectrum sensing. The purpose of the assumption for perfect SINR estimation is to identify the theoretic performance bound and simplify analyses. For the impacts of imperfect channel estimation on the performances of wireless communications links, researchers can

refer to [95, 96].

3. QoS Requirements for the Secondary Users

Due to the time-varying wireless channel and opportunistic spectrum access, the deterministic QoS metrics such as hard delay bound and hard queue-length bound are difficulty to guarantee. In contrast, we focus on statistical QoS metrics such as the queue-length-bound violation probability and buffer-overflow probability [3,8,53]. The queueing system for the secondary sender is depicted in Fig. 52. In particular, the secondary sender is responsible for transmitting a data stream to the secondary receiver. The secondary sender maintains a queue to buffer the arrival data, where the queue buffer size is denoted by L nats. The discrete-time arrival-rate process of the data stream is denoted by $A[t]$ (nats/frame), and the departure-rate process is denote by $R[t]$ (nats/frame), where

$$R[t] = \begin{cases} BT\log(1+\gamma_{\rm r}), & \text{if transmitting;} \\ 0, & \text{if not transmitting.} \end{cases} \tag{9.3}$$

In Eq. (9.3), $BT\log(1+\gamma_{\rm r})$ is the Shannon capacity given the SINR $\gamma_{\rm r}$. In this chapter, we assume that $A[t]$ is the constant-rate process and focus on the impacts of spectrum sensing techniques on the transmission capability of the opportunistic spectrum access in cognitive radio networks. Next, we introduce the QoS requirements for the scenarios with $L = \infty$ and $L < \infty$, respectively.

a. $L = \infty$

Since no buffer flow happens in this case, we focus on the queue-length-bound violation probability. The queue length and the required queue-length bound are denoted by Q and $Q_{\rm th}$, respectively. The probability of violating this queue-length bound

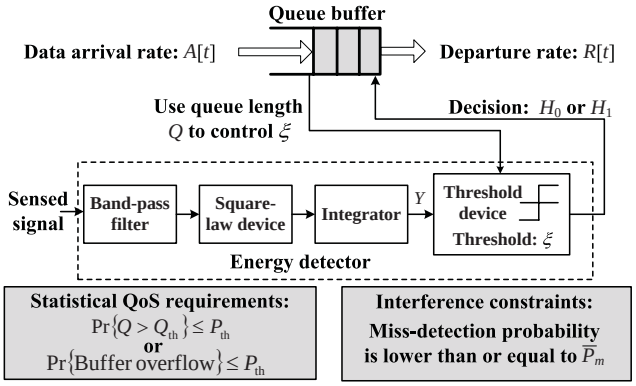


Fig. 52. Queue-aware spectrum sensing framework for the SUs.

cannot exceed a specified threshold P_{th}, i.e., the inequality given by

$$\Pr\{Q > Q_{\text{th}}\} \leq P_{\text{th}} \tag{9.4}$$

needs to hold.

b. $L < \infty$

For this scenario, we require that the buffer-overflow probability satisfies the following inequality:

$$\Pr\{\text{buffer overflow}\} \leq P_{\text{th}}. \tag{9.5}$$

C. Queue-Aware Spectrum Sensing Framework for the Secondary Users

We propose a queue-aware spectrum-sensing framework for the secondary sender, as illustrated in Fig. 52. We employ the energy detector for spectrum sensing, which is marked with the dash-lined box in Fig. 52. Specifically, the secondary sender first passes the sensed signal through the band-pass filter to remove the out-of-band noises. Then, the square-law device and integrator are applied to obtain the energy of the sensed signal within the certain observation interval. For more details of the

energy detector, please refer to [50, 52]. We mainly concentrate on the output of the integrator, which is denoted by Y. The energy detector uses Y to test two hypotheses H_0 and H_1, where H_1 represents that the PU is transmitting data and H_0 denotes the idle state of the PU. Moreover, Y follows the distributions below [50, 52]:

$$Y \sim \begin{cases} \chi^2_{2m}, & H_0 \text{ (the PU is transmitting data);} \\ \chi^2_{2m}(2\gamma_{\rm ps}), & H_1 \text{ (the PU is idle),} \end{cases} \tag{9.6}$$

where χ^2_{2m} denotes the central chi-square distribution and $\chi^2_{2m}(x)$ represents the non-central chi-square distributions with the non-centrality parameter equal to x. In the above equation, $2m$ is the degree of freedoms. More specifically, m represents the number of samples from the energy detector [52]. Then, we use a threshold ξ to decide the spectrum occupancy as follows:

$$Y \underset{H_0}{\overset{H_1}{\gtrless}} \xi. \tag{9.7}$$

When the decision is H_1, the secondary sender transmits data in the buffer with the rate determined by Eq. (9.3), which will cause interference to the PU. Accordingly, given ξ the miss-detection probability, denoted by $P_{\rm m}(\xi)$, is defined by $P_{\rm m}(\xi) \triangleq \Pr\{Y < \xi \,|\, H_1\}$. Then, we can use the miss-detection probability to characterize the interference probability to the PU.

The traditional energy detector typically uses the fixed ξ to guarantee the miss-detection probability equal to the specified threshold. In contrast, we dynamically regulate the threshold ξ as a function of the queue length Q at the secondary sender, as shown in Fig. 52. Denoting the dynamic threshold by $\xi(Q)$, the average miss-detection probability is then equal to $\mathbb{E}_Q\{P_{\rm m}(\xi(Q))\} = \mathbb{E}_Q\{\Pr\{Y < \xi(Q) \,|\, H_1\}\}$, where $\mathbb{E}_Q\{\cdot\}$ denotes the expectation over Q. The principle of our dynamic threshold based scheme is explained as follows. On the one hand, when the queue length

is small, implying that the current traffic burden is not heavy as compared to the channel service capability. Accordingly, the secondary sender can use a relatively conservative strategy, i.e., a smaller threshold, resulting in a small miss-detection probability and imposing less interferences to the PU. On the other hand, as the queue length becomes large, the secondary sender uses a more aggressive strategy, i.e., a larger threshold for energy detection. Moreover, the threshold adjustment rule needs to be carefully designed such that the average interference probability is upper-bounded by $\overline{P}_{\mathrm{m}}$ while satisfying the statistical QoS requirements.

D. Queue-Aware Spectrum Sensing Schemes for the Secondary Sender

Following the framework described in Section C, we need to identify the dynamic threshold $\xi(Q)$ as a function of queue length Q. Given a fixed threshold ξ, the detection probability, denoted by $P_{\mathrm{d}}(\xi)$, in Rayleigh fading channel was given [50, 52] by

$$P_{\mathrm{d}}(\xi) = e^{-\frac{\xi}{2}} \sum_{k=0}^{m-2} \frac{1}{k\,!} \left(\frac{\xi}{2}\right)^k + \left(\frac{1+\overline{\gamma}}{\overline{\gamma}}\right)^{m-1} \times \left(e^{-\frac{\xi}{2(1+\overline{\gamma})}} - e^{-\frac{\xi}{2}} \sum_{k=0}^{m-2} \frac{1}{k\,!} \left(\frac{\xi\overline{\gamma}}{2(1+\overline{\gamma})}\right)^k \right),$$

where $P_{\mathrm{m}}(\xi) = 1 - P_{\mathrm{d}}(\xi)$. When $\xi(Q)$ is applied, the average miss-detection probability needs to be upper-bounded by $\overline{P}_{\mathrm{m}}$. To increase the chances of the spectrum access for SUs, we set the targeted average miss-detection probability just equal to $\overline{P}_{\mathrm{m}}$. Thus, we have

$$\int_0^\infty \mathcal{P}_{\mathrm{m}}(q) g(q) dq = \overline{P}_{\mathrm{m}}, \tag{9.8}$$

where $g(q)$ denotes the probability density function (pdf) of the queue length distribution, and $\mathcal{P}_{\mathrm{m}}(Q) \triangleq P_{\mathrm{m}}(\xi(Q))$ represents the designed miss-detection probability given Q. Since $P_{\mathrm{d}}(\xi)$ is a monotonically increasing function of ξ, so is $P_{\mathrm{m}}(\xi)$. Thus,

the inverse function of $P_{\rm m}(\xi)$, denoted by $\widetilde{\xi}(P_{\rm m})$, exists, which can be readily obtained through numerical searching techniques. Then, we can focus on controlling $\mathcal{P}_{\rm m}(Q)$ to satisfy Eq. (9.8). After determining $\mathcal{P}_{\rm m}(Q)$, we can obtain the dynamic threshold for energy detection as follows:

$$\xi(Q) = \widetilde{\xi}(\mathcal{P}_{\rm m}(Q)). \tag{9.9}$$

Following that discussion Chapter II, the queue-length-bound violation probability of a dynamic queueing system can be approximated by

$$\Pr\{Q > q\} \approx e^{-\theta q}. \tag{9.10}$$

Accordingly, we can write the pdf $g(q)$ of Q as

$$g(q) \approx \theta e^{-\theta q}. \tag{9.11}$$

The parameter $\theta > 0$ describes the exponentially decaying speed of $\Pr\{Q > q\}$ as q increases, which is called the QoS exponent [8, 53]. Having obtained the pdf of Q, we design the queue-aware sensing schemes for scenarios with $L = \infty$ and $L < \infty$, respectively, in the following sections.

1. The Scenario with Infinite Buffer Size

Based on Eqs. (9.4) and (9.10), in order to guarantee the QoS requirements, the QoS exponent θ needs to satisfy

$$\theta \geq -\frac{1}{Q_{\rm th}} \log(P_{\rm th}). \tag{9.12}$$

Then, we use the boundary value $\theta = -\log(P_{\rm th})/Q_{\rm th}$ as the guidance to design our queue-aware spectrum sensing scheme. We further require that control function $\mathcal{P}_{\rm m}(Q)$ has the following properties: (i) $\mathcal{P}_{\rm m}(Q)$ is an increase function of Q; (ii) since

$\mathcal{P}_{\rm m}(Q)$ is essentially a probability, it must be upper-bounded by 1; (*iii*) $\mathcal{P}_{\rm m}(Q)$ is a continuous function of Q. Note that property (*i*) follows the principles proposed in Section C. The smaller queue length corresponds to the more conservative strategy, while the larger queue length generates the more aggressive strategy. Property (*iii*) is applied to obtain the flexible dynamic-threshold policies. Following the above properties, we design a control function $\mathcal{P}_{\rm m}(Q)$ given by

$$\mathcal{P}_{\rm m}(Q) = \begin{cases} \dfrac{Q}{\phi}, & \text{if } 0 \leq Q \leq \phi; \\ 1, & \text{if } Q > \phi, \end{cases} \tag{9.13}$$

where $\phi > 0$. We choose the linear function in that given $\xi(Q)$ in a time frame, the average queue length increment/decrement is linear to the miss-detection probability. Under this control policy, when the queue length is larger than or equal to ϕ, the SU transmits data regardless of the current status of the PU to decrease the queue-length-bound violation probability. When the queue length is equal to 0, the miss-detection probability will also become 0, implying that the secondary sender does not attempt to use the spectrum in this case. If the queue length fall in the interval given by $(0, \phi)$, the miss-detection probability for the current time frame varies linearly to Q.

Plugging Eq. (9.13) into Eq. (9.8) we obtain

$$\begin{aligned} \int_0^\infty \mathcal{P}_{\rm m}(q) g(q) dq &= \int_0^\phi \frac{\theta q}{\phi} e^{-\theta q} \int_\phi^\infty \theta e^{-\theta q} \\ &= \frac{1}{\theta\phi}\left(1 - e^{-\theta\phi}\right) = \overline{P}_{\rm m}. \end{aligned} \tag{9.14}$$

Solving this equation, we get the analytical expression of ϕ as follows:

$$\phi = \frac{1}{\theta \overline{P}_{\rm m}} \left[1 + \overline{P}_{\rm m} \mathcal{W}\left(-\frac{1}{\overline{P}_{\rm m}} e^{-\frac{1}{\overline{P}_{\rm m}}} \right) \right], \tag{9.15}$$

where $\mathcal{W}(\cdot)$ is the Lambert W function [74] which is known as the inverse function of

$Z(W) = We^W$. Applying ϕ into Eq. (9.13), we get the control policy $\mathcal{P}_{\rm m}(Q)$ for the scenario with infinite buffer size. The corresponding dynamic-threshold policy can be obtained through Eq. (9.9).

Although we use the boundary θ in Eq. (9.12) to design our queue-aware spectrum sensing scheme, the actual QoS exponent under our proposed policy will vary with the traffic load. However, through simulations in Section E we will show that our proposed schemes can support higher traffic loads than the traditional energy-detection based scheme under the same QoS requirement and interference constraint.

2. The Scenario with Finite Buffer Size

When L is finite, the maximum queue length is equal to L, and the arrival data will be dropped if the queue is full. Based on the theory of statistical QoS, the queue-length distribution can be characterized by

$$\begin{cases} g(q) \approx \theta e^{-\theta q}, & \text{if } q < L; \\ g(q) = 0, & \text{if } q > L; \\ \Pr\{Q = q\} \approx e^{-\theta q}, & \text{if } q = L. \end{cases} \tag{9.16}$$

Accordingly, the buffer-overflow probability can be approximated by $\Pr\{\text{Buffer overflow}\} \approx 1 - \Pr\{Q < L\}$. Then in order to satisfy Eq. (9.16), the QoS exponent θ needs to satisfy

$$\theta \geq -\frac{1}{L}\log(P_{\rm th}). \tag{9.17}$$

We also select the boundary value $\theta = -\log(P_{\rm th})/L$ to design our dynamic-threshold policy.

When the queue length approaches L, we need to increase the chances for transmissions, such that the buffer-overflow probability can be effectively decreased. Then,

a natural strategy is to set $\mathcal{P}_{\rm m}(L) = 1$. Accordingly, when the queue length decreases to 0, we set $\mathcal{P}_{\rm m}(0) = 0$. Similar to the design for the scenario with infinite buffer size, we let $\mathcal{P}_{\rm m}(Q)$ linearly vary with Q for $Q \in [0, L]$ and thus get

$$\mathcal{P}_{\rm m}(Q) = \frac{Q}{L}, \quad \forall\ 0 \leq Q \leq L. \tag{9.18}$$

However, with the policy of Eq. (9.18), the average miss-detection probability is not equal to the design target $\overline{P}_{\rm m}$ under the given QoS exponent θ, which causes either too much interferences to the PU or insufficient utilization of the wireless spectrum. Accordingly, we need to modify the policy given by Eq. (9.18) for the cases with $L < \phi$ and $L \geq \phi$, respectively, where ϕ is determined by Eq. (9.15).

a. For the case of $L < \phi$

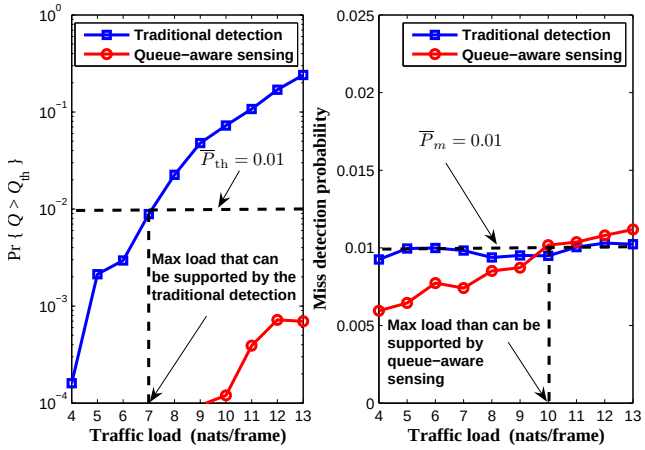
In this case, when applying the policy given by Eq. (9.18), the integration result on the left-hand side of Eq. (9.8) is larger than $\overline{P}_{\rm m}$. To satisfy the interference constraint, we introduce a multiplier $\mu \in (0, 1]$ into Eq. (9.18) and derive a new control policy as follows:

$$\mathcal{P}_{\rm m}(Q) = \frac{\mu Q}{L}, \quad \forall\ 0 \leq Q \leq L. \tag{9.19}$$

Plugging Eqs. (9.19) and (9.16) into Eq. (9.8) and solving for μ, we obtain the analytical solution to μ as:

$$\mu = \frac{\theta L \overline{P}_{\rm m} e^{\theta L}}{-1 + e^{\theta L}}. \tag{9.20}$$

Note that under this new policy, the maximum detection probability in any time frame is smaller than 1, which is forced by the interference constraints.



(a) (b)

Fig. 53. The SU's transmission performance with the *infinite* buffer size, where $\Pr\{Q > Q_{\rm th}\} \leq P_{\rm th} = 0.01$, $Q_{\rm th} = 500$ nats, and $\overline{P}_{\rm m} = 0.01$. (a) The queue-length-bound violation probability $\Pr\{Q > Q_{\rm th}\}$ versus the traffic load $\overline{A}$. (b) The miss-detection probability $P_{\rm m}$ versus the traffic load $\overline{A}$.

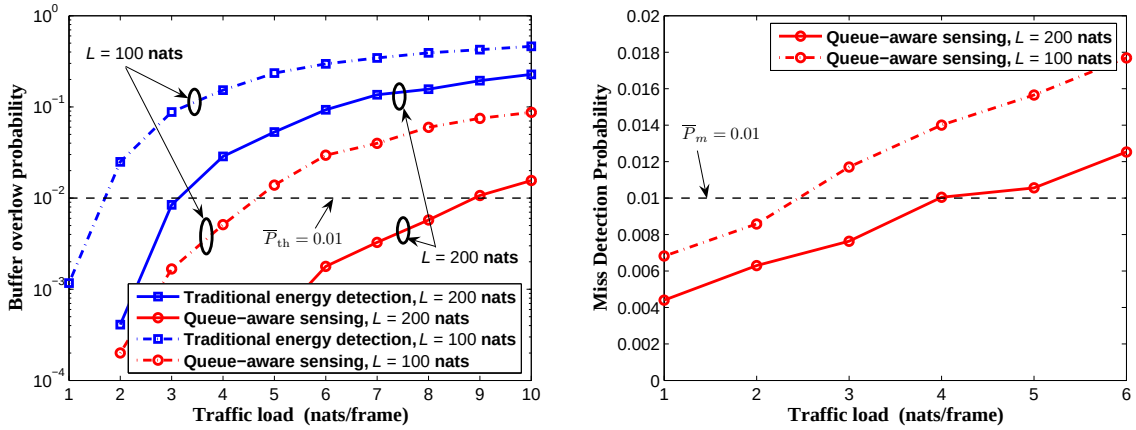
b. For the case of $L \geq \phi$

If applying Eq. (9.18) in this case, the integration results on the left-hand side of Eq. (9.8) is lower than $\overline{P}_{\rm m}$, which decreases the spectrum utilization. Then, further increasing $\mathcal{P}_{\rm m}(Q)$ for $Q < L$ and following the similar design principles applied in Section D-1, we get the dynamic threshold policy in this case as follows:

$$\mathcal{P}_{\rm m}(Q) = \begin{cases} \dfrac{Q}{\phi}, & \text{if } 0 \leq Q \leq \phi; \\ 1, & \text{if } \phi < Q \leq L. \end{cases} \tag{9.21}$$

E. Simulation Evaluations

We use simulations to evaluate the performance of our proposed queue-aware sensing schemes. For comparative analyses, we also simulate the traditional energy-detection based scheme. In the simulations, we set bandwidth of the licensed spectrum as



(a) (b)

Fig. 54. The SU's transmission performance with the *finite* buffer size, where $P_{\rm th} = 0.01$ and $\overline{P}_{\rm m} = 0.01$. (a) The buffer-overflow probability versus the traffic load $\overline{A}$. (b) The average miss-detection probability versus the traffic load $\overline{A}$.

$B = 10^5$ Hz and the time-frame length as $T = 1$ ms. The probability transition parameters in Eq. (9.1) are set to $\alpha = 0.5$ and $\beta = 0.7$. For wireless channels, the average SNR's of $\gamma_{\rm ps}$, $\gamma_{\rm pr}$, and $\gamma_{\rm pr}$ are given by $\overline{\gamma}_{\rm ps} = 10$ dB, $\overline{\gamma}_{\rm pr} = 10$ dB, and $\overline{\gamma}_{\rm sr} = 15$ dB, respectively. The arrival traffic $A[t]$ for the SU has a constant arrival rate, denoted by $\overline{A}$ (nats/frame), which represents the traffic load. Moreover, the miss-detection probability needs to be upper-bounded by $\overline{P}_{\rm m} = 0.01$. Figs. 53 and 54 simulate the scenarios with infinite and finite buffer size, respectively.

Figures. 53(a) and 53(b) plot the queue-length-bound violation probability and miss-detection probability, respectively, versus the traffic load, where $P_{\rm th} = 0.01$ and $Q_{\rm th} = 500$ nats. Fig. 53(a) shows that the queue-length-bound violation probability increases as the traffic load gets larger. We can observe from Fig. 53(a) that our proposed scheme generates much lower queue-length-bound violation probability than the traditional energy detection scheme. From Fig. 53(b), we can see that the miss-detection probability of the traditional energy detection scheme does not vary with

traffic loads, which is expected in that the threshold for the hypotheses test does not vary with queue length. The miss-detection probability of our proposed queue-aware sensing scheme increases as the traffic load gets large. This is also expected because when the traffic load is higher, the queue-length is usually larger, resulting in a more aggressive detection strategy in our scheme. On condition that both the QoS and interference constraints are satisfied, as shown in Fig. 53(a), the maximum traffic load that can be supported by the traditional energy-detection based scheme is only 7 nats/frame. In contrast, by using our proposed queue-aware sensing scheme, the maximum traffic load that can be supported reaches 10 nats/frame (see Fig. 53(b)), which significantly outperforms the traditional scheme.

Figures. 54(a) and 54(b) depict the queue-length-bound violation probability and miss-detection probability, respectively, versus the traffic load, where the buffer size is finite with $P_{\text{th}} = 0.01$. We do not draw the curve for traditional scheme in Fig. 54(b), because the miss-detection probability of the traditional scheme is always fixed at $\overline{P}_{\text{m}}$, which is independent of the buffer size. Figs. 54(a) and 54(b) demonstrate the similar tendency to Figs. 53(a) and 53(b), respectively. Moreover, Fig. 54(a) shows that as the buffer size gets larger, the buffer-overflow probabilities for both schemes increase. However, regardless of the variation of the buffer size, our proposed queue-aware sensing scheme can always achieve much lower buffer-overflow probability than the traditional detection scheme. Based on Figs. 54(a) and 54(b), when $L = 200$ nats, the maximum traffic loads that can be supported by our proposed scheme and the traditional scheme are 4 and 3 nats/frame, respectively. When $L = 100$ nats, the maximum sustainable traffic loads under our proposed scheme and the traditional scheme are 2 and 1 nats/frame, respectively. In both cases, our queue-aware sensing scheme shows the great superiority over the traditional energy detection scheme in supporting higher system throughput under the specified

QoS and interference constraints.

F. Summary

We proposed the queue-aware spectrum sensing schemes for SUs in cognitive radio networks. The energy threshold to decide the spectrum occupancy status dynamically varies with the queue length of the secondary sender, such that statistical QoS requirements can be effectively satisfied while upper-bounding the interference probability to the primary user. We developed the queue-aware threshold control policies for the scenarios with infinite buffer and finite buffer at the secondary sender, respectively. Simulations evaluations demonstrated that under the specified statistical QoS requirements and interference constraints, our proposed schemes can support much higher data traffic loads for SUs than the traditional energy-detection based scheme.

CHAPTER X

CONCLUSIONS

A. Summary of the Dissertation

In this dissertation, we study the adaptive resource allocation problems for statistical QoS provisioning in several typical mobile wireless networks. In Chapter I, we discussed the background, motivations, and related works. In Chapter II, we gave an introduction to the theory of statistical delay-QoS guarantees and the dual concepts of effective capacity and effective bandwidth, which served as the foundation for our works in Chapters IV, V, VII, VIII, and IX.

In Chapter III, we derived the optimal time-sharing rate policy for mobile multicast with i.i.d. fading channels. In i.i.d. fading environments, to maximize average multicast goodput is equivalent to maximizing the sum of achieved rates over receivers in each fading state independently. The derived optimal policy has good scalability in term of the multicast group size. As the multicast group size approaches infinity, the derived optimal policy converges to a constant rate policy with a non-zero goodput. By using a SNR-plane partition based method, we also derived the optimal time-sharing policy for two-receiver cases with non-i.i.d. fading channels. To solve the problem that the statistical channel information is usually unavailable to the sender, we developed a sub-grouping based suboptimal rate policy, which can effective apply the algorithm derived in i.i.d. fading environments into non-i.i.d. fading environments across multicast receivers. Simulation and numerical analyses show that our proposed policies significantly outperform the existing rate adaptation schemes.

In Chapter IV, we proposed the efficient framework for mobile multicast over broadcast fading channels by integrating the effective-capacity theory, multicast rate

adaptation, and loss-rate control. Subject to the QoS exponent and average loss-rate constraint, we formulated an effective-capacity maximization problem via channel-aware rate adaptation. For rate adaptation, we employed the time-sharing and superposition-coding techniques, respectively, to handle the heterogeneous qualities over channels across multicast receivers. We also developed a novel *pre-drop* scheme to implement the more efficient QoS-driven wireless multicasting. Under the developed framework, we derived the optimal time-sharing based and superposition-coding adaptive multicast policies. Simulation evaluations demonstrated the trade-off between the effective capacity and QoS metrics and showed the superiority of our derived optimal policies over the fixed-dominating-position based policy and the constant-rate policies.

In Chapter V, we proposed a framework to model the wireless transmission of multi-layer video stream with statistical delay QoS guarantees. A separate queue is maintained for each video layer and the same statistical delay QoS-requirement needs to be satisfied by all video layers, where the statistical delay QoS is characterized by the delay-bound and its corresponding violation probability through the effective bandwidth/capacity theory. Under the proposed framework, we derived a set of optimal rate adaptation and time-slot allocation schemes for video unicast/multicast with and/or without loss tolerance, which minimizes the time-slot resource consumption. We also conducted extensive simulation experiments to demonstrate the impact of statistical QoS provisionings on wireless resource allocation by using our derived optimal adaptive transmission schemes.

In Chapter VI, we developed and analyzed an adaptive hybrid ARQ-FEC graph-code-based erasure-correcting protocol for the QoS-driven multicast services over mobile wireless networks. The key features of our proposed scheme are two-fold: the low complexity and dynamic adaptation to packet-loss levels. The low complexity is

achieved by using the graph code. To increase the error-control efficiency, we proposed a two-dimensional adaptive error-control scheme, which dynamically adjusts both the error-control redundancy and code-mapping structures in each adaptation step according to packet-loss levels. By deriving and identifying the closed-form nonlinear analytical expression between the optimal check-node degree and the packet-loss level for any given code-block length, we proposed the nonuniformed adaptive coding structures to achieve high error-control efficiency. Furthermore, we developed a loss covering strategy to determine the error-control redundancy in each transmission round and derive the corresponding analytical expressions of the error-control redundancy. Using the proposed two-dimensional nonuniformed adaptive error-control scheme, we developed an efficient hybrid ARQ-FEC protocol for multicast. We evaluated the proposed protocol through simulation experiments. The simulation results show that our proposed adaptive scheme can achieve high error-control efficiency for QoS-driven multicast services while introducing low computational complexity and implementation overhead.

In Chapter VII, we derived the optimal channel-aware time-slot length and power allocation policy in cellular networks to maximize the sum effective capacity while satisfying the proportional-effective-capacity constraint and guaranteeing the diverse statistical QoS requirements from different mobile users. We also developed a suboptimal but simpler equal-length time-division policy. Simulation results demonstrated the impact of QoS provisionings on the resource allocation across different mobile users and the network performance, and showed that our derived optimal adaptation policy significantly outperforms the equal-length time-division policy.

In Chapter VIII, we proposed the QoS-aware BS-selection schemes for the distributed wireless MIMO downlink to minimize the BS usages and to reduce the interfering range caused by the distributed MIMO system, while satisfying diverse sta-

tistical delay-QoS constraints characterized by the delay-bound violation probability and the effective capacity technique. For the single-user scenario, we developed the scheme using the incremental BS selection and time-sharing strategy and proposed the scheme employing the ordered-gain based BS-selection and probabilistic transmission strategy. The former scheme archives better performance, while the latter scheme is easier to implement. For the multi-user scenario, we developed the joint priority BS-selection, block-diagonalization precoding, and probabilistic transmission scheme. We also studied the TDMA based BS selection scheme for multi-user link. Abundant simulation results show that our proposed schemes can effectively support the incoming traffic load under the specified QoS requirements and significantly outperforms baseline schemes in terms of minimizing the average BS usage and the interfering range.

In Chapter IX, we proposed the queue-aware spectrum sensing schemes for SUs in cognitive radio networks. The energy threshold to decide the spectrum occupancy status dynamically varies with the queue length of the secondary sender, such that statistical QoS requirements can be effectively satisfied while upper-bounding the interference probability to the primary user. We developed the queue-aware threshold control policies for the scenarios with infinite buffer and finite buffer at the secondary sender, respectively. Simulations evaluations demonstrated that under the specified statistical QoS requirements and interference constraints, our proposed schemes can support much higher data traffic loads for SUs than the traditional energy-detection based scheme.

B. Future Work

1. Resource Allocation and Statistical Delay-QoS Provisionings for Multi-Hop Networks

In this dissertation, we exploited QoS-driven adaptive resource allocation schemes for several types of wireless networks. Although our works shed light on how to effectively design resource allocation schemes for statistical delay-QoS provisioning, there are still many widely cited open problems for delay-QoS guarantees in wireless networks. Our works mainly focused on single-hop wireless communications scenarios. However, how to integrate statistical delay-QoS analyses into resource allocation for multi-hop communications is still a challenging task. While effective capacity/bandwidth theory have established a powerful framework to evaluate the statistical delay QoS, they cannot be directly applied into multi-hop environments due to the following reasons: 1) each node on the route has its own buffer; 2) the total delay for the end user is a result from concatenated-queue model rather than a single-queue model. These facts cause significant challenges in analytical analyses. Consequently, it is critically important to extend the effective capacity/bandwidth theory to multi-hop networks for the design of efficient delay-QoS driven resource allocation schemes.

2. Resource Allocation and Statistical Delay-QoS Provisioning in Cooperative Wireless Networking

In addition to the problem of statistical delay-QoS provisioning in multi-hop wireless networks, the statistical delay-QoS provisioning in cooperative wireless communications and networks is also a very important and promising research direction. Not like the scenarios discussed in this dissertation, where resource allocation are controlled by the centralized base station or server, in cooperative wireless networks mobile

users will interact to cooperatively delivery data over wireless fading channels. Over the last few years, fundamental research has demonstrated the advantage of cooperative wireless communications in significantly improving the capacity and reliability for data transmissions [22, 97–99]. Moreover, the recent research focuses of cooperative wireless communications have shifted from the simpler link-level scenarios to the more complex wireless-networks regimes. However, diverse delay-QoS provisioning problems in cooperative wireless networking has neither been well addressed nor thoroughly studied. As a result, there is the urgent need to explore this research areas towards the future wireless networking techniques.

3. Quality-of-Service Provisioning versus Quality-of-Experience Provisioning

While quality-of-service (QoS) provisioning for mobile wireless communications has been widely studied, quality-of-experience (QoE) provisioning in wireless networks has attracted more and more research attention [100,101]. QoS provisioning typically has objective performance metric and measure methods, which can be quantitatively measured by the network infrastructures. In contrast, QoE represents the users' subjective satisfaction levels for their received services, which are often difficult to quantitatively evaluate. Correspondingly, developing new approaches/frameworks for designing various resource allocation schemes in wireless networks to address QoE provisioning is a very promising research direction.

REFERENCES

[1] *ITU-T Rec. G.114: One-way transmission time*, International Telecommunication Union Std.

[2] 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.

[3] C.-S. Chang, *Performance Guarantees in Communication Networks*. London: Springer-Verlag, 2000.

[4] F. Kelly, S. Zachary, and I. Ziedins, *Stochastic Networks: Theory and Applications, Royal Statistical Society Lecture Notes Series*. Oxford: Oxford University Press, U.K., 1996, vol. 4.

[5] C. Courcoubetis and R. Weber, "Effective bandwidth for stationary sources," *Probability in Engineering and Information Sciences*, vol. 9, no. 2, pp. 285–294,, 1995.

[6] M. 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. 384–395, Feb. 2001.

[7] 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.

[8] 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, Jul. 2003.

[9] A. J. Goldsmith and P. P. Varaiya, "Capacity of fading channels with channel side information," *IEEE Transactions on Information Theory*, vol. 43, no. 6, pp. 1986–1992, Nov. 1997.

[10] A. J. Goldsmith, "The capacity of downlink fading channels with variable rate and power," *IEEE Transactions on Vehicular Technology*, vol. 46, no. 3, pp. 569–580, Mar. 1997.

[11] A. J. Goldsmith and M. Effros, "The capacity region of broadcast channels with intersymbol interference and colored Gaussian noise," *IEEE Transactions on Information Theory*, vol. 47, no. 1, pp. 219–240, Jan. 2001.

[12] N. Jindal and A. Goldsmith, "Capacity and dirty paper coding for Gaussian broadcast channels with common information," in *Proc. Int. Symp. Information Theory ISIT 2004*, Chicago, IL, USA, Jun. 27-Jul. 2 2004, p. 215.

[13] D. N. Tse, "Optimal power allocation over parallel Gaussian broadcast channels," in *Proc. Symp. IEEE Int Information Theory 1997*, Ulm, Germany, Jun. 1997, p. 27.

[14] D. N. Tse, "Optimal power allocation over parallel Gaussian broadcast channels," 1997, full paper version. [Online]. Available: http://www.eecs.berkeley.edu/~dtse/broadcast2.pdf.

[15] L. Li and A. J. 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.

[16] 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.

[17] E. Biglieri, G. Caire, and G. Taricco, "Limiting performance of block-fading channels with multiple antennas," *IEEE Transactions on Information Theory*, vol. 47, no. 4, pp. 1273–1289, May 2001.

[18] L. Li and A. J. 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.

[19] J. Tang and X. Zhang, "Quality-of-service driven power and rate adaptation over wireless links," *IEEE Transactions on Wireless Communications*, vol. 6, no. 8, pp. 3058–3068, Aug. 2007.

[20] J. Tang and X. Zhang, "Quality-of-service driven power and rate adaptation for multichannel communications over wireless links," *IEEE Transactions on Wireless Communications*, vol. 6, no. 12, pp. 4349–4360, Dec. 2007.

[21] J. Tang and X. Zhang, "QoS-driven power allocation over parallel fading channels with imperfect channel estimations in wireless networks," in *Proc. INFOCOM 2007. 26th IEEE Int. Conf. Computer Communications*, Anchorage, Alaska, USA, May 2007, pp. 62–70.

[22] 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*, vol. 25, no. 4, pp. 645–656, May 2007.

[23] D. Wu and R. Negi, "Downlink scheduling in a cellular network for quality-of-service assurance," *IEEE Transactions on Vehicular Technology*, vol. 53, no. 5,

pp. 1547–1557, Sep. 2004.

[24] G. H. Forman and J. Zahorjan, "The challenges of mobile computing," *Computer*, vol. 27, no. 4, pp. 38–47, Apr. 1994.

[25] M. Hauge and O. Kure, "Multicast in 3G networks: Employment of existing ip multicast protocols in UMTS," in *WOWMOM '02: Proceedings of the 5th ACM international workshop on Wireless mobile multimedia*, New York, NY, USA, Sep. 2002, pp. 96–103.

[26] "IEEE Standard for Local and Metropolitan Area Networks Part 16: Air Interface for Fixed Broadband Wireless Access Systems," May 2004, IEEE Std 802.16-2004 (Revision of IEEE Std 802.16-2001).

[27] N. Jindal and Z.-Q. Luo, "Capacity limits of multiple antenna multicast," in *Proc. IEEE Int Information Theory Symp*, Seattle, WA, USA, Jul. 2006, pp. 1841–1845.

[28] Y. Sun and K. J. R. Liu, "Transmit diversity techniques for multicasting over wireless networks," in *Proc. WCNC Wireless Communications and Networking Conf. 2004 IEEE*, vol. 1, Atlanta, GA, USA, Mar. 2004, pp. 593–598.

[29] N. D. Sidiropoulos, T. N. Davidson, and Z.-Q. Luo, "Transmit beamforming for physical-layer multicasting," *IEEE Transactions on Signal Processing*, vol. 54, no. 6, pp. 2239–2251, Jun. 2006.

[30] Y. Park, Y. Seok, N. Choi, Y. Choi, and J.-M. Bonnin, "Rate-adaptive multimedia multicasting over IEEE 802.11 wireless LANs," in *Proc. Consumer Communications and Networking Conference, CCNC 2006*, Las Vegas, NV, USA, Jan. 2006, pp. 178–182.

[31] P. K. Gopala and H. El Gamal, "On the throughput-delay tradeoff in cellular multicast," in *Proc. Int. Wireless Networks, Communications and Mobile Computing Conf.*, vol. 2, Maui, HI, USA, Jun. 2005, pp. 1401–1406.

[32] X. Zhang, K. G. Shin, D. Saha, and D. D. Kandlur, "Scalable flow control for multicast ABR services in ATM networks," *IEEE/ACM Transactions on Networking*, vol. 10, no. 1, pp. 67–85, Feb. 2002.

[33] X. Zhang and K. G. Shin, "Delay analysis of feedback-synchronization signaling for multicast flow control," *Networking, IEEE/ACM Transactions on*, vol. 11, no. 3, pp. 436–450, Jun.s 2003.

[34] X. Zhang and K. G. Shin, "Markov-chain modeling for multicast signaling delay analysis," *Networking, IEEE/ACM Transactions on*, vol. 12, no. 4, pp. 667–680, Aug. 2004.

[35] S. McCanne, M. Vetterli, and V. Jacobson, "Low-complexity video coding for receiver-driven layered multicast," *IEEE Journal on Selected Areas in Communications*, vol. 15, no. 6, pp. 983–1001, Aug. 1997.

[36] J. Villalon, P. Cuenca, L. Orozco-Barbosa, Y. Seok, and T. Turletti, "Cross-layer architecture for adaptive video multicast streaming over multirate wireless LANs," *IEEE Journal on Selected Areas in Communications*, vol. 25, no. 4, pp. 699–711, May 2007.

[37] A. Majumda, D. G. Sachs, I. V. Kozintsev, K. Ramchandran, and M. M. Yeung, "Multicast and unicast real-time video streaming over wireless lans," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 12, no. 6, pp. 524–534, Jun. 2002.

[38] Y. S. Chan, J. W. Modestino, Q. Qu, and X. Fan, "An end-to-end embedded approach for multicast/broadcast of scalable video over multiuser cdma wireless networks," *IEEE Transactions on Multimedia*, vol. 9, no. 3, pp. 655–667, Apr. 2007.

[39] R. Knopp and P. A. Humblet, "Information capacity and power control in single-cell multiuser communications," in *Proc. IEEE International Conference on Communications (ICC)*, vol. 1, Seattle, WA, USA, Jun. 1995, pp. 331–335.

[40] A. Sanderovich, S. Shamai, and Y. Steinberg, "Distributed MIMO receiver—Achievable rates and upper bounds," *IEEE Transactions on Information Theory*, vol. 55, no. 10, pp. 4419–4438, Oct. 2009.

[41] D. R. Brown and H. V. Poor, "Time-slotted round-trip carrier synchronization for distributed beamforming," *IEEE Transactions on Signal Processing*, vol. 56, no. 11, pp. 5630–5643, Nov. 2008.

[42] P. Shang, G. Zhu, L. Tan, G. Su, and T. Li, "Transmit antenna selection for the distributed MIMO systems," in *Proc. Int. Conf. Networks Security, Wireless Communications and Trusted Computing NSWCTC '09*, vol. 2, Wuhan, Hubei, China, Apr. 2009, pp. 449–453.

[43] R. Chen, R. W. Heath, and J. G. Andrews, "Transmit selection diversity for unitary precoded multiuser spatial multiplexing systems with linear receivers," *IEEE Transactions on Signal Processing*, vol. 55, no. 3, pp. 1159–1171, Mar. 2007.

[44] R. Mudumbai, D. R. Brown, U. Madhow, and H. V. Poor, "Distributed transmit beamforming: Challenges and recent progress," *Communications Magazine, IEEE*, vol. 47, no. 2, pp. 102–110, Feb. 2009.

[45] E. Telatar, "Capacity of multi-antenna Gaussian channels," *European Trans. Telecomm.*, vol. 10, no. 6, pp. 585–596, Nov. 1999.

[46] M. Gharavi-Alkhansari and A. B. Gershman, "Fast antenna subset selection in MIMO systems," *IEEE Transactions on Signal Processing*, vol. 52, no. 2, pp. 339–347, 2004.

[47] S. Sanayei and A. Nosratinia, "Antenna selection in MIMO systems," *IEEE Communications Magazine*, vol. 42, no. 10, pp. 68–73, Oct. 2004.

[48] S. Haykin, "Cognitive radio: Brain-empowered wireless communications," *Selected Areas in Communications, IEEE Journal on*, vol. 23, no. 2, pp. 201–220, Feb. 2005.

[49] Q. Zhao, L. Tong, and A. Swami, "Decentralized cognitive mac for dynamic spectrum access," in *Proc. First IEEE Int. Symp. New Frontiers in Dynamic Spectrum Access Networks DySPAN 2005*, Baltimore, MD, USA, Nov. 2005, pp. 224–232.

[50] A. Ghasemi and E. S. Sousa, "Collaborative spectrum sensing for opportunistic access in fading environments," in *Proc. First IEEE Int. Symp. New Frontiers in Dynamic Spectrum Access Networks DySPAN 2005*, Baltimore, MD, USA, Nov. 2005, pp. 131–136.

[51] D. Cabric, A. Tkachenko, and R. W. Brodersen, "Spectrum sensing measurements of pilot, energy, and collaborative detection," in *Proc. IEEE Military Communications Conf. MILCOM 2006*, Washington, DC, USA, Oct. 2006, pp. 1–7.

[52] F. F. Digham, M.-S. Alouini, and M. K. Simon, "On the energy detection of unknown signals over fading channels," *IEEE Transactions on Communications*, vol. 55, no. 1, pp. 21–24, Jan. 2007.

[53] 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, 2006.

[54] Q. Du and X. Zhang, "Time-sharing based rate adaptation for multicast over wireless fading channels in mobile wireless networks," in *Proc. 40th Annual Conf. Information Sciences and Systems, CISS 2006*, Princeton, NJ, USA, Mar. 2006, pp. 1385–1390.

[55] Q. Du and X. Zhang, "Fixed/variable power multicast over heterogeneous fading channels in cellular networks," in *Proc. IEEE Int. Conf. Communications ICC '08*, Beijing, China, May 2008, pp. 2182–2186.

[56] Q. Du and X. Zhang, "Effective capacity optimization with layered transmission for multicast in wireless networks," in *Proc. Int. Wireless Communications and Mobile Computing Conf. IWCMC '08*, Crete Island, Greece, Aug. 2008, pp. 267–272.

[57] Q. Du and X. Zhang, "Effective capacity of superposition coding based mobile multicast in wireless networks," in *Proc. IEEE Int. Conf. Communications ICC '09*, Dresden, Germany, Jun. 2009.

[58] X. Zhang and Q. Du, "Cross-layer modeling for QoS-driven multimedia multicast/broadcast over fading channels in [Advances in Mobile Multimedia]," *IEEE Communications Magazine*, vol. 45, no. 8, pp. 62–70, Aug. 2007.

[59] Q. Du and X. Zhang, "Statistical QoS provisionings for wireless unicast/multicast of multi-layer video streams," *IEEE Journal on Selected Areas in Communications*, vol. 28, no. 3, pp. 420–433, Mar. 2010.

[60] X. Zhang and Q. Du, "Adaptive low-complexity erasure-correcting code-based protocols for QoS-driven mobile multicast services over wireless networks," *IEEE Transactions on Vehicular Technology*, vol. 55, no. 5, pp. 1633–1647, Sep. 2006.

[61] Q. Du and X. Zhang, "Resource allocation for downlink statistical multiuser qos provisionings in cellular wireless networks," in *Proc. INFOCOM 2008. The 27th IEEE Conf. Computer Communications*, Phoenix, AZ, USA, Apr. 2008, pp. 2405–2413.

[62] Q. Du and X. Zhang, "QoS-aware base-station selections for distributed MIMO links in broadband wireless networks," *submitted to IEEE Journal on Selected Areas in Communications*, 2010.

[63] Q. Du and X. Zhang, "Queue-aware spectrum sensing for interference-constrained transmissions in cognitive radio networks," in *Proc. IEEE International Conference on Communications (ICC)*, Cape Town, South Africa, May 2010.

[64] G. L. Choudhury, 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.

[65] U. Varshney, "Multicast over wireless networks," *Commun. ACM*, vol. 45, no. 12, pp. 31–37, Dec. 2002.

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

[67] J. Nonnenmacher, E. W. Biersack, and D. Towsley, "Parity-based loss recovery for reliable multicast transmission," *IEEE/ACM Transactions on Networking*, vol. 6, no. 4, pp. 349–361, Aug. 1998.

[68] J. W. Byers, M. Luby, and M. Mitzenmacher, "A digital fountain approach to asynchronous reliable multicast," *IEEE Journal on Selected Areas in Communications*, vol. 20, no. 8, pp. 1528–1540, Oct. 2002.

[69] 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, Mar. 2004.

[70] M. K. Simon and M.-S. Alouini, *Digital Communication over Fading Channels*. Hoboken, NJ: John Wiley & Sons, Inc., 2005.

[71] T. Cover, "Broadcast channels," *IEEE Transactions on Information Theory*, vol. 18, no. 1, pp. 2–14, Jan. 1972.

[72] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integral, Series, and Products*. New York, NY: Academic Press, 1992.

[73] O. Kallenberg, *Foundations Modern Probability*, 2nd, Ed. New York: Springer-Verlag, 2000.

[74] F. Chapeau-Blondeau and A. Monir, "Numerical evaluation of the Lambert W function and application to generation of generalized Gaussian noise with

exponent 1/2," *IEEE Transactions on Signal Processing*, vol. 50, no. 9, pp. 2160–2165, Sep. 2002.

[75] M. J. Gans and B. Chen, "Beaconing in MIMO broadcast channels," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP '05)*, Philadelphia, PA, USA, Mar. 2005.

[76] Q. Du and X. Zhang, "Cross-layer design based rate control for mobile multicast in cellular networks," in *Proc. IEEE Global Telecommunications Conf. GLOBECOM '07*, Washington DC, USA, Nov. 2007, pp. 5180–5184.

[77] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge: Cambridge University Press, U.K., 2004.

[78] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to Algorithms*, 2nd, Ed. Cambridge, MA: MIT Press, 2001.

[79] M. S. Bazaraa, H. D. Sherali, and C. M. Shetty, *Nonlinear Programming: Theory and Algorithms*, 3rd ed. Hoboken, NJ: John Wiley & Sons, Inc., 2006.

[80] J. Shin, J. W. Kim, and C.-C. J. Kuo, "Quality-of-service mapping mechanism for packet video in differentiated services network," *IEEE Transactions on Multimedia*, vol. 3, no. 2, pp. 219–231, Jun. 2001.

[81] H. Schwarz, D. Marpe, and T. Wiegand, "Overview of the scalable video coding extension of the H.264/AVC standard," *IEEE Transactions on Circuits and Systems*, vol. 17, no. 9, pp. 1103–1120, Sep. 2007.

[82] A. Shokrollahi, "Raptor codes," *IEEE Transactions on Information Theory*, vol. 52, no. 6, pp. 2551–2567, Jun. 2006.

[83] *Multimedia Broadcast/Multicast Service: Protocols and Codecs*, 3GPP Std., Sep. 2009, tS 26.346, Release 9, v9.0.0.

[84] S. Floyd, V. Jacobson, C.-G. Liu, S. McCanne, and L. Zhang, "A reliable multicast framework for light-weight sessions and application level framing," *IEEE/ACM Transactions on Networking*, vol. 5, no. 6, pp. 784–803, Dec. 1997.

[85] C. Huitema, "The case for packet level FEC," in *PfHSN '96: Proceedings of the TC6 WG6.1/6.4 Fifth International Workshop on Protocols for High-Speed Networks*, London, UK, Oct. 1996, pp. 109–120.

[86] N. Nikaein, H. Labiod, and C. Bonnet, "MA-FEC: a QoS-based adaptive fec for multicast communication in wireless networks," in *Proc. IEEE Int. Conf. Communications ICC 2000*, vol. 2, New Orleans, LA, USA., Jun. 2000, pp. 954–958.

[87] M. G. Luby, M. Mitzenmacher, M. A. Shokrollahi, and D. A. Spielman, "Efficient erasure correcting codes," *IEEE Transactions on Information Theory*, vol. 47, no. 2, pp. 569–584, Feb. 2001.

[88] M. Luby, "LT codes," in *Proc. 43rd Annual IEEE Symp. Foundations of Computer Science*, Nov. 2002, pp. 271–280.

[89] A. Leon-Garcia and I. Widjaja, *Communication Networks: Fundamentals Concepts and Key Architectures*. Boston, MA: McGraw-Hill, 2000.

[90] S. Boyd, L. Xiao, and A. Mutapcic, "Subgradient methods," Oct. 2003. [Online]. Available: http://www.stanford.edu/class/ee392o/subgrad_method.pdf

[91] H. Weingarten, Y. Steinberg, and S. Shamai, "The capacity region of the gaussian multiple-input multiple-output broadcast channel," *IEEE Transactions on*

Information Theory, vol. 52, no. 9, pp. 3936–3964, Sep. 2006.

[92] Q. H. Spencer, A. L. Swindlehurst, and M. Haardt, "Zero-forcing methods for downlink spatial multiplexing in multiuser mimo channels," *IEEE Transactions on Signal Processing*, vol. 52, no. 2, pp. 461–471, Feb. 2004.

[93] C. Meyer, *Matrix Analysis and Applied Linear Algebra*. Philadelphia, PA: Society for Industrial and Applied Mathematics, 2000.

[94] T. S. Rappaport, *Wireless Communications: Principles & Practice*, 2nd, Ed. Upper Saddle River, NJ: Prentice Hall, 2001.

[95] J. Wang and J. Chen, "Performance of wideband CDMA systems with complex spreading and imperfect channel estimation," *IEEE Journal on Selected Areas in Communications*, vol. 19, no. 1, pp. 152–163, Jan. 2001.

[96] 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.

[97] 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, Nov. 2003.

[98] 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, Nov. 2003.

[99] 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, Dec. 2004.

[100] M. Fiedler, T. Hossfeld, and P. Tran-Gia, "A generic quantitative relationship between quality of experience and quality of service," *IEEE Network*, vol. 24, no. 2, pp. 36–41, Mar. 2010.

[101] P. Brooks and B. Hestnes, "User measures of quality of experience: why being objective and quantitative is important," *IEEE Network*, vol. 24, no. 2, pp. 8–13, Mar. 2010.

[102] J. R. Norris, *Markov Chains*. Cambridge: Cambridge University Press, U.K., 1997.

APPENDIX A

PROOF OF PROPOSITION 1

Proof. Define $\overline{R}_{\text{sum}} \triangleq \sum_{n=1}^{N} \overline{R}_n\big(\boldsymbol{\lambda}(\boldsymbol{\gamma})\big)$. Based on Eq. (3.4) we get an upper-bound of $\overline{R}\big(\boldsymbol{\lambda}(\boldsymbol{\gamma})\big)$ given by

$$\overline{R}\big(\boldsymbol{\lambda}(\boldsymbol{\gamma})\big) \leq \frac{1}{N}\overline{R}_{\text{sum}}\big(\boldsymbol{\lambda}(\boldsymbol{\gamma})\big) \leq \mathcal{J} \triangleq \frac{1}{N}\max_{\boldsymbol{\lambda}(\boldsymbol{\gamma})}\Big\{\overline{R}_{\text{sum}}\big(\boldsymbol{\lambda}(\boldsymbol{\gamma})\big)\Big\}. \tag{A.1}$$

Applying Eq. (3.3), we can simplify the right-hand side of Eq. (A.1) as

$$\max_{\boldsymbol{\lambda}(\boldsymbol{\gamma})}\Big\{\overline{R}_{\text{sum}}\big(\boldsymbol{\lambda}(\boldsymbol{\gamma})\big)\Big\} \Leftrightarrow \max_{\boldsymbol{\lambda}(\boldsymbol{\gamma})}\left\{\sum_{n=1}^{N}\mathbf{c}_n(\boldsymbol{\gamma})^{\tau}\boldsymbol{\lambda}(\boldsymbol{\gamma})\right\}, \tag{A.2}$$

which implies that for each $\boldsymbol{\gamma}$, maximizing $\overline{R}_{\text{sum}}\big(\boldsymbol{\lambda}(\boldsymbol{\gamma})\big)$ is equivalent to maximizing $\sum_{n=1}^{N}\mathbf{c}_n(\boldsymbol{\gamma})^{\tau}\boldsymbol{\lambda}(\boldsymbol{\gamma})$. Furthermore, we derive

$$\sum_{n=1}^{N}\mathbf{c}_n(\boldsymbol{\gamma})^{\tau}\boldsymbol{\lambda}(\boldsymbol{\gamma}) = \sum_{i=1}^{N}\widehat{\lambda}_i(\boldsymbol{\gamma})\left[\sum_{k=1}^{N}c\big(\widehat{\gamma}_k, B\log(1+\widehat{\gamma}_i)\big)\right] = \sum_{i=1}^{N}\widehat{\lambda}_i(\boldsymbol{\gamma})\widehat{\Theta}_i, \tag{A.3}$$

where $\widehat{\Theta}_i = B\sum_{\forall\, k,\widehat{\gamma}_k\geq\widehat{\gamma}_i}\log(1+\widehat{\gamma}_i)$. Because $\widehat{\Theta}_i$ does not vary with $\boldsymbol{\lambda}(\boldsymbol{\gamma})$, in order to optimize $\sum_{n=1}^{N}\mathbf{c}_n(\boldsymbol{\gamma})^{\tau}\boldsymbol{\lambda}(\boldsymbol{\gamma})$, we need to allocate all time proportions to the maximum $\widehat{\Theta}_i$. Thus, the policy maximizing the right-hand side of Eq. (A.3), denoted by $\widehat{\boldsymbol{\lambda}}^{*}(\boldsymbol{\gamma})$ (the permuted version), is given by

$$\widehat{\lambda}^{*}{}_i(\boldsymbol{\gamma}) = 1, \quad \text{if } i = i^*; \quad \widehat{\lambda}^{*}{}_i(\boldsymbol{\gamma}) = 0, \quad \text{if } i \neq i^*; \quad i^* = \arg\max_{1\leq i\leq N}\Big\{\widehat{\Theta}_i\Big\}, \tag{A.4}$$

Since $\widehat{\gamma}_1 \geq \widehat{\gamma}_2 \geq \cdots \geq \widehat{\gamma}_N$, we get

$$\begin{aligned} i^* &= \arg\max_{1\leq i\leq N}\big\{\widehat{\Theta}_i\big\} \\ &= \arg\max_{1\leq i\leq N}\left\{B\sum_{\forall\, k,\widehat{\gamma}_k\geq\widehat{\gamma}_i}\log(1+\widehat{\gamma}_i)\right\} = \arg\max_{1\leq i\leq N}\big\{iB\log(1+\widehat{\gamma}_i)\big\}. \end{aligned} \tag{A.5}$$

Note that i^* may not be unique for a given CSI vector $\boldsymbol{\gamma}$. If there exist multiple indices $i_1, i_2, \ldots, i_S$ (S is an integer, $S \leq N$) such that $i_s B \log(1+\widehat{\gamma}_{i_s}) = \max_{1\leq i\leq N}\{iB\log(1+\widehat{\gamma}_i)\}$ holds for all $s = 1, 2, \ldots, S$, we set $i^* = \max\{i_1, i_2, \ldots, i_S\}$ without loss generality. Because $\{\gamma_n\}_{n=1}^N$ are i.i.d. and $\widehat{\boldsymbol{\gamma}}$ completely determines $\widehat{\boldsymbol{\lambda}}^*(\boldsymbol{\gamma})$, $\widehat{\boldsymbol{\lambda}}^*(\boldsymbol{\gamma})$ benefits all receivers evenly and we obtain $\overline{R}_1\big(\widehat{\boldsymbol{\lambda}}^*(\boldsymbol{\gamma})\big) = \overline{R}_2\big(\widehat{\boldsymbol{\lambda}}^*(\boldsymbol{\gamma})\big) = \cdots = \overline{R}_N\big(\widehat{\boldsymbol{\lambda}}^*(\boldsymbol{\gamma})\big)$. Then through Eq. (A.1) we obtain $\overline{R}_{\text{sum}} = N\mathcal{J}$, and further derive

$$\overline{R}^{\text{gp}}\big|_{\widehat{\boldsymbol{\lambda}}^*(\boldsymbol{\gamma})} = \overline{R}_n\big(\widehat{\boldsymbol{\lambda}}^*(\boldsymbol{\gamma})\big) = \frac{1}{N}\overline{R}_{\text{sum}} = \mathcal{J}. \quad \forall 1 \leq n \leq N, \tag{A.6}$$

Therefore, $\widehat{\boldsymbol{\lambda}}^*(\boldsymbol{\gamma})$ achieves the upper-bound of the average multicast goodput, and Proposition 1 follows. □

APPENDIX B

PROOF OF PROPOSITION 2

Proof. For convenience of presentation, given multicast group size N we rewrite the average multicast goodput as $\overline{R}_{\rm gp}\left(N, \boldsymbol{\lambda}\left(\boldsymbol{\gamma}_{\langle N\rangle}\right)\right)$. Moreover, we rewrite the corresponding OPTS policy as $\boldsymbol{\lambda}^*\left(\boldsymbol{\gamma}_{\langle N\rangle}\right)$. For given N, we split the entire multicast group into two sub-groups: sub-group 1 includes users $1, 2, \ldots, N-1$ and sub-group 2 includes user N. Accordingly, we get $\boldsymbol{\gamma}_{\langle N-1\rangle} = (\gamma_1, \gamma_2, \ldots, \gamma_{N-1})^\tau$ and $\boldsymbol{\gamma}_{\langle N\rangle} = \left(\boldsymbol{\gamma}^\tau_{\langle N-1\rangle}, \gamma_N\right)^\tau$. Following Definition 2, we derive

$$\begin{aligned}\overline{R}_{\rm gp}\left(N, \boldsymbol{\lambda}^*\left(\boldsymbol{\gamma}_{\langle N\rangle}\right)\right) &= \min_{1\le n\le N}\left\{\overline{R}_n\left(\boldsymbol{\lambda}^*\left(\boldsymbol{\gamma}_{\langle N\rangle}\right)\right)\right\} \\ &= \min\left\{\min_{1\le n\le N-1}\left\{\overline{R}_n\left(\boldsymbol{\lambda}^*\left(\boldsymbol{\gamma}_{\langle N\rangle}\right)\right)\right\}, \overline{R}_N\left(\boldsymbol{\lambda}^*\left(\boldsymbol{\gamma}_{\langle N\rangle}\right)\right)\right\}. \quad \text{(B.1)}\end{aligned}$$

We first focus only on sub-group 1. Based on results in Section D-1, the OPTS policy for sub-group 1 is $\boldsymbol{\lambda}^*\left(\boldsymbol{\gamma}_{\langle N-1\rangle}\right)$ obtained through Proposition 1. Therefore, we have

$$\begin{aligned}\min_{1\le n\le N-1}\left\{\overline{R}_n\left(\boldsymbol{\lambda}^*\left(\boldsymbol{\gamma}_{\langle N\rangle}\right)\right)\right\} &\le \min_{1\le n\le N-1}\left\{\overline{R}_n\left(\boldsymbol{\lambda}^*\left(\boldsymbol{\gamma}_{\langle N-1\rangle}\right)\right)\right\} \\ &= \overline{R}_{\rm gp}\left(N-1, \boldsymbol{\lambda}^*\left(\boldsymbol{\gamma}_{\langle N-1\rangle}\right)\right) \quad \text{(B.2)}\end{aligned}$$

Plugging Eq. (B.2) into Eq. (B.1), we get

$$\begin{aligned}\overline{R}_{\rm gp}\left(N, \boldsymbol{\lambda}^*\left(\boldsymbol{\gamma}_{\langle N\rangle}\right)\right) &\le \min\left\{\overline{R}_{\rm gp}\left(N-1, \boldsymbol{\lambda}^*\left(\boldsymbol{\gamma}_{\langle N-1\rangle}\right)\right), \overline{R}_N\left(\boldsymbol{\lambda}^*\left(\boldsymbol{\gamma}_{\langle N\rangle}\right)\right)\right\} \\ &\le \overline{R}_{\rm gp}\left(N-1, \boldsymbol{\lambda}^*\left(\boldsymbol{\gamma}_{\langle N-1\rangle}\right)\right), \quad \text{(B.3)}\end{aligned}$$

and thus Proposition 2 follows. □

APPENDIX C

PROOF OF PROPOSITION 4

Proof. For convenience, we use V_a^*, V_b^*, V_c^*, and V_d^* to denote V_a, V_b, V_c, and V_d, with $\eta_0 = \eta_0^*$, respectively. Consider any policy $\boldsymbol{\lambda}'(\boldsymbol{\gamma}) \neq \boldsymbol{\lambda}^*(\boldsymbol{\gamma})$ with $\overline{R}_1\big(\boldsymbol{\lambda}'(\boldsymbol{\gamma})\big) = \overline{R}_2\big(\boldsymbol{\lambda}'(\boldsymbol{\gamma})\big)$. We have

$$\begin{aligned}
&\overline{R}_2\big(\boldsymbol{\lambda}^*(\boldsymbol{\gamma})\big) - \overline{R}_2\big(\boldsymbol{\lambda}'(\boldsymbol{\gamma})\big) \\
&\quad = \mathbb{E}\left\{-\lambda_1'(\boldsymbol{\gamma})\mu_2(\boldsymbol{\gamma}) | \boldsymbol{\gamma} \in V_a^* \cup V_c^*\right\} + \mathbb{E}\left\{(1-\lambda_1'(\boldsymbol{\gamma}))\mu_2(\boldsymbol{\gamma}) | \boldsymbol{\gamma} \in V_b^* \cup V_d^*\right\}. \quad \text{(C.1)}
\end{aligned}$$

Using the definition of compensation efficiency (see Eq. (3.24)), we derive

$$\begin{aligned}
&\overline{R}_1\big(\boldsymbol{\lambda}^*(\boldsymbol{\gamma})\big) - \overline{R}_1\big(\boldsymbol{\lambda}'(\boldsymbol{\gamma})\big) \\
&= \mathbb{E}\left\{-\lambda_1'(\boldsymbol{\gamma})\eta(\boldsymbol{\gamma})(-\mu_2(\boldsymbol{\gamma})) | \boldsymbol{\gamma} \in V_a^* \cup V_c^*\right\} \\
&\qquad\qquad + \mathbb{E}\left\{(1-\lambda_1'(\boldsymbol{\gamma}))\eta(\boldsymbol{\gamma})(-\mu_2(\boldsymbol{\gamma})) | \boldsymbol{\gamma} \in V_b^* \cup V_d^*\right\} \\
&\overset{(a)}{\geq} -\eta_0^* \Big[\mathbb{E}\left\{-\lambda_1'(\boldsymbol{\gamma})\mu_2(\boldsymbol{\gamma}) | \boldsymbol{\gamma} \in V_a^* \cup V_c^*\right\} + \mathbb{E}\left\{(1-\lambda_1'(\boldsymbol{\gamma}))\mu_2(\boldsymbol{\gamma}) | \boldsymbol{\gamma} \in V_b^* \cup V_d^*\right\}\Big] \\
&= -\eta_0^* \left[\overline{R}_2\big(\boldsymbol{\lambda}^*(\boldsymbol{\gamma})\big) - \overline{R}_2\big(\boldsymbol{\lambda}'(\boldsymbol{\gamma})\big)\right], \quad \text{(C.2)}
\end{aligned}$$

where (a) follows because $\mu_2(\boldsymbol{\gamma}) \leq 0$ holds for all $\boldsymbol{\gamma}$ (see Eq. (3.22)), $\eta_0^* \geq \eta(\boldsymbol{\gamma})$ for $\boldsymbol{\gamma} \in V_a^* \cup V_c^*$, and $\eta_0^* \leq \eta(\boldsymbol{\gamma})$ for $\boldsymbol{\gamma} \in V_b^* \cup V_d^*$. Comparing Eq. (C.2) with the facts of $\overline{R}_1\big(\boldsymbol{\lambda}'(\boldsymbol{\gamma})\big) = \overline{R}_2\big(\boldsymbol{\lambda}'(\boldsymbol{\gamma})\big)$ and $\overline{R}_1\big(\boldsymbol{\lambda}^*(\boldsymbol{\gamma})\big) = \overline{R}_2\big(\boldsymbol{\lambda}^*(\boldsymbol{\gamma})\big)$, we derive $\overline{R}^{\text{gp}}|_{\boldsymbol{\lambda}^*(\boldsymbol{\gamma})} \geq \overline{R}^{\text{gp}}|_{\boldsymbol{\lambda}'(\boldsymbol{\gamma})}$, which completes the proof of Proposition 4. □

APPENDIX D

PROOF OF LEMMA 1

Proof. Applying Definition 9 to Eq. (4.16), we obtain Eq. (4.22) of Claim 1. Following the principles for optimizing non-differentiable concave functions [79, pp. 128], $\widetilde{R}_s$ is the solution to

$$0 \in \partial \widetilde{g}_{\text{sum}}(\widetilde{R}_s). \tag{D.1}$$

In order to satisfy Eq. (D.1), Eq. (4.23) needs to hold for certain integer $1 \leq k \leq \mathcal{N}$. The existences of such a k is guaranteed by Eq. (4.17). The uniqueness of such a k is obtained through applying the strict monotonicity of $\eta_{\alpha(j)}$ given by Eq. (4.17). If $\eta_{\alpha(k)} < 0$, $\widetilde{g}_{\text{sum}}(R_s)$ is a strictly increasing function for $R_s \in [r_{\pi(N)}, r_{\alpha(k)}]$, and a strictly decreasing function for $R_s \in [r_{\alpha(k)}, r_{\pi(1)}]$. As a result, $\widetilde{R}_s = r_{\alpha(k)}$ achieves the unique peak of $\widetilde{g}_{\text{sum}}(R_s)$. If $\eta_{\alpha(k)} = 0$, we have $\eta_{\alpha(k+1)} > 0 = \eta_{\alpha(k)} > \eta_{\alpha(k-1)}$, suggesting that $\widetilde{g}_{\text{sum}}(R_s)$ is a strictly increasing function for $R_s \in [r_{\alpha(\mathcal{N})}, r_{\alpha(k)}]$, remains unchanged within $R_s \in [r_{\alpha(k)}, r_{\alpha(k-1)}]$, and becomes a strictly decreasing function within $R_s \in [r_{\alpha(k-1)}, r_{\alpha(1)}]$. Consequently, the maximizer $\widetilde{R}_s$ can be any real number within $[r_{\alpha(k)}, r_{\alpha(k-1)}]$, which competes the proof of Claim 2 and thus Lemma 1 follows. □

APPENDIX E

PROOF OF LEMMA 2

Proof. Consider any feasible TS policy $(\boldsymbol{\lambda}', R_d')$ of problem **IV**-$\boldsymbol{P}$**1**-a with group loss rate equal to q_0', where $q_0' \leq q_{\text{th}}$ holds. We denote A' the point in the ISR-ISG plane generated under $\boldsymbol{\lambda}'$ and suppose that A' is not an upper-boundary point of the convex hull spanned by $\{A_{\pi(i)}\}_{i=1}^{N}$ for some $\boldsymbol{\gamma}$.

Note that any feasible $(R_s, \widetilde{g}_{\text{sum}}(R_s))$ of **IV**-$\boldsymbol{P}$**1**-b must be an upper-boundary point. Thus, in order to prove Lemma 2 we need to show that for any $(\boldsymbol{\lambda}', R_d')$, there exists a certain policy $(\boldsymbol{\lambda}'', R_d'')$ which satisfies the following three conditions:

Condition (i): $\boldsymbol{\lambda}''$ generates an upper-boundary point for all $\boldsymbol{\gamma}$.

Condition (ii): $(\boldsymbol{\lambda}'', R_d'')$ is feasible to **IV**-$\boldsymbol{P}$**1**-a;

Condition (iii): $e^{-\theta(R_s''+R_d'')} \leq e^{-\theta(R_s'+R_d')}$, $\forall\, \boldsymbol{\gamma}$.

In the rest of this proof, we use $(\cdot)'$ and $(\cdot)''$ to denote the values of corresponding variables under the policies $(\boldsymbol{\lambda}', R_d')$ and $(\boldsymbol{\lambda}'', R_d'')$, respectively. Consider any $\boldsymbol{\gamma}$ under which A' is not an upper-boundary point, as shown in Fig. 55. Starting from A', which is marked with a hollow square, we draw a ray with slope $N(1-q_0')$ towards the direction where the sending rate increases. Then, we have either of the following two cases, depending on the position of A':

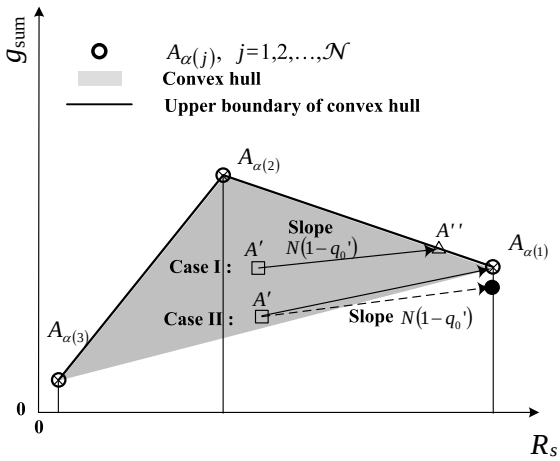


Fig. 55. Illustration in the ISR-ISG plane for the proof of Lemma 2, where $\mathcal{N} = 3$ in this example.

Case I: The ray intersects the upper boundary at A'', which is marked with a hollow triangle in Fig. 55;

Case II: There is no intersection between the ray (marked with a dashed line in Fig. 55) and the upper boundary. The ray intersects with the vertical line $R_s = R_{\alpha(1)}$ at a point below $A_{\alpha(1)}$, which is marked with a black solid circle. In this case, we let $A'' := A_{\alpha(1)}$.

Next, we construct a new adaptive transmission policy $(\boldsymbol{\lambda}'', R_d'')$ through the strategy below. If A' is on the upper boundary, we set $\boldsymbol{\lambda}'' := \boldsymbol{\lambda}'$. Otherwise, we set $\boldsymbol{\lambda}''$ as the TS policy generating A'' as described in Cases I and II. Moreover, we let $R_d'' = R_d'$ for all $\boldsymbol{\gamma}$.

Under the above strategy, Condition (i) is satisfied. Defining $\Delta R_s \triangleq R_s'' - R_s'$ and $\Delta g_{\text{sum}} \triangleq g_{\text{sum}}'' - g_{\text{sum}}'$, we have

$$\Delta R_s \geq 0, \quad \Delta g_{\text{sum}} \geq N(1-q_0')\Delta R_s, \quad \text{and} \quad R_d' = R_d'', \quad \forall\, \boldsymbol{\gamma}. \tag{E.1}$$

Then, we derive

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\gamma}}\Big\{\rho(R_s''+R_d'')-g_{\text{sum}}''\Big\} &= \mathbb{E}_{\boldsymbol{\gamma}}\Big\{\rho(R_s'+R_d')-g_{\text{sum}}'\Big\}+\mathbb{E}_{\boldsymbol{\gamma}}\Big\{\rho\Delta R_s-\Delta g_{\text{sum}}\Big\}\\ &\overset{(a)}{\leq} \mathbb{E}_{\boldsymbol{\gamma}}\Big\{\rho\Delta R_s-\Delta g_{\text{sum}}\Big\}\\ &\overset{(b)}{\leq} \mathbb{E}_{\boldsymbol{\gamma}}\Big\{N(1-q_{\text{th}})\Delta R_s-N(1-q_0')\Delta R_s\Big\} \overset{(c)}{\leq} 0, \quad \text{(E.2)}\end{aligned}$$

where (a) holds because of $q_0' \leq q_{\text{th}}$, (b) results from $\Delta g_{\text{sum}} \geq N(1-q_0')\Delta R_s$, (c) follows by applying $q_0' \leq q_{\text{th}}$ and $\Delta R_s \geq 0$. Eq. (E.2) shows that the new policy $(\boldsymbol{\lambda}'', R_d'')$ is feasible to problem **IV-*P*1**-a, and thus Condition (ii) follows. Furthermore, because of $\Delta R_s \geq 0$, Condition (iii) holds, which completes the proof of Lemma 2. □

APPENDIX F

PROOF OF THEOREM 1

Proof. If $\psi^* = 0$, the solutions to Eqs. (4.27)-(4.28) become $(R_s^* + R_d^*) \to \infty$, implying that the group loss-rate constraint is violated. Therefore, we must have $\psi^* > 0$. Next, we consider the following two cases to solve Eqs. (4.27)-(4.28).

A. For the case with $R_d^* > 0$

If $R_d^* > 0$, plugging Eq. (4.31) into Eq. (4.28) we get

$$\psi^* \rho = \theta e^{-\theta(R_s^* + R_d^*)}, \quad \text{if } R_d^* > 0. \tag{F.1}$$

Applying Eq. (F.1) into Eq. (4.30), we derive

$$\partial \ell_{R_s}(R_s^*, R_d^*; \psi^*) = \begin{cases} \Big\{ -\psi^* \eta_{\alpha(j)} \Big\}, & \text{if } r_{\alpha(j)} < R_s^* < r_{\alpha(j-1)} \ \& \ \mathcal{N} \geq j \geq 2; \\ \Big[-\psi^* \eta_{\alpha(j+1)}, -\psi^* \eta_{\alpha(j)} \Big], & \text{if } R_s^* = r_{\alpha(j)} \ \& \ \mathcal{N} \geq j \geq 1, \end{cases} \tag{F.2}$$

Because $\psi^* > 0$, we can see from Eqs. (4.30) and (4.22) that $0 \in \partial \ell_{R_s}(R_s^*, R_d^*; \psi^*)$ in this case is equivalent to $0 \in \partial \widetilde{g}_{\text{sum}}(R_s^*)$. Comparing $0 \in \partial \widetilde{g}_{\text{sum}}(R_s^*)$ with Eq. (D.1), we see that

$$R_s^* = \widetilde{R}_s, \qquad \text{if } R_d^* > 0. \tag{F.3}$$

Solving Eq. (F.1), we can express R_d^* through R_s^* by

$$R_d^* = -\frac{1}{\theta} \log \left(\frac{\psi^* \rho}{\theta} \right) - R_s^*, \quad \text{if } R_d^* > 0. \tag{F.4}$$

B. For the case with $R_d^* = 0$

Plugging $R_d^* = 0$ into Eq. (4.27), we get

$$R_s^* = \widehat{R}_s, \qquad \text{if } R_d^* = 0. \tag{F.5}$$

To satisfy Eqs. (4.28) and (4.31), the inequality

$$\psi^* \rho > \theta e^{-\theta(R_s^* + R_d^*)}, \quad \text{if } R_d^* = 0 \tag{F.6}$$

needs to hold. In other words, we have

$$R_d^* = 0, \quad \text{if } -\frac{1}{\theta} \log\left(\frac{\psi^* \rho}{\theta}\right) < R_s^*. \tag{F.7}$$

Obtaining the analytical expressions for (R_s^*, R_d^*) in the above two cases, we need to examine which solution is optimal given a $\boldsymbol{\gamma}$. For the case of $R_d^* = 0$, we define

$$\zeta_{\boldsymbol{\gamma}} \triangleq \psi^* \rho - \theta e^{-\theta(R_s^* + R_d^*)} > 0.$$

Applying $\zeta_{\boldsymbol{\gamma}}$ to Eq. (4.30), we derive

$$\begin{aligned} &\partial \ell_{R_s}(R_s^*, R_d^*; \psi^*) \\ &= \begin{cases} \left\{\zeta_{\boldsymbol{\gamma}} - \psi^* \eta_{\alpha(j)}\right\}, & \text{if } r_{\alpha(j)} < R_s^* < r_{\alpha(j-1)} \ \& \ \mathcal{N} \geq j \geq 2; \\ \left[\zeta_{\boldsymbol{\gamma}} - \psi^* \eta_{\alpha(j)},\ \zeta_{\boldsymbol{\gamma}} - \psi^* \eta_{\alpha(j-1)}\right], & \text{if } R_s^* = r_{\alpha(j)} \ \& \ \mathcal{N} \geq j \geq 1, \end{cases} \end{aligned} \tag{F.8}$$

Note that the solutions of R_s^* to $0 \in \partial \ell_{R_s}(R_s^*, R_d^*; \psi^*)$ under Eqs. (F.2) and (F.8) are $\widetilde{R}_s$ and $\widehat{R}_s$, respectively. Then, comparing Eq. (F.2) with Eq. (F.8) under $\zeta_{\boldsymbol{\gamma}} > 0$, we must have

$$\begin{cases} \widetilde{R}_s \geq \widehat{R}_s \iff R_d^* = 0; \\ \widetilde{R}_s \leq \widehat{R}_s \iff R_d^* > 0. \end{cases} \tag{F.9}$$

Combing Eqs. (F.3), (F.5), and (F.9), we obtain Eq. (4.32). Further summarizing

Eqs. (F.4), (F.7), and (F.9), we get Eq. (4.33). In addition, the definition of the group loss rate given by Eq. (4.9) indicates that Eq. (4.35) is required by Eq. (4.29), which completes the proof of Theorem 1. □

APPENDIX G

PROOF OF THEOREM 3

Proof. We define $\widetilde{\psi} \triangleq \psi^* \rho/\theta$. Plugging $\psi^* = \widetilde{\psi}\theta/\rho$ and Eq. (4.30) into Eqs. (4.27), the condition $0 \in \partial \ell_{R_s}(R_s^*, R_d^*; \psi^*)$ for the optimal solution becomes

$$\begin{cases} 0 &= \widetilde{\psi}\left(1 - \frac{\eta_{\alpha(j)}}{\rho}\right) - 1, & \text{if } r_{\alpha(j)} < R_s^* < r_{\alpha(j-1)} \ \& \ \mathcal{N} \geq j \geq 2; \\ 0 &\in \left[\widetilde{\psi}\left(1 - \frac{\eta_{\alpha(j+1)}}{\rho}\right) - 1,\ \widetilde{\psi}\left(1 - \frac{\eta_{\alpha(j)}}{\rho}\right) - 1\right], & \text{if } R_s^* = r_{\alpha(j)} \ \& \ \mathcal{N} \geq j \geq 1, \end{cases} \tag{G.1}$$

Further defining $\eta_{\text{th}} \triangleq \rho(1 - 1/\widetilde{\psi})$ and applying $\widetilde{\psi} = \rho/(\rho - \eta_{\text{th}})$ into Eq. (G.1), we have

$$\begin{cases} 0 &= \frac{\rho - \eta_{\alpha(j)}}{\rho - \eta_{\text{th}}} - 1, & \text{if } r_{\alpha(j)} < R_s^* < r_{\alpha(j-1)} \ \& \ \mathcal{N} \geq j \geq 2; \\ 0 &\in \left[\frac{\rho - \eta_{\alpha(j+1)}}{\rho - \eta_{\text{th}}} - 1,\ \frac{\rho - \eta_{\alpha(j)}}{\rho - \eta_{\text{th}}} - 1\right], & \text{if } R_s^* = r_{\alpha(j)} \ \& \ \mathcal{N} \geq j \geq 1, \end{cases} \tag{G.2}$$

The monotonic property of $\eta_{\alpha(j)}$ shown in Eq. (4.17) suggests that there exists a unique integer $\overline{k}$ such Eq. (4.42) holds. Applying $\overline{k}$ and solving Eq. (G.2), we get the optimal sending rate R_s^* expressed by Eq. (4.41). For $\eta_{\text{th}} \in (0, \rho)$, we have $\widetilde{\psi} \in (1, \infty)$ accordingly. Plugging $\widetilde{\psi} > 1$ and $\psi^* = \widetilde{\psi}\theta/\rho$ into Eq. (4.28) and letting $\theta \to 0$, we derive $R_d^* = 0$ for all $\boldsymbol{\gamma}$, as given in Eq. (4.40).

Lemma 3 shows that $q_0(\eta_{\text{th}}, \zeta_{\boldsymbol{\gamma}})$ is a decreasing function of η_{th} with $\lim_{\eta_{\text{th}} \to 0} q_0(\eta_{\text{th}}, 1) = \widetilde{q}_0 > q_{\text{th}}$. Moreover, letting $\eta_{\text{th}} \to \rho = N(1 - q_{\text{th}})$ in Eq. (H.3), we can obtain $\lim_{\eta_{\text{th}} \to \rho} q_0(\eta_{\text{th}}, 1) \leq q_{\text{th}}$. Therefore, there must exist certain η_{th} and $\zeta_{\boldsymbol{\gamma}}$ such that $q_0|_{(R_s, R_d) = (R_s^*, R_d^*)} = q_{\text{th}}$ holds, which is equivalent to Eq. (4.29). Since the policy characterized by Eq. (4.40)-(4.42) satisfy all three conditions given by Eqs. (4.27)-(4.28), it is the optimal solution to $\boldsymbol{P1}$-b under $\theta \to 0$. The proof of Theorem 3 is completed.

□

APPENDIX H

PROOF OF LEMMA 3

Proof. For any multicast policy characterized by Eqs. (4.40)-(4.42), we can derive

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\gamma}}\Big\{\widetilde{g}_{\text{sum}}(R_s^*)\Big\} &= \mathbb{E}_{\boldsymbol{\gamma}}\left\{Nr_{\alpha(\mathcal{N})}+\sum_{j=\overline{k}+1}^{\mathcal{N}}\eta_{\alpha(j)}\left(r_{\alpha(j-1)}-r_{\alpha(j)}\right)+\eta_{\alpha(\overline{k})}(R_s^*-r_{\alpha(\overline{k})})\right\}\\ &\overset{(a)}{\geq} \eta_{\text{th}}\mathbb{E}_{\boldsymbol{\gamma}}\{R_s^*\} = \frac{\eta_{\text{th}}}{N\left(1-q_0(\eta_{\text{th}},\varsigma_{\boldsymbol{\gamma}})\right)}\mathbb{E}_{\boldsymbol{\gamma}}\Big\{\widetilde{g}_{\text{sum}}(R_s^*)\Big\},\end{aligned} \tag{H.1}$$

where (a) follows because of $\eta_{\alpha(j)} \geq \eta_{\text{th}}$ for $j = \overline{k}+1,\ldots,\mathcal{N}$. Through Eq. (H.1), we get

$$\eta_{\text{th}} \leq N\left(1-q_0(\eta_{\text{th}},\varsigma_{\boldsymbol{\gamma}})\right). \tag{H.2}$$

We use the superscripts $(\cdot)'$ and $(\cdot)''$ to mark the corresponding variables for the policies under η'_{th} and η''_{th}, respectively. Following Eq. (H.1), we get

$$\begin{cases}\mathbb{E}_{\boldsymbol{\gamma}}\Big\{\widetilde{g}_{\text{sum}}(R'_s)\Big\} = \mathbb{E}_{\boldsymbol{\gamma}}\left\{Nr_{\alpha(\mathcal{N})}+\sum_{j=\overline{k}'+1}^{\mathcal{N}}\eta_{\alpha(j)}\left(r_{\alpha(j-1)}-r_{\alpha(j)}\right)+\eta_{\alpha(\overline{k}')}\left(R'_s-r_{\alpha(\overline{k}')}\right)\right\};\\ \mathbb{E}_{\boldsymbol{\gamma}}\Big\{\widetilde{g}_{\text{sum}}(R''_s)\Big\} = \mathbb{E}_{\boldsymbol{\gamma}}\left\{Nr_{\alpha(\mathcal{N})}+\sum_{j=\overline{k}''+1}^{\mathcal{N}}\eta_{\alpha(j)}\left(r_{\alpha(j-1)}-r_{\alpha(j)}\right)+\eta_{\alpha(\overline{k}'')}\left(R''_s-r_{\alpha(\overline{k}'')}\right)\right\}.\end{cases}$$

Then, we derive

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\gamma}}\Big\{\widetilde{g}_{\text{sum}}(R''_s)\Big\}-\mathbb{E}_{\boldsymbol{\gamma}}\Big\{\widetilde{g}_{\text{sum}}(R'_s)\Big\} = \mathbb{E}_{\boldsymbol{\gamma}}\Bigg\{&\eta_{\alpha(\overline{k}')}\left(r_{\alpha(\overline{k}'-1)}-R'_s\right)\\ &+\sum_{j=\overline{k}''+1}^{\overline{k}'-1}\eta_{\alpha(j)}\left(r_{\alpha(j-1)}-r_{\alpha(j)}\right)+\eta_{\alpha(\overline{k}'')}\left(R''_s-r_{\alpha(\overline{k}'')}\right)\Bigg\}.\end{aligned} \tag{H.3}$$

Since $\eta'_{\text{th}} > \eta''_{\text{th}}$, by applying the monotonic property of $\eta_{\alpha(j)}$ given by Eq. (4.17) and

the criterion to determine $\overline{k}$ in Eq. (4.42), we must have $\overline{k}' \leq \overline{k}''$ and

$$\eta'_{\text{th}} \geq \eta_{\alpha(\overline{k}')} \geq \eta_{\alpha(j)} \geq \eta_{\alpha(\overline{k}'')}, \qquad \forall\, j,\ \overline{k}' \geq j \geq \overline{k}''. \tag{H.4}$$

Plugging Eq. (H.4) into Eq. (H.3), we obtain

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{\gamma}}\Big\{\widetilde{g}_{\text{sum}}(R''_s)\Big\} - \mathbb{E}_{\boldsymbol{\gamma}}\Big\{\widetilde{g}_{\text{sum}}(R'_s)\Big\} \\
&\leq\ \eta'_{\text{th}}\mathbb{E}_{\boldsymbol{\gamma}}\left\{\left(r_{\alpha(\overline{k}'-1)} - R'_s\right) + \textstyle\sum_{j=\overline{k}''+1}^{\overline{k}'-1}\left(r_{\alpha(j-1)} - r_{\alpha(j)}\right) + \left(R''_s - r_{\alpha(\overline{k}'')}\right)\right\} \\
&=\ \eta'_{\text{th}}\Big(\mathbb{E}_{\boldsymbol{\gamma}}\{R''_s\} - \mathbb{E}_{\boldsymbol{\gamma}}\{R'_s\}\Big) \overset{(a)}{\leq}\ N(1 - q_0(\eta'_{\text{th}}, \varsigma'_{\boldsymbol{\gamma}}))\Big(\mathbb{E}_{\boldsymbol{\gamma}}\{R''_s\} - \mathbb{E}_{\boldsymbol{\gamma}}\{R'_s\}\Big),
\end{aligned} \tag{H.5}$$

where (a) results from Eq. (H.2). Furthermore, we derive

$$\begin{aligned}
&N(1 - q_0(\eta''_{\text{th}}, \varsigma''_{\boldsymbol{\gamma}})) \\
&\quad= \frac{\mathbb{E}_{\boldsymbol{\gamma}}\{\widetilde{g}_{\text{sum}}(R''_s)\}}{\mathbb{E}_{\boldsymbol{\gamma}}\{R''_s\}} \\
&\quad= \frac{\mathbb{E}_{\boldsymbol{\gamma}}\{\widetilde{g}_{\text{sum}}(R'_s)\} + \mathbb{E}_{\boldsymbol{\gamma}}\{\widetilde{g}_{\text{sum}}(R''_s)\} - \mathbb{E}_{\boldsymbol{\gamma}}\{\widetilde{g}_{\text{sum}}(R'_s)\}}{\mathbb{E}_{\boldsymbol{\gamma}}\{R'_s\} + \mathbb{E}_{\boldsymbol{\gamma}}\{R''_s\} - \mathbb{E}_{\boldsymbol{\gamma}}\{R'_s\}} \\
&\quad\leq \frac{N(1 - q_0(\eta'_{\text{th}}, \varsigma'_{\boldsymbol{\gamma}}))\mathbb{E}_{\boldsymbol{\gamma}}\{R'_s\} + N(1 - q_0(\eta'_{\text{th}}, \varsigma'_{\boldsymbol{\gamma}}))\Big(\mathbb{E}_{\boldsymbol{\gamma}}\{R''_s\} - \mathbb{E}_{\boldsymbol{\gamma}}\{R'_s\}\Big)}{\mathbb{E}_{\boldsymbol{\gamma}}\{R'_s\} + \mathbb{E}_{\boldsymbol{\gamma}}\{R''_s - R'_s\}} \\
&\quad= N(1 - q_0(\eta'_{\text{th}}, \varsigma'_{\boldsymbol{\gamma}})),
\end{aligned} \tag{H.6}$$

implying that Eq. (4.43) holds. Given $\eta_{\text{th}} \to N$, Eq. (4.42) suggests $R_s^* = r_{\alpha(\mathcal{N})} = r_{\pi(N)}$, because $\eta_{\alpha(\mathcal{N}+1)} = N$ and $\eta_{\alpha(j)} < N$ for all $j = 1, 2, \ldots, \mathcal{N}$ (see Eq. (4.17)). With $R_s^* = r_{\pi(N)}$ and $R_d^* = 0$, there is no loss for all multicast receivers and thus we have $\lim_{\eta_{\text{th}} \to N} q_0(\eta_{\text{th}}, \varsigma_{\boldsymbol{\gamma}}) = 0$. In contrast, as $\eta_{\text{th}} \to 0$, Eq. (4.42) reduces to Eq. (4.23), implying that $R_s^* = \widetilde{R}_s$. Then, together with $R_d^* = 0$, we get $\lim_{\eta_{\text{th}} \to 0} q_0(\eta_{\text{th}}, 1) = \widetilde{q}_0$, which completes the proof of Lemma 3. $\square$

APPENDIX I

PROOF OF THEOREM 4

Proof. We define $\kappa \triangleq -\log(\psi^*\rho/\theta)/\theta$. Plugging $\psi^* = \theta e^{-\theta\kappa}/\rho$ and Eq. (4.30) into Eq. (4.27) and letting $\theta \to 0$, the condition of $0 \in \partial\ell_{R_s}(R_s^*, R_d^*; \psi^*)$ reduces to

$$\begin{cases} 0 &= \eta_{\alpha(j)}, & \text{if } r_{\alpha(j)} < R_s^* < r_{\alpha(j-1)} \ \& \ \mathcal{N} \geq j \geq 2; \\ 0 &\in \left[-\eta_{\alpha(j+1)}, -\eta_{\alpha(j)}\right], & \text{if } R_s^* = r_{\alpha(j)} \ \& \ \mathcal{N} \geq j \geq 1. \end{cases} \tag{I.1}$$

Solving this equation, we get $R_s^* = \widetilde{R}_s$. Moreover, applying $\psi^* = \theta e^{-\theta\kappa}/\rho$ into Eqs. (4.28) and (4.31) under $\theta \to 0$, the condition $0 \in \partial\ell_{R_d}(R_s^*, R_d^*; \psi^*)$ changes to

$$\begin{cases} e^{-\theta R_s^*} &< e^{-\theta\kappa}, & \text{if } R_d^* = 0; \\ 0 &= e^{-\theta(R_s^*+R_d^*)} - e^{-\theta(R_s+R_d)}, & \text{if } R_d^* > 0, \end{cases} \tag{I.2}$$

solving which we obtain

$$R_d^* = \begin{cases} 0, & \text{if } \widetilde{R}_s \geq \kappa; \\ \kappa - R_d^*, & \text{if } \widetilde{R}_s \leq \kappa. \end{cases} \tag{I.3}$$

Summarizing the above results, we get Eq. (4.46). Also, $q_0|_{(R_s,R_d)=(R_s^*,R_d^*)} = q_{\text{th}}$ has to hold, as required by Eq. (4.35) in Theorem 1. Because the larger pre-drop rate improves the instantaneous throughput but does not affect the instantaneous sum goodput, the group loss rate under the policy given by Eq. (4.46) is an increasing function of κ. Eq. (4.46) also implies that the average sum goodput vary continuously with κ. Therefore, given any $q_{\text{th}} \geq \widetilde{q}_0$, there must exist a certain κ to guarantee $q_0|_{(R_s,R_d)=(R_s^*,R_d^*)} = q_{\text{th}}$. The proof of Theorem 4 is completed. □

APPENDIX J

PROOF OF THEOREM 5

Proof. Recall that we define $\xi \triangleq \log(\psi^*\rho/\theta)/\theta$. Then, rewriting the conditions given by Eqs. (4.27) and (4.28) in terms of ξ, we get

$$\begin{cases} 0 = e^{-\theta\xi}\left(1 - \frac{\eta_{\alpha(j)}}{\rho}\right) - e^{-\theta(R_s^*+R_d^*)}, & \text{if } r_{\alpha(j)} < R_s^* < r_{\alpha(j-1)} \ \& \ \mathcal{N} \geq j \geq 2; \\ 0 \in \Big[e^{-\theta\xi}\left(1 - \frac{\eta_{\alpha(j+1)}}{\rho}\right) - e^{-\theta(R_s^*+R_d^*)}, & \\ \qquad\qquad e^{-\theta\xi}\left(1 - \frac{\eta_{\alpha(j)}}{\rho}\right) - e^{-\theta(R_s^*+R_d^*)}\Big], & \text{if } R_s^* = r_{\alpha(j)} \ \& \ \mathcal{N} \geq j \geq 1. \end{cases} \tag{J.1}$$

and

$$\begin{cases} 0 \ < \ e^{-\theta\xi} - e^{-\theta R_s^*}, & \text{if } R_d^* = 0; \\ 0 \ = \ \xi - (R_s^* + R_d^*), & \text{if } R_d^* > 0, \end{cases} \tag{J.2}$$

respectively. Then, the solution can be categorized into the following five cases.

A. For the case of $R_s^* = r_{\alpha(j)}$ and $R_s^* + R_d^* = \xi$

In order to meet the second line of Eq. (J.1), $1 - \eta_{\alpha(j+1)}/\rho < 1 \leq 1 - \eta_{\alpha(j)}/\rho$ has to hold, which is equivalent to $\eta_{\alpha(j+1)} > 0 \geq \eta_{\alpha(j)}$, implying $r_{\alpha(j)} = \widetilde{R}_s$. As a result, we get

$$(R_s^*, R_d^*) = (\widetilde{R}_s, \xi - \widetilde{R}_s), \quad \text{if } \left(R_s^* = r_{\alpha(j)}\right) \wedge (R_s^* + R_d^* = \xi) \tag{J.3}$$

for some j. Further applying $R_s^* + R_d^* = \xi$ into Eq. (J.2), we get $R_d^* > 0$ and thus

$$\widetilde{R}_s < \xi. \tag{J.4}$$

B. For the case of $R_s^* = r_{\alpha(j)}$ and $R_s^* + R_d^* > \xi$

The second line of Eq. (J.1) implies that $1 - \eta_{\alpha(j+1)}/\rho < 0 \leq 1 - \eta_{\alpha(j)}/\rho$ must hold, which is equivalent to $\eta_{\alpha(j+1)} > \rho \geq \eta_{\alpha(j)}$. We use $\widehat{k}$ to denote the unique integer j to satisfy $\eta_{\alpha(j+1)} > \rho \geq \eta_{\alpha(j)}$. Then, we get $R_s^* = R_\rho$, where R_ρ is given by Eq. (4.48). Moreover, the condition $R_s^* + R_d^* > \xi$ requires that $R_d^* = 0$ according to Eq. (J.2). Summarizing these results, we obtain

$$(R_s^*, , R_d^*) = (R_\rho, 0), \quad \text{if } \left(R_s^* = r_{\alpha(j)}\right) \wedge \left(R_s^* + R_d^* > \xi\right), \tag{J.5}$$

for some j, where we must have

$$\xi < R_\rho. \tag{J.6}$$

C. For the case of $R_s^* = r_{\alpha(j)}$ and $R_s^* + R_d^* < \xi$

Under this condition, it is easy to verify that Eq. (J.1) does not have solution as $\theta \to \infty$. Thus, this optimal solution will never fall into this category.

D. For the case of $r_{\alpha(j)} < R_s^* < r_{\alpha(j-1)}$ and $\eta_{\alpha(j)} \neq 0$

If $1 - \eta_{\alpha(j)}/\rho \leq 0$, i.e., $\eta_{\alpha(j)} \geq \rho$, it is clear that the solution to Eq. (J.1) does not exist; moreover, if $\eta_{\alpha(j)} < 0$, we have

$$0 = e^{-\theta\xi}\left(1 - \frac{\eta_{\alpha(j)}}{\rho}\right) - e^{-\theta(R_s^* + R_d^*)} > e^{-\theta\xi} - e^{-\theta(R_s^* + R_d^*)}, \tag{J.7}$$

under which Eq. (J.2) does not have solution, either. Consequently, $\rho > \eta_{\alpha(j)} > 0$ needs to hold, and thus the first line of Eq. (J.1) suggests

$$R_s^* + R_d^* = \xi - \frac{1}{\theta}\log\left(1 - \frac{\eta_{\alpha(j)}}{\rho}\right) > \xi. \tag{J.8}$$

Furthermore, $R_s^* + R_d^* > \xi$ results in $R_d^* = 0$ based on Eq. (J.2). Further letting $\theta \to \infty$ in Eq. (J.8), we have $R_s^* \to \xi$. Thus, the optimal R_s^* and R_d^* are given by

$$(R_s^*, R_d^*) = (\xi, 0), \quad \text{if } \left(r_{\alpha(j)} < R_s^* < r_{\alpha(j-1)}\right) \tag{J.9}$$

for some j. In addition, since $\rho > \eta_{\alpha(j)} > 0$ holds, according to the monotonicity of $\eta_{\alpha(j)}$ we get

$$r_{\alpha(\widehat{k})} \leq R_s^* = \xi \leq \widetilde{R}_s. \tag{J.10}$$

E. For the case of $r_{\alpha(j)} < R_s^* < r_{\alpha(j-1)}$ and $\eta_{\alpha(j)} = 0$

Solving the first line of Eq. (J.1) under $\eta_{\alpha(j)} = 0$, we get $R_s^* + R_d^* = \xi > \widetilde{R}_s$. Note that the requirement $\xi > \widetilde{R}_s$ coincides with the results in case *1)*, which is given by Eq. (J.3). Therefore, when $\xi > \widetilde{R}_s$ and $\eta_{\alpha(j)} = 0$ hold for some j, the solution for R_d^* and R_s^* may not be unique. However, because the optimization problem is convex, these solutions result in the same optimal value of the objective function. Then, we ignore this case and use the solution in case *1)* to characterize the optimal multicast policy.

As discussed in the above, the solution of (R_s^*, R_d^*) will fall into cases *1)*, *2)*, or *4)*. Since $\rho = N(1-q_{\text{th}}) > 0$, we must have $r_{\alpha(i)} \leq R_\rho$. Then, comparing Eqs. (J.4), (J.6), and (J.10), we can see that cases *1)*, *2)*, and *4)* are mutual exclusive. Summarizing the results for these three cases, we get Eq. (4.47) and thus Theorem 5 follows. □

APPENDIX K

PROOF OF LEMMA 4

Proof. Optimization theory shows that [77, Chs. 4.2.1 and 5.1] the Lagrangian dual function $z(\varrho; R_s)$ is *convex*, and thus also *continuous*, over the Lagrangian multiplier ϱ. On the one hand, letting $\varrho = N - 1$ for the Greedy algorithm in Fig. I, we get $\widetilde{\mu}_{\pi(N)} = 1$ and $\widetilde{\mu}_{\pi(i)} = 0$ for $i \neq N$, resulting in $R_s - \mathcal{R}_s(N-1) = BT \log(1+\gamma_{\pi(N)}) \geq 0$. On the other hand, $\varrho \to \infty$ generates $\widetilde{\mu}_{\pi(1)} = 1$ and $\widetilde{\mu}_{\pi(i)} = 0$ for $i \neq 1$, resulting in $R_s - \mathcal{R}_s(N-1) = R_s - BT \log(1 + \gamma_{\pi(1)}) \leq 0$. Consequently, we can always find a certain $\overline{\varrho}$ to satisfy Eq. (4.56).

Lagrangian duality theory also shows that [79, Th. 6.3.4] $R_s - \mathcal{R}_s(\varrho)$ is a subgradient of the Lagrangian dual function $z(\varrho; R_s)$, i.e., $R_s - \mathcal{R}_s(\varrho) \in \partial z(\varrho; R_s)$, where $\partial z(\varrho; R_s)$ is the subdifferential of $z(\varrho; R_s)$ w.r.t. ϱ. Since we have $R_s - \mathcal{R}_s(\overline{\varrho}) = 0$, the equation $0 \in z(\overline{\varrho}; R_s)$ holds, implying that $\overline{\varrho}$ is the minimizer of the Lagrangian dual problem given by $\min_{\varrho}\{z(\varrho; R_s)\}$. Then, we have

$$\overline{\varrho} = \arg\min_{\varrho} \{z(\varrho; R_s)\}. \tag{K.1}$$

Note that when $R_s - \mathcal{R}_s(\overline{\varrho}) = 0$, from Eqs. (4.52), (4.53), and (K.1) we can see $\max_{\boldsymbol{\mu}}\{\nu(\boldsymbol{\mu}; \varrho, R_s)\} = \min\{z(\overline{\varrho}; R_s)\}$, suggesting there is no duality gap [79, Chs. 6.1-6.2] between the primary problem $\boldsymbol{P3}$ and its dual problem. As a result, $\widetilde{\boldsymbol{\mu}}(\overline{\varrho})$ is the optimal to $\boldsymbol{P3}$, and Eq. (4.57) holds, which completes the proof of Lemma 4. □

APPENDIX L

PROOF OF LEMMA 5

Proof. In order to prove the concavity of $\widetilde{g}_{\text{sum}}(R_s)$, we need to show that for any $R_s \in [BT\log(1+\gamma_{\pi(N)}), BT\log(1+\gamma_{\pi(1)})]$, we can find a real number u to satisfy the following equality [79, Th. 3.2.6]:

$$\widetilde{g}_{\text{sum}}(R_s') \leq \widetilde{g}_{\text{sum}}(R_s) + u(R_s' - R_s), \tag{L.1}$$

for all $R_s' \in [BT\log(1+\gamma_{\pi(N)}), BT\log(1+\gamma_{\pi(1)})]$. For convenience, we use $\overline{\varrho}\,'$ to denote any solution to Eq. (4.56) under $R_s = R_s'$. Then, we derive

$$\begin{aligned}
\widetilde{g}_{\text{sum}}(R_s') - \widetilde{g}_{\text{sum}}(R_s) &\overset{(a)}{=} z\big(\overline{\varrho}\,'; R_s'\big) - z\big(\overline{\varrho}\,; R_s\big) \\
&\overset{(b)}{\leq} z\big(\overline{\varrho}\,; R_s'\big) - z\big(\overline{\varrho}\,; R_s\big) \\
&\overset{(c)}{=} \mathcal{G}_{\text{sum}}\big(\overline{\varrho}\,\big) + \overline{\varrho}\,\Big(R_s' - \mathcal{R}_s\big(\overline{\varrho}\,\big)\Big) - \Big[\mathcal{G}_{\text{sum}}\big(\overline{\varrho}\,\big) + \overline{\varrho}\,\Big(R_s - \mathcal{R}_s\big(\overline{\varrho}\,\big)\Big)\Big] \\
&= \overline{\varrho}\,(R_s' - R_s),
\end{aligned} \tag{L.2}$$

where (a) holds by applying Eq. (4.57), (b) follows because $\overline{\varrho}\,'$ minimizes $z\big(\varrho; R_s'\big)$ as shown in Eq. (K.1), and (c) results from Eq. (4.55). Since Eq. (L.2) is equivalent to Eq. (L.1) with $w = \overline{\varrho}$, we have proven the concavity of $\widetilde{g}_{\text{sum}}(R_s)$.

Based on Definition 9 and Eq. (L.2), any $\overline{\varrho}$ satisfying Eq. (4.56) is a subgradient of $\widetilde{g}_{\text{sum}}(R_s)$. Thus, we have

$$\partial\widetilde{g}_{\text{sum}}(R_s) \supset \{\overline{\varrho}\,|\mathcal{R}_s(\overline{\varrho}\,) = R_s\}. \tag{L.3}$$

On the other hand, consider any $u \in \widetilde{g}_{\text{sum}}(R_s)$. Because $\widetilde{g}_{\text{sum}}(R_s)$ is a concave function, $\partial\widetilde{g}_{\text{sum}}(R_s)$ must be a convex set [79, Ch. 3.2.3] and thus $w \in [\overline{\varrho}_{\min}, \overline{\varrho}_{\max}]$, where $\overline{\varrho}_{\min}$

and $\overline{\varrho}_{\max}$ represent the minimum and maximum elements in $\widetilde{g}_{\text{sum}}(R_s)$, respectively. As discussed in the proof of Lemma 4, $z(\varrho; R_s)$ is a convex function of ϱ with a subgradient equal to $R_s - \mathcal{R}_s(\varrho)$, implying that $\mathcal{R}_s(\varrho)$ must be a monotonically decreasing function of ϱ. Then, we can get $R_s = \mathcal{R}_s(\overline{\varrho}_{\max}) \leq \mathcal{R}_s(w) \leq \mathcal{R}_s(\overline{\varrho}_{\min}) = R_s$, resulting in $\mathcal{R}_s(w) = R_s$ and thus

$$\partial \widetilde{g}_{\text{sum}}(R_s) \subset \{\overline{\varrho} \,|\mathcal{R}_s(\overline{\varrho}) = R_s\}. \tag{L.4}$$

Combining Eqs. (L.3)-(L.4), we get Eq. (4.58). The proof of Lemma 5 is completed. □

APPENDIX M

Derivations of Eqs. (5.52)-(5.53)

If $t^* > 0$ is given, then w becomes convex over (z_ℓ, R_ℓ), and thus (z_ℓ^*, R_ℓ^*) must satisfy

$$\begin{cases} \left.\frac{\partial w}{\partial z_\ell}\right|_{z_\ell=z_\ell^*} = 0, & \text{if } z_\ell^* > 0; \\ \left.\frac{\partial w}{\partial z_\ell}\right|_{z_\ell=z_\ell^*} > 0, & \text{if } z_\ell^* = 0, \end{cases} \qquad \text{and} \qquad 0 \in \partial w_{R_\ell} \tag{M.1}$$

To obtain Eqs. (5.52)-(5.53), we need to study the cases with $\phi_\ell = 0$ and $\phi_\ell > 0$, respectively, as follows.

A. For the case of $\phi_\ell > 0$:

Based on Eqs. (5.47)-(5.48), for any given $\phi_\ell > 0$ and z_ℓ^*, the variable x^* to satisfy $0 \in \partial w_{R_\ell}$ is derived by

$$x^* = \frac{1}{\phi_\ell}\left(\phi_\ell N\left(1 - q_{\text{th}}^{(\ell)}\right) - \lambda_\ell \beta_\ell e^{-\beta_\ell (z_\ell^* + t_\ell R_\ell)}\right). \tag{M.2}$$

Moreover, because $\widetilde{g}_s(R_\ell)$ is a concave function and $\widetilde{R}$ maximizes $\widetilde{g}_s(R_\ell)$, we have

$$R_\ell^* < \widetilde{R}, \ \text{ if } x^* > 0; \qquad R_\ell^* = \widetilde{R}, \ \text{ if } x^* = 0; \qquad R_\ell^* > \widetilde{R}, \ \text{ if } x^* < 0. \tag{M.3}$$

Comparing Eq. (5.48) with Eq. (5.50), we get

$$\left.\frac{\partial w}{\partial z_\ell}\right|_{z_\ell=z_\ell^*} = x^* \phi_\ell. \tag{M.4}$$

Then, we need to further consider the following two different cases:

1. If $x^* > 0$

Eq. (M.4) results in $(\partial w/\partial z_\ell)\,|_{z_\ell=z_\ell^*} > 0$. Comparing this with Eq. (M.1), we get $z_\ell^* = 0$, which implies $R_\ell^* = \widehat{R}_\ell$ according to Eq. (M.1). Moreover, Eq. (M.3) shows that $\widehat{R}_\ell = R_\ell^* < \widetilde{R}$ when $x^* > 0$. Thus, if $\widehat{R}_\ell < \widetilde{R}$, we have $R_\ell^* = \widehat{R}_\ell$ and $z_\ell^* = 0$, which generates Eq. (5.53).

2. If $x^* \leq 0$

By Eq. (M.4), we get the $(\partial w/\partial z_\ell)\,|_{z_\ell=z_\ell^*} \leq 0$. Comparing this with Eq. (M.1), we obtain $z_\ell^* > 0$ and $(\partial w/\partial z_\ell)\,|_{z_\ell=z_\ell^*} = 0$. Plugging $(\partial w/\partial z_\ell)\,|_{z_\ell=z_\ell^*} = 0$ into Eq. (M.4), x^* is set to be zero, resulting in $R_\ell^* = \widetilde{R}$ based on Eq. (M.3). Then, solving $\partial w/\partial z_\ell = 0$, we obtain Eq. (5.52).

B. For the case of $\phi_\ell = 0$:

When $\phi_\ell = 0$, the solution to Eq. (M.1) only needs to satisfy $(z_\ell^* + t_\ell^* R_\ell^*) = \infty$. However, since R_ℓ is upper-bounded by $R_{\pi(1)}$, $z_\ell^* = \infty$ must hold. Plugging $z_\ell^* = \infty$ into $0 \in \partial w_{R_\ell}$, we get $R_\ell^* = \widetilde{R}$. Combining the results for $\phi_\ell = 0$ and $\phi_\ell > 0$, we complete the derivations of Eqs. (5.52)-(5.53).

APPENDIX N

Derivations of $\left(t_\ell^{(2)}, R_\ell^{(2)}, z_\ell^{(2)}\right)$ under $R_\ell = \widetilde{R}$ and $\left(t_\ell^{(3)}, R_\ell^{(3)}, z_\ell^{(3)}\right)$ under $z_\ell = 0$

A. Derivation of $\left(t_\ell^{(2)}, R_\ell^{(2)}, z_\ell^{(2)}\right)$ under $R_\ell = \widetilde{R}$

When $R_\ell = \widetilde{R}$ is given, we have $R_\ell^{(2)} = \widetilde{R}$ and w is convex over the 2-tuple (t_ℓ, z_ℓ), where $\widetilde{R}$ is determined by Eq. (5.55). Then, the minimizer $\left(t_\ell^{(2)}, \widetilde{R}, z_\ell^{(2)}\right)$ is the unique solution to $\partial w/\partial t_\ell = 0$ and $\partial w/\partial z_\ell = 0$ under the boundary conditions $t_\ell \geq 0$ and $z_\ell \geq 0$. More specifically, we need to guarantee the followings:

$$\begin{cases} \left.\frac{\partial w}{\partial t_\ell}\right|_{t_\ell = t_\ell^{(2)}} > 0, & \text{if } t_\ell^{(2)} = 0; \\ \left.\frac{\partial w}{\partial t_\ell}\right|_{t_\ell = t_\ell^{(2)}} = 0, & \text{if } t_\ell^{(2)} > 0; \end{cases} \quad \text{and} \quad \begin{cases} \left.\frac{\partial w}{\partial z_\ell}\right|_{z_\ell = z_\ell^{(2)}} > 0, & \text{if } z_\ell^{(2)} = 0; \\ \left.\frac{\partial w}{\partial z_\ell}\right|_{z_\ell = z_\ell^{(2)}} = 0, & \text{if } z_\ell^{(2)} > 0. \end{cases} \tag{N.1}$$

Then, applying Eqs. (5.49)-(5.50), we derive the expressions of $\left(t_\ell^{(2)}, \widetilde{R}, z_\ell^{(2)}\right)$ for the following three different cases.

1. For the case of $1 + \psi_{\boldsymbol{\gamma}} \leq \phi_\ell \widetilde{g}_{\mathrm{s}}(\widetilde{R}) - \phi_\ell N \left(1 - q_{\mathrm{th}}^{(\ell)}\right) \widetilde{R}$:

In this case, we get $\partial w/\partial t_\ell < 0$ for all $t_\ell > 0$, resulting in $t_\ell^{(2)} = \infty$. Moreover, $t_\ell^{(2)} = \infty$ leads to $\partial w/\partial z_\ell > 0$ for all z_ℓ based on Eq. (5.50), which implies $z_\ell^{(2)} = 0$ in order to satisfy Eq. (N.1). Thus in this case, we get $t_\ell^{(2)} = \infty$ and $z_\ell^{(2)} = 0$.

2. For the case of $\phi_\ell \widetilde{g}_{\mathrm{s}}(\widetilde{R}) - \phi_\ell N \left(1 - q_{\mathrm{th}}^{(\ell)}\right) \widetilde{R} < 1 + \psi_{\boldsymbol{\gamma}} < \phi_\ell \widetilde{g}_{\mathrm{s}}(\widetilde{R})$:

This condition guarantees that $\partial w/\partial t_\ell < \widetilde{R} \times \partial w/\partial z_\ell$. Comparing this condition with Eq. (N.1), the equation $z_\ell^{(2)} = 0$ needs to hold. Then, solving $\partial w/\partial t_\ell = 0$, we get

$$t_\ell^{(2)} = \left[\frac{1}{-\beta_\ell \widetilde{R}} \log \left(\frac{1 + \psi_{\boldsymbol{\gamma}} - \phi_\ell \widetilde{g}_{\mathrm{s}}(\widetilde{R}) + \phi_\ell N \left(1 - q_{\mathrm{th}}^{(\ell)}\right) \widetilde{R}}{\lambda_\ell \beta_\ell \widetilde{R}} \right) \right]^+ . \tag{N.2}$$

3. For the case of $1+\psi_{\boldsymbol{\gamma}} \geq \phi_\ell \widetilde{g}_{\mathrm{s}}(\widetilde{R})$:

This condition is equivalent to $\partial w/\partial t_\ell \geq \widetilde{R} \times \partial w/\partial z_\ell$. Comparing this condition with Eq. (N.1), the equation $t_\ell^{(2)} = 0$ has to be satisfied. Then, solving $\partial w/\partial z_\ell = 0$, we get

$$z_\ell^{(2)} = \left[\frac{1}{-\beta_\ell} \log \left(\frac{\phi_\ell N \left(1 - q_{\mathrm{th}}^{(\ell)}\right)}{\lambda_\ell \beta_\ell} \right) \right]^+ . \tag{N.3}$$

Combining the results for the above three cases, the derivation for $\left(t_\ell^{(2)}, R_\ell^{(2)}, z_\ell^{(2)}\right)$ is completed.

B. Derivation of $\left(t_\ell^{(3)}, R_\ell^{(3)}, z_\ell^{(3)}\right)$ under $z_\ell = 0$

According to Eqs. (5.35) and (5.36), the domain of R_ℓ is the union of a number of sets, which are denoted by $\Omega \triangleq \{r_i, i = 1, 2, \ldots, \mathcal{N}\}$ and $\Lambda_j \triangleq (r_j, r_{j-1})$, $j = 2, 3, \ldots, \mathcal{N}$. We can first find the minimizer in each of the above sets, and then compare these minimizers obtained to get the minimizer over the entire domain.

1. For the case of $R_\ell \in \Omega$:

Given $R_\ell = r_i$ for some r_i, the time slot (at the ℓth video layer) minimizing w, denoted by $t_{\ell,\{i\}}$, is obtained through solving $\partial w/\partial t_\ell = 0$, which is given by

$$t_{\ell,\{i\}} = \begin{cases} \infty, & \text{if } 1+\psi_{\boldsymbol{\gamma}} \leq \phi_\ell \widetilde{g}_{\mathrm{s}}(r_i) - \phi_\ell N \left(1 - q_{\mathrm{th}}^{(\ell)}\right) r_i; \\ \left[\frac{1}{-\beta_\ell r_i} \log \left(\frac{1+\psi_{\boldsymbol{\gamma}} - \phi_\ell \widetilde{g}_{\mathrm{s}}(r_i) + \phi_\ell N \left(1 - q_{\mathrm{th}}^{(\ell)}\right) r_i}{\lambda_\ell \beta_\ell r_i} \right) \right]^+, & \text{if } 1+\psi_{\boldsymbol{\gamma}} > \phi_\ell \widetilde{g}_{\mathrm{s}}(r_i) - \phi_\ell N \left(1 - q_{\mathrm{th}}^{(\ell)}\right) r_i. \end{cases} \tag{N.4}$$

2. For the case of $R_\ell \in [r_j, r_{j-1}]$:

When $R_\ell \in \Lambda_j = (r_j, r_{j-1})$, $\widetilde{g}_s(R_\ell)$ is differentiable and we have

$$\frac{\partial w}{\partial R_\ell} = t_\ell\Big(\phi_\ell N\left(1 - q_{\text{th}}^{(\ell)}\right) - \lambda_\ell \beta_\ell e^{-\beta_\ell t_\ell R_\ell} - \phi_\ell \eta_j\Big). \tag{N.5}$$

Therefore, the minimizer is either a stationary point of w or R_ℓ is located at the boundary, i.e, $R_\ell = r_j$ or r_{j-1}. If $\phi_\ell N\left(1 - q_{\text{th}}^{(\ell)}\right) - \phi_\ell \eta_j \leq 0$, we have $\partial w/\partial R_\ell \leq 0$ for all $R_\ell \in \Lambda_j$, implying that the minimizer must be located at the boundary of the closure of Λ_j; otherwise, jointly solving $\partial w/\partial R_\ell = 0$ and $\partial w/\partial t_\ell = 0$, we get

$$\begin{cases} t_\ell R_\ell &= -\frac{1}{\beta_\ell}\log\left(\frac{\phi_\ell N\left(1-q_{\text{th}}^{(\ell)}\right)-\phi_\ell\eta_j}{\lambda_\ell\beta_\ell}\right); \\ \frac{\partial w}{\partial t_\ell} &= 1 + \psi_{\boldsymbol{\gamma}} - \phi_\ell \widetilde{g}_s(r_j) + \phi_\ell \eta_j r_j. \end{cases} \tag{N.6}$$

In the above equation, $\partial w/\partial t_\ell$ is a constant. As a result, the minimizer t_ℓ must satisfy either $t_\ell = 0$ or $t_\ell = \infty$, both of which result in $\partial w/\partial R_\ell \geq 0$ for all $R_\ell \in \Lambda_j$, implying that the optimal R_ℓ must lie on the boundary of the closure of Λ_j.

When R_ℓ is on the boundary of the closure of Λ_j, the minimizer has been derived in Appendix 1. As a result, we do not need to examine the sets Λ_j, $j = 2, 3, \ldots, \mathcal{N}$. Accordingly, we have

$$\left(t_\ell^{(3)}, R_\ell^{(3)}, z_\ell^{(3)}\right) = \arg \min_{\left(t_{\ell,\{i\}}, r_i, 0\right),\ i=1,2,\ldots,\mathcal{N}} \Big\{w\Big\}, \tag{N.7}$$

which completes the derivation of $\left(t_\ell^{(3)}, R_\ell^{(3)}, z_\ell^{(3)}\right)$.

APPENDIX O

PROOF OF THEOREM 9

Proof. Because losses for different data packets are *i.i.d.*, we can express $\psi_1(k, \theta, \gamma, 1)$ as

$$\psi_1(k, \theta, \gamma, 1) = \frac{\lambda}{\Lambda}, \tag{O.1}$$

where λ is the number of loss patterns under which one of the lost data packets can be repaired by a single received check packet. Λ is the total number of loss patterns.

As described in Section B, in order to repair one lost data packet with one check packet for a loss pattern, the following conditions must be satisfied. 1) Among γ data packets which are connected with the same check packet in the bipartite graph, there is only one lost packet. If $\gamma > k - \theta + 1$, at least two data packets are connected with the check packet and then no losses can be repaired. 2) Among $(k - \gamma)$ data packets which are not connected with the check packet, $(\theta - 1)$ of them are other lost data packets. Based on the above discussions, we derive λ given by

$$\lambda = \begin{cases} \binom{\gamma}{1}\binom{k-\gamma}{\theta-1}, & \text{if } \gamma \leq k - \theta + 1; \\ 0, & \text{if } \gamma > k - \theta + 1. \end{cases} \tag{O.2}$$

Also, through the above definition of Λ, we have $\Lambda = \binom{k}{\theta}$. Then, using Eq. (O.1), we get

$$\psi_1(k, \theta, \gamma, 1) = \begin{cases} \frac{\theta\gamma(k-\theta)\,!(k-\gamma)!}{(k-\gamma-\theta+1)\,!k\,!}, & \text{if } \gamma \leq k - \theta + 1; \\ 0, & \text{if } \gamma > k - \theta + 1, \end{cases} \tag{O.3}$$

which completes the proof of Claim 1. For $1 \leq \gamma \leq k-1$, we define

$$\Delta(\gamma) \triangleq \psi_1(k,\theta,\gamma+1,1) - \psi_1(k,\theta,\gamma,1). \tag{O.4}$$

Plugging (O.3) into (O.4), and letting $\Delta(\gamma) \leq 0$, we derive

$$\Delta(\gamma) \leq 0 \;\; \Leftrightarrow \;\; \left\lceil \frac{(k+1)-\theta}{\theta} \right\rceil \leq \gamma \leq k-1. \tag{O.5}$$

Thus, the following inequalities hold:

$$\psi_1(k,\theta,1,1) \leq \psi_1(k,\theta,2,1) \leq \cdots \leq \psi_1\left(k,\theta,\left\lceil \frac{(k+1)-\theta}{\theta} \right\rceil,1\right); \tag{O.6}$$

$$\psi_1\left(k,\theta,\left\lceil \frac{(k+1)-\theta}{\theta} \right\rceil,1\right) \geq \psi_1\left(k,\theta,\left\lceil \frac{(k+1)-\theta}{\theta} \right\rceil+1,1\right) \geq \cdots \geq \psi_1(k,\theta,k,1). \tag{O.7}$$

Therefore, $\gamma = \lceil ((k+1)-\theta)/\theta \rceil$ maximizes $\psi_1(k,\gamma,\theta,1)$ with the given θ. Then, $\gamma_1^*(k,\theta)$ is given by

$$\gamma_1^*(k,\theta) = \arg\max_{1\leq\gamma\leq k} \psi_1(k,\theta,\gamma,1) = \left\lceil \frac{(k+1)-\theta}{\theta} \right\rceil, \tag{O.8}$$

which completes the proof of Claim 2.

Note that $\gamma > k-\theta+1 \Leftrightarrow \theta > k-\gamma+1$. Thus, we have

$$\begin{aligned}
\psi_1(k,\theta,\gamma,1) &= \begin{cases} \frac{\theta\gamma(k-\theta)\,!(k-\gamma)!}{(k-\gamma-\theta+1)\,!k\,!}, & \text{if } \gamma \leq k-\theta+1; \\ 0, & \text{if } \gamma > k-\theta+1, \end{cases} \\
&= \begin{cases} \frac{\gamma\theta(k-\gamma)\,!(k-\theta)!}{(k-\theta-\gamma+1)\,!k\,!}, & \text{if } \theta \leq k-\gamma+1; \\ 0, & \text{if } \theta > k-\gamma+1, \end{cases} \\
&= \psi_1(k,\gamma,\theta,1).
\end{aligned} \tag{O.9}$$

This proves that the dynamics of $\psi_1(k,\theta,\gamma,1)$ is symmetric with respect to θ and γ.

Note that we have $\psi_1(k, \theta, \gamma, 1) \leq 1$. Next, we solve the equation $\psi_1(k, \theta, \gamma, 1) = 1$ for $\gamma \leq k - \theta + 1$ to see whether some (θ, γ) achieves the upper bound 1. Equivalently, we need to solve $\binom{\gamma}{1}\binom{k-\gamma}{\theta-1} = \binom{k}{\theta}$. Also, because $\binom{k}{\theta} = \sum_{i=0}^{\min\{\gamma,\theta\}} \binom{\gamma}{i}\binom{k-\gamma}{\theta-i}$ holds for any $1 \leq \gamma \leq k$, we need to guarantee $\min\{\gamma, \theta\} = 1$ for $\binom{\gamma}{1}\binom{k-\gamma}{\theta-1} = \binom{k}{\theta}$. Thus, either $\theta = 1$ or $\gamma = 1$ must be satisfied. Plugging $\theta = 1$ and $\gamma = 1$ into Eq. (6.10), respectively, we can derive $\psi_1(k, 1, k, 1) = 1$ and $\psi_1(k, k, 1, 1) = 1$. Thus, 1 is the least upper bound of $\psi_1(k, \theta, \gamma, 1)$. Moreover, through Eq. (6.11), we have $\gamma_1^*(k, 1) = k$ and through and through Eq. (6.11), we have $\gamma_1^*(k, 1) = k$ and $\gamma_1^*(k, k) = 1$. Then, the proof of Theorem 9 is completed. □

APPENDIX P

PROOF OF THEOREM 10

Proof. We model the loss-covering procedure by a random process $\{X_n\}$ taking value in the state space specified by $\{0, 1, 2 \cdots, k-\theta+1\}$ as shown in Fig. 56, which describes the loss-covering status of the data packets. State i, $0 \leq i < k-\theta+1$, represents the total number i out of k data packets having been covered. State i, $i = (k-\theta+1)$, represents the *target state*, where at least $(k-\theta+1)$ data packets or, equivalently, at least one lost data packet have been covered by the generated check packets. We call the data packets which have not been covered *uncovered* data packets.

The random variable X_n denotes the covering state after the nth check packet of the current TR has been generated. If $X_{n_0-1} < k-\theta+1$ and $X_{n_0} = k-\theta+1$ for some n_0, we say that we reach the target state after n_0 check packets having been generated. It is clear that the number $T(k, \theta, \gamma)$ described in the Eq. (6.18) is equal to $E\{n_0\}$.

Next, we show $\{X_n\}$, $n \geq 0$, is a Markov Chain. Note that if $X_n = i_n$, $(k - i_n)$ equals the number of data packets which have not been covered after the nth data packet of the current TR has been generated. Then, we have

$$\begin{aligned}
&\Pr\left\{X_{n+1} = i_{n+1} \middle| X_n = i_n, X_{n-1} = i_{n-1}, \ldots, X_0 = i_0\right\} \\
&= \Pr\left\{X_{n+1} - X_n = i_{n+1} - i_n \middle| X_n = i_n, X_{n-1} = i_{n-1}, \ldots, X_0 = i_0\right\} \\
&\overset{(a)}{=} \Pr\Big\{(i_{n+1} - i_n) \text{ out of } (k - i_n) \text{ data packets } \textit{uncovered} \text{ by the previous } n \text{ check} \\
&\qquad \text{packets are covered by the } (n+1)\text{-th check packet} \Big| X_n = i_n, \\
&\qquad\qquad X_{n-1} = i_{n-1}, \ldots, X_0 = i_0\Big\}. \qquad \text{(P.1)}
\end{aligned}$$

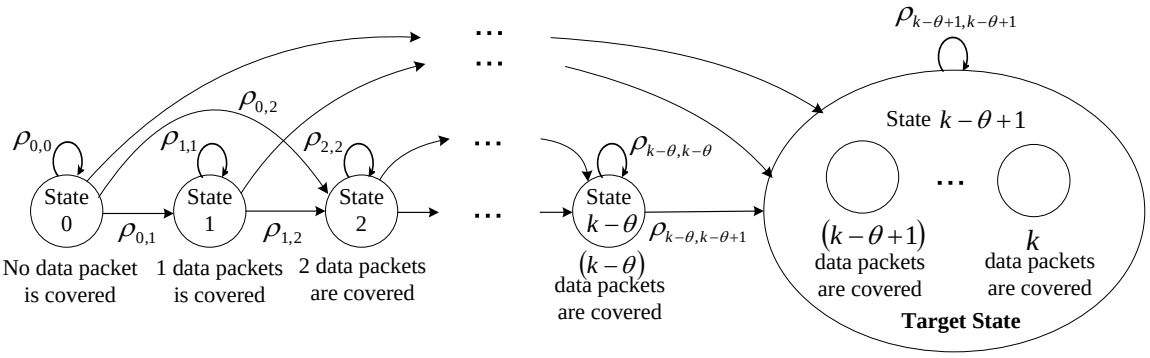


Fig. 56. State transition diagram of the Markov Chain for the covering status.

Given X_n, $(k - X_n)$ is fixed. Also, as described in Section C-2, the random construction of the check packet is independent of the constructions of other check packets. So, if X_n is given, the conditional probability in (a) of Eq. (P.1) is independent of $X_{n-1}, X_{n-2}, \ldots, X_0$. Thus, we can derive

$$\begin{aligned}
&\Pr\left\{X_{n+1} = i_{n+1} \middle| X_n = i_n, X_{n-1} = i_{n-1}, \ldots, X_0 = i_0\right\} \\
&= \Pr\Big\{(i_{n+1} - i_n) \text{ out of } (k - i_n) \text{ data packets } \textit{uncovered} \text{ by the previous } n \text{ check} \\
&\qquad \text{packets are covered by the } (n+1)\text{-th check packet} \Big| X_n = i_n\Big\}. \\
&= \Pr\left\{X_{n+1} - X_n = i_{n+1} - i_n \middle| X_n = i_n\right\} \\
&= \Pr\left\{X_{n+1} = i_{n+1} \middle| X_n = i_n\right\}. \qquad \text{(P.2)}
\end{aligned}$$

Therefore, $\{X_n\}$, $n \geq 0$, is a Markov Chain. Clearly, the Markov Chain is homogeneous in terms of n. We define the transition probability, denoted by $\rho_{i,j}$, as follows.

$$\rho_{i,j} \triangleq \Pr\left\{X_{n+1} = j \middle| X_n = i\right\}, \quad n \geq 0, 0 \leq i, j \leq k - \theta + 1. \qquad \text{(P.3)}$$

For convenience, we write Eq. (6.20) again in the following.

$$\rho_{i,j} = \begin{cases} \binom{i}{\gamma-j+i}\binom{k-i}{j-i}/\binom{k}{\gamma}, & \text{if } 0 \le j-i \le \gamma \le j \text{ and } j < k-\theta+1; \\ \sum_{v=k-\theta+1}^{\min\{i+\gamma,k\}} \frac{\binom{i}{\gamma-v+i}\binom{k-i}{v-i}}{\binom{k}{\gamma}}, & \text{if } j = k-\theta+1 \text{ and } i+\gamma \ge k-\theta+1. \\ 0, & \text{otherwise,} \end{cases}$$

Note that $\binom{i}{\gamma-j+i}\binom{k-i}{j-i}$ is the number of ways of constructing the check packet such that $X_{n+1} = j$ with given $X_n = i$, while $\binom{k}{\gamma}$ is the total number of ways constructing a check packet. Hence, $\rho_{i,j}$ is equal to the ratio of $\binom{i}{\gamma-j+i}\binom{k-i}{j-i}$ to $\binom{k}{\gamma}$, which is shown in the first line of Eq. (6.20). The condition $0 \le j-i \le \gamma \le j$ is obtained by solving $i \ge \gamma - j + i \ge 0$ and $k - i \ge j - i \ge 0$ such that the expressions of $\binom{i}{\gamma-j+i}$ and $\binom{k-i}{j-i}$ are meaningful. For the special case $j = k-\theta+1$, $\rho_{i,j}$ is derived as follows.

$$\begin{aligned} &\rho_{i,k-\theta+1} \\ &= \Pr\left\{X_{n+1} = k-\theta+1 \middle| X_n = i\right\} \\ &= \Pr\Big\{\text{At least } (k-\theta+1-i) \text{ out of } (k-i) \textit{ uncovered} \text{ data packets are covered} \\ &\qquad\qquad \text{by the } (n+1)\text{-th check packet} \Big| X_n = i\Big\} \\ &= \sum_{v=k-\theta+1}^{\min\{i+\gamma,k\}} \Pr\Big\{(v-i) \text{ out of } (k-i) \textit{ uncovered} \text{ data packets are covered by} \\ &\qquad\qquad \text{the } (n+1)\text{-th check packet} \Big| X_n = i\Big\} \\ &= \sum_{v=k-\theta+1}^{\min\{i+\gamma,k\}} \frac{\binom{i}{\gamma-v+i}\binom{k-i}{v-i}}{\binom{k}{\gamma}}. \end{aligned} \tag{P.4}$$

It is clear that when the conditions of the first two cases in Eq. (6.20) are not satisfied, the covering state cannot transfer from state i to state j with only one new check packet, and thus we get $\rho_{i,j} = 0$.

Then, the probability-transition matrix, expressed by a $(k-\theta+2) \times (k-\theta+2)$

square matrix $\boldsymbol{\rho}$, is determined by

$$\boldsymbol{\rho} = \begin{bmatrix} \rho_{0,0} & \rho_{0,1} & \cdots & \rho_{0,k-\theta+1} \\ 0 & \rho_{1,1} & \cdots & \rho_{1,k-\theta+1} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}. \tag{P.5}$$

Note that $\boldsymbol{\rho}$ is an upper triangular matrix because X_n is an increasing sequence in terms of n.

We define a set of variables h_i, $0 \leq i \leq k-\theta+1$: if the current covering status is i (equivalently, we have covered i data packets), on average, the sender needs other h_i check packets to reach the target covering state $(k-\theta+1)$ (equivalently, we have covered at least $(k-\theta+1)$ data packets). Then, h_i for $0 \leq i \leq k-\theta+1$ is expressed as

$$h_i \triangleq E\left\{ j \left| X_n = i,\, X_{n+j} = k-\theta+1, j \geq 0, n \geq 0 \right. \right\}. \tag{P.6}$$

Clearly, we have $h_{k-\theta+1} = 0$ and $T(k,\theta,\gamma) = h_0$. If $\gamma \geq k-\theta+1$, it is clear that we need only one check packet to satisfy the covering criterion. Thus, we obtain Eq. (6.18).

We define $\mathbf{h} = (h_0, h_1, \ldots, h_{k-\theta+1})^\tau$, where $(\cdot)^\tau$ denotes the matrix transpose operator. $\mathbf{h}$ is the solution to the following linear equations [102]

$$\begin{cases} \mathbf{h} = \mathbf{z} + \boldsymbol{\rho}\mathbf{h}; \\ h_{k-\theta+1} = 0, \end{cases} \tag{P.7}$$

where $\mathbf{z}$ is a $(k-\theta+2)$-dimension column vector $(1, 1, \ldots, 1, 0)^\tau$.

As shown in Eq. (P.5), $\boldsymbol{\rho}$ is an upper triangular matrix. Hence, we can get the

solution to $\mathbf{h}$ by the following iterative equations

$$\begin{cases} h_i = \dfrac{1}{1-\rho_{i,i}} \left(1 + \displaystyle\sum_{j=i+1}^{k-\theta+1} \rho_{i,j} h_j \right), \quad i = 0, 1, 2, \ldots, k-\theta; \\ h_{k-\theta+1} = 0, \end{cases}$$

which complete the proof of Eq. (6.19), and thus Theorem 10 follows. □

APPENDIX Q

PROOF OF THEOREM 13

Proof. We construct the Lagrangian function of **VIII-*A*1′**, denoted by $\mathcal{J}_{A1}(\mathcal{L},\lambda)$, follows:

$$\mathcal{J}_{A1}(\mathcal{L},\lambda)=\mathbb{E}_{\mathbf{H}}\{J_{A1}(\mathcal{L},\lambda)\}$$

with

$$J_{A1}(\mathcal{L},\lambda)=\mathcal{L}+\lambda\left(e^{-\theta_1\widetilde{R}(\mathcal{L})}-e^{-\theta_1\overline{C}_1}\right), \tag{Q.1}$$

where λ is the Lagrangian multiplier associated with the constraint of **VIII-*A*1′**. Then, the optimal $\mathcal{L}^*$ and the optimal Lagrangian multiplier λ^* are solutions to the following equations [79]:

$$\begin{cases} 0 \in \partial_{\mathcal{L}} J_{A1}(\mathcal{L},\lambda), \quad \forall\,\mathbf{H}; \\ 0 = \mathbb{E}_{\mathbf{H}}\left\{e^{-\theta_1\widetilde{R}(\mathcal{L})}-e^{-\theta_1\overline{C}_1}\right\}. \end{cases} \tag{Q.2}$$

where $\partial_{\mathcal{L}} J_{A1}(\mathcal{L},\lambda)$ denotes the *subdifferential* [79] of the function $J_{A1}(\mathcal{L},\lambda)$ with respect to $\mathcal{L}$. Note that the subdifferential is defined for nondifferentiable convex functions (e.g., piece-wise linear functions), which is the counterpart concept for the gradient of differentiable convex functions. Based on [79], for a convex function $f(\mathbf{b})$ defined on $\mathbf{b}\in\mathcal{B}\subset\mathbb{R}$, where $\mathbb{R}$ is real-number set and $\mathcal{B}$ is a convex set, an $n\times 1$ real-valued vector $\boldsymbol{\varpi}$ is a *subgradient* of $h(\mathbf{b})$ if $h(\mathbf{b}')\geq h(\mathbf{b})+\boldsymbol{\varpi}^{\tau}(\mathbf{b}'-\mathbf{b})$ for all $\mathbf{b}'\in\mathcal{B}$. Accordingly, the collection of subgradients at $\mathbf{b}$ is the *subdifferential* of $h(\mathbf{b})$. For more details and properties of subdifferential, please refer to Definition 9.

Applying the piece-wise linear property and the concavity of $\widetilde{R}(\mathcal{L})$, we derive

$$\partial_{\mathcal{L}} J_{A1}(\mathcal{L},\lambda) = \begin{cases} \left[1-\theta_1\lambda\nu_j e^{-\theta_1\widetilde{R}(m_j)},\, 1-\theta_1\lambda\nu_{j+1}e^{-\theta_1\widetilde{R}(m_j)}\right), & \text{if } \mathcal{L}=m_j, \\ & j=0,1,\ldots,\mathcal{K}; \\ \left\{1-\theta_1\lambda\nu_j e^{-\theta_1\widetilde{R}(\mathcal{L})}\right\}, & \text{if } \mathcal{L}\in(m_{j-1},m_j), \\ & j=1,\ldots,\mathcal{K}. \end{cases} \tag{Q.3}$$

Plugging Eq. (Q.3) into Eq. (Q.2) and solving for the optimal solution, we get Eq. (8.21). Also, the equality of Eq. (8.14) needs to hold as required by Eq. (Q.2). which completes the proof of Theorem 13. □

APPENDIX R

PROOF OF THEOREM 14

Proof. It is clear that both the objective and constraint functions of **VIII-*A2*** are linear to $\boldsymbol{\phi}$. Thus, **VIII-*A2*** is a convex optimization problem. We construct the Lagrangian function of the **VIII-*A2***, denoted by $\mathcal{J}_{A2}(\boldsymbol{\phi}, \lambda)$, as

$$\mathcal{J}_{A2}(\boldsymbol{\phi}, \lambda) = \mathbb{E}_{\mathbf{H}} \left\{ \sum_{L=0}^{K_{\rm bs}} \phi_L L \right\} + \lambda \mathbb{E}_{\mathbf{H}} \left\{ \left(\sum_{L=0}^{K_{\rm bs}} \phi_L e^{-\theta_1 R(\Omega_L)} \right) - e^{-\theta_1 \overline{C}_1} \right\}$$

for $\sum_{L=0}^{K_{\rm bs}} \phi_L = 1$, where $\lambda \geq 0$ is the Lagrangian multiplier associated with Eq. (8.23). Since **VIII-*A2*** is a convex optimization problem, the optimal solution and Lagrangian multiplier of $(\boldsymbol{\phi}^*, \lambda^*)$ satisfies

$$\begin{cases} \boldsymbol{\phi}^* &= \arg \min\limits_{\boldsymbol{\phi}: \sum_{L=0}^{K_{\rm bs}} \phi_L = 1} \left\{ \mathcal{J}_{A2}(\boldsymbol{\phi}, \lambda^*) \right\}; \\ 0 &= \mathbb{E}_{\mathbf{H}} \left\{ \left(\sum_{L=0}^{K_{\rm bs}} \phi_L^* e^{-\theta_1 R(\Omega_L)} \right) - e^{-\theta_1 \overline{C}_1} \right\}. \end{cases} \tag{R.1}$$

Following the above equations, we further derive

$$\begin{aligned} \boldsymbol{\phi}^* &= \arg \min_{\boldsymbol{\phi}: \sum_{L=0}^{K_{\rm bs}} \phi_L = 1} \left\{ \mathbb{E}_{\mathbf{H}} \left\{ \sum_{L=0}^{K_{\rm bs}} \phi_L L + \lambda \left(\sum_{L=0}^{K_{\rm bs}} \phi_L e^{-\theta_1 R(\Omega_L)} \right) - e^{-\theta_1 \overline{C}_1} \right\} \right\}; \\ &= \arg \min_{\boldsymbol{\phi}: \sum_{L=0}^{K_{\rm bs}} \phi_L = 1} \left\{ \mathbb{E}_{\mathbf{H}} \left\{ \sum_{L=0}^{K_{\rm bs}} \left(L + \lambda e^{-\theta_1 R(\Omega_L)} \right) \phi_L \right\} \right\}. \end{aligned} \tag{R.2}$$

Eq. (R.2) suggests that the objective function is simply a linearly-weighted summation over $\{\phi_i\}_{i=1}^{K_{\rm bs}}$. Then, to minimize the objective function, ϕ_L associated with the minimum weight needs to be maximized. Following this principle, we define $L^* = \arg\min_L \left\{ L + \lambda^* e^{-\theta_1 R(\Omega_L)} \right\}$ as given in Eq. (8.25) and thus obtain Eq. (R.2) in Theorem 14. Furthermore, Eq. (R.1) implies that the equality of Eq. (8.23) holds, and thus Theorem 14 follows. □

VITA

Qinghe Du received his B.S. and his M.S. from Xi'an Jiaotong University, and graduated in 2010 with his Ph.D. degree under the supervision of Prof. Xi Zhang in the Networking and Information Systems Laboratory, Department of Electrical and Computer Engineering at Texas A&M University. His research interests include mobile wireless communications and networks with emphasis on mobile multicast, statistical QoS provisioning, QoS-driven resource allocations, cognitive radio techniques, and cross-layer design over wireless networks. His work, co-authored with his Ph.D. advisor Prof. Xi Zhang, received the Best Paper Award in the IEEE GLOBECOM 2007 for the paper Cross-Layer Design Based Rate Control for Mobile Multicast in Cellular Networks. He has published multiple papers in IEEE Transactions, IEEE Journal on Selected Areas in Communications, IEEE Communications Magazine, IEEE INFOCOM, IEEE GLOBECOM, IEEE ICC, etc. He may be reached at Nancun 4-4-701, Northwestern Polytechnical University, 127 West Youyi Road, Xi'an, Shaanxi, 710072, P. R. China. His e-mail address is duqinghe@tamu.edu.

The typist for this dissertation was Qinghe Du.