Title

Cross-layer optimization for transmission of delay- sensitive and bursty traffic in wireless systems

Permalink

https://escholarship.org/uc/item/2824r937

Author

Kittipiyakul, Somsak

Publication Date

2008

UNIVERSITY OF CALIFORNIA, SAN DIEGO

Cross-Layer Optimization for Transmission of Delay-Sensitive and Bursty Traffic in Wireless Systems

A dissertation submitted in partial satisfaction of the

requirements for the degree

Doctor of Philosophy

in

Electrical Engineering (Communication Theory and Systems)

by

Somsak Kittipiyakul

Committee in charge:

Professor Tara Javidi, Chair
Professor Rene Cruz
Professor Massimo Franceschetti
Professor Sonia Martínez
Professor Ruth Williams

2008

The dissertation of Somsak Kittipiyakul is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

____

____

____

____

____

Chair

University of California, San Diego

2008

To my beloved family

TABLE OF CONTENTS

Signature Page iii

Dedication iv

Table of Contents v

List of Figures ix

Acknowledgments x

Vita, Publications, and Fields of Study xiii

Abstract of the Dissertation xv

Chapter 1 Introduction 1

Chapter 2 Delay-Optimal Server Allocation in Multi-Queue Multi-Server Systems With Time-Varying Connectivities 12

2.1 Problem Formulation and Assumptions 16

2.1.1 Model and Notations 16

2.1.2 Problem Formulation 19

2.2 Instantaneous Throughput Maximizing vs. Load-Balancing Policies 20

2.2.1 Feasible and Non-Idling Allocations 21

2.2.2 Instantaneous Maximum Throughput Policies (MT) 22

2.2.3 Load-Balancing (LB) Policies 22

2.2.4 Example 23

2.3 Special Case: On-Off Channel 24

2.3.1 MTLB Policy 24

2.3.2 Existence of MTLB Policy 25

2.3.3 Construction of MTLB Policy 25

2.4 Optimality of MTLB Policy for Two Users 29

2.4.1 Class of Cost Functions 30

2.4.2 Dynamic Programming Formulation 31

2.4.3 Proof of the Optimality 33

2.5 Optimality of MTLB Policy under Fluid Relaxation 37

2.6 Application of MTLB Policy in OFDMA Wireless Systems 39

2.6.1 Heuristic Policies and Value of QSI versus CSI 40
2.6.2 Numerical Comparisons and Simulation 41
2.7 Summary 44

Chapter 3 High-SNR Analysis of Outage-Limited Communications with Bursty and Delay-Limited Information 46
3.0.1 Prior Work and Our Contribution 49
3.0.2 Overview of the Results 53
3.0.3 Outline of the Chapter 54
3.0.4 Notations 55
3.1 System Model 56
3.1.1 Channel and Coding Model 57
3.1.2 Smoothly-Scaling Bit-Arrival Process 59
3.1.3 Performance Measure and System Objective 65
3.2 Asymptotic Analysis of Probability of Delay Violation 66
3.3 Main Result: Optimal Asymptotic Total Probability of Error 68
3.3.1 Approximation of the Optimal Negative SNR Exponent . . 71
3.4 Applications of the Result 71
3.4.1 SISO Rayleigh Fast-Fading Channel 72
3.4.2 Cooperative Wireless Networking with Optimal Clustering . 75
3.4.3 MIMO Quasi-Static Communications 78
3.5 Summary 80

Chapter 4 Optimal Operating Point for MIMO Multiple Access Channel With Bursty Traffic 82
4.1 System Model 85
4.1.1 Source Model 87
4.1.2 MIMO-MAC Channel Model and PHY Model 89
4.1.3 Rate Scheduler 92
4.1.4 Arrival Rate Scaling and Stability Condition 94
4.1.5 Objective 95
4.2 Problem Analysis 97
4.2.1 Asymptotic $P_{\text{ch}}(d)$ 97
4.2.2 Asymptotic $P_{\text{delay}}(d)$ 98
4.2.3 Minimizing Asymptotic Total Error Probability 103
4.2.4 Statistical-Multiplexing and Optimal Diversity Gain 105
4.2.5 Resource Pooling and Statistical-Multiplexing 106

4.3 Example: Compound Poisson Sources and $K = 2$ 108
4.4 Summary 114

Chapter 5 Many-Sources Large Deviations for Max-Weight Scheduling . . 115
5.1 Problem Formulation 118
5.1.1 Main Results 121
5.2 Background and Assumptions 122
5.2.1 Topology for Sample Paths 122
5.2.2 Large Deviations Principle 123
5.2.3 Sample Path LDP of Arrival Processes 124
5.2.4 Garcia's Extended Contraction Principle 125
5.3 Analysis: LDP's for Workloads 127
5.3.1 LDP for Finite-Horizon Workloads 127
5.3.2 LDP for Infinite-Horizon Workloads 128
5.4 I.I.D. Increments: Simplified Rate Functions 129
5.4.1 Infinite-Horizon Rate Function 130
5.4.2 Finite-Horizon Rate Function 131
5.4.3 Properties of the Minimum-Cost Sample Paths 133
5.4.4 Example: Calculation of I_2 134
5.4.5 Bounds on I_t 138
5.4.6 Comparison of the Bounds: Numerical Examples 140
5.5 Summary 143

Chapter 6 Conclusions 144
6.1 Future Work 145
6.1.1 Future Work for Chapter 2 145
6.1.2 Future Work for Chapter 3 147
6.1.3 Future Work for Chapter 4 148
6.1.4 Future Work for Chapter 5 149

Appendix A Appendix for Chapter 2 151
A.1 Existence Proof of MTLB: Proof of Theorem 2.1 151
A.1.1 Alternating, Balancing, and Throughput-Increasing Paths . 152
A.1.2 Proof of Existence of MTLB Policy 154
A.1.3 Necessary and Sufficient Condition for MTLB Policy 156
A.1.4 Proof of Proposition 2.9 159
A.2 Supporting Lemmas for Theorem 2.2 160
A.3 Supporting Lemmas and Proof of Theorem 2.3 171

Appendix B Appendix for Chapter 3 178
B.1 Proof of Proposition 3.6 178
B.2 Proof of Lemma 3.10 179
B.3 Proof of Theorem 3.1 184
B.4 Proof of Lemma B.1 186
B.5 Proof of Lemma B.2 190
B.6 Proof of Lemma B.3 193

Appendix C Appendix for Chapter 4 198
C.1 Additional Assumption on Source Model 198
C.2 Proofs of Lemmas and Proposition 199
C.2.1 Proof of Lemma 4.9 199
C.2.2 Proof of Lemma 4.12 200
C.2.3 Proof of Theorem 4.1 204
C.2.4 Proof of Proposition 4.18 205

Appendix D Appendix for Chapter 5 206
D.1 Proof of Lemma 5.12 206
D.2 Proof of Lemma 5.13 211

Bibliography 214

LIST OF FIGURES

Figure 2.1 Multi-queue-multi-server allocation problem 17

Figure 2.2 Example of MTLB construction 27

Figure 2.3 Average queue backlog for bounded pareto distribution. 43

Figure 2.4 Average queue backlog for the Poisson distribution. 44

Figure 3.1 System model for point-to-point communication 48

Figure 3.2 Optimal negative SNR exponent for SISO, Rayleigh fast-fading channel 74

Figure 3.3 Comparisons of the exact values d^*, T^*, and r^* and their integer-relaxed approximations 76

Figure 3.3 Comparisons of the exact values, continued. 77

Figure 3.4 Snapshot of a cooperative relay wireless network 79

Figure 3.5 MIMO, quasi-static, coherent 2x2 channel. d^* v.s. λ for different values of D and μ. 80

Figure 4.1 System model and the two causes of bit loss: delay violation and channel decoding error. 86

Figure 4.2 Example of the multiplexing gain region $\mathcal{R}(d)$ 92

Figure 4.3 The MIMO-MAC multiplexing gain region $\mathcal{R}(d)$ and the multiplexing gain region of the fictitious system $\mathcal{R}_{\text{fic}}(d)$ 102

Figure 4.4 Queuing models of the upper and lower bounds of $\Pr(D^1 > D)$ for the case of $K = 2$ users and the antenna-pooling regime . . 102

Figure 4.5 The relation between (λ, D) to the system performance. 107

Figure 4.6 Plots of the exponents of $P_{\text{delay}}(d)$ and $P_{\text{ch}}(d)$ for two different average arrival rates ($\lambda = 0.5$ and 1) and delay bound $D = 20$ 111

Figure 4.7 Similar plots as in Figure 4.6 but with delay bound $D = 150$. 112

Figure 4.8 3D and contour plots characterizing the statistical-multiplexing gain 113

Figure 5.1 Two properties of the minimum-cost sample paths 135

Figure 5.2 Example of accumulated arrival and workload paths for I_2 . . 137

Figure 5.3 Example of the rate function $I_{10}(\mathbf{b})$ and its upper and lower bounds, and their corresponding optimizing t^* and optimal trajectories, when $\mathbf{b} = (b^1, b^2 = 1)$. 141

Figure 5.3 Example of the rate functions, continued. 142

Figure A.1 Example of an alternating path and the alternating allocation 155

ACKNOWLEDGMENTS

I would like to thank my advisor, Professor Tara Javidi, for her advice, encouragement, and support throughout my graduate studies. None of my dissertation research would have been possible without her wisdom and guidance. I also would like to thank my committee members, Professor Rene Cruz, Professor Massimo Franceschetti, Professor Ruth Williams, and Professor Sonia Martínez, for their valuable comments, suggestions, and time.

In addition, I would like to thank Professor Richard Ladner at University of Washington, Professor Tami Tamir at the Interdisciplinary Center, Israel, and Professor Mingyan Liu at Univeristy of Michigan, for their helpful suggestions. Chapter 3 was a result of joint work with Professor Petros Elia at EURECOM, France, whose insightful and careful way of looking at problems I really appreciated. I would like to thank Dr. Vijay Subramanian at Hamilton Institute, Ireland, for his guidance and suggestions in the last chapter of my work. I would also like to thank my lab mates for their help in listening to and improving my work.

My thanks also go to those friends in San Diego and Seattle for their friendship and encouragement, in particular to Nut, Ae, Am, Aui, Thawee, P'Ake, P'Korn, Dew, Roong, Muay, Ple, Nui, Noong, P'Mee, Num, Rathinakumar Appuswamy, Sumit Bhardwaj, Eric Wong, Chia-Wei Chang, and Jaewook Shim. Special thanks to my old friends, in particular to Jimmy, Jocelyn, Meaw, Mod, Namo, and Tanong, for their enduring friendship.

I would like to take this opportunity to express my gratitude to all the

people who have shaped and supported me throughout my life: my teachers at Wat Phai Ngoen Chotanaram School and Wat Suthiwararam School, in particular to Ajahn Nuwee Choovisitkul, Ajahn Khwanta Niemsa-ing, Ajahn Somsri Seaksard, Ajahn Boonchoo Suthinopparattanakul, and my late Ajahn Jittima; Ah Pae and Jeh Yu for the work opportunity that helped support my family and my study; Doctor Narong Roongwithu for his support during my high school study; my friends and teachers at Deerfield Academy and MIT, especially my Master thesis advisor Professor Donald E. Troxel; my host family, Uncle Ted Komosa; and my colleagues at Shin Satellite.

Most importantly, I thank my family, which has always been my invaluable treasure and inspiration. No amount of words can express my gratitude. I thank my loving wife for her unwavering love, understanding, friendship, encouragement, and care of our beloved son; my older brother for the role model in my study; my younger brothers and sisters and my wife's family for their support; and to my parents who, with incomparable patience and understanding, always love us no matter what. I attribute everything I may have achieved to my family.

This dissertation was supported in part by the Center for Wireless Communications, UCSD, UC Discovery Grant No. Com04-10176, ARO-MURI Grant No. W911NF-04-1-0224, NSF CAREER Award No. CNS-0533035, and AFOSR Grant No. FA9550-05-01-0430.

Chapter 2, in part, is under review for publication in IEEE Transactions on Information Theory. It also appears, in part, in IEEE Communications Letters.

Chapter 3, in full, is accepted for publication in IEEE Transactions on Information Theory. Chapter 4, in full, appears in IEEE Transactions on Wireless Communications. Chapter 5, in full, will appear in the Proceedings of the 46th Annual Allerton Conference on Communication, Control, and Computing, 2008. The dissertation author was the primary investigator and author of the aforementioned papers.

VITA

1996	S.B. and M.Eng., Electrical Engineering and Computer Science, Massachusetts Institute of Technology
1996–2003	Systems Engineer, Shin Satellite Plc., Nonthaburi, Thailand
2001	B.P.A., Public Administration, Sukhothai Thammathirat Open University, Nonthaburi, Thailand
2003–2004	Graduate Research Assistant, University of Washington, Seattle
2005–2008	Graduate Student Researcher, University of California, San Diego
2008	Ph.D., Electrical Engineering, University of California San Diego

PUBLICATIONS

S. Kittipiyakul, T. Javidi, and V. G. Subramanian, "Many-sources large deviations for max-weight scheduling," to appear in *46th Annual Allerton Conference on Communication, Control, and Computing (Allerton'08).*

S. Kittipiyakul, P. Elia, and T. Javidi, "High-SNR analysis of outage-limited communications with bursty and delay-limited information," is accepted for publication in *IEEE Trans. Inf. Theory.*

S. Kittipiyakul and T. Javidi, "Relay scheduling and cooperative diversity for delay-sensitive and bursty traffic," in *45th Annual Allerton Conference on Communication, Control, and Computing*, Monticello, Illinois, USA, Sep. 2007.

P. Elia, S. Kittipiyakul, and T. Javidi, "Cooperative diversity in wireless networks with stochastic and bursty traffic," in *IEEE Int. Symp. Information Theory*, Nice, France, Jun. 2007.

P. Elia, S. Kittipiyakul, and T. Javidi, "On the Responsiveness-Diversity-Multiplexing tradeoff," in *5th Intl. Symp. on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks*, Limassol, Cyprus, Apr. 2007.

S. Kittipiyakul and T. Javidi, "Optimal operating point for MIMO multiple access channel with bursty traffic," *IEEE Trans. Wireless Commun.*, vol. 6, no. 12, pp. 4464–4474, Dec. 2007.

S. Kittipiyakul and T. Javidi, “Optimal operating point in MIMO channel for delay-sensitive and bursty traffic,” in *IEEE Int. Symp. Information Theory*, Seattle, Washington, USA, Jul. 2006.

S. Kittipiyakul and T. Javidi, “Delay-optimal server allocation in multi-queue multi-server systems with time-varying connectivities,” submitted to *IEEE Trans. Inf. Theory.*

S. Kittipiyakul and T. Javidi, “Resource allocation in OFDMA with time-varying channel and bursty arrivals,” *IEEE Commun. Lett.*, vol. 11, no. 9, pp. 708–710, Sept. 2007.

S. Kittipiyakul and T. Javidi, “Subcarrier allocation in OFDMA systems: beyond water-filling,” in *2004 Asilomar Conference on Signals, Systems, and Computers*, Nov. 2004.

S. Kittipiyakul and T. Javidi, “A fresh look at optimal subcarrier allocation in OFDMA systems,” in *IEEE Conference on Decision and Control (CDC 2004)*, Dec. 2004.

S. Kittipiyakul, “Automated remote microscope for inspection of integrated circuits,” M.Eng. Thesis, MIT, 1996.

FIELDS OF STUDY

Major Field: Electrical Engineering (Communication Theory and Systems)

Studies in Wireless Communications and Networks
Professor Tara Javidi

ABSTRACT OF THE DISSERTATION

Cross-Layer Optimization for Transmission of Delay-Sensitive and Bursty Traffic in Wireless Systems

by

Somsak Kittipiyakul

Doctor of Philosophy in Electrical Engineering

(Communication Theory and Systems)

University of California, San Diego, 2008

Professor Tara Javidi, Chair

High demands on the quality of service (QoS), in terms of throughput, delay, and packet loss, in wireless systems have fueled substantial research interest in jointly considering different layers of protocol stack. Integrated design of wireless systems from physical to application layers is challenging due to high variability of wireless channels and stochastic and delay-sensitive nature of traffic. Following this research interest, the current dissertation considers resource allocation in wireless systems for transmission of delay-sensitive and bursty traffic in two main areas.

In the first chapter of this dissertation, a subcarrier allocation problem in OFDMA downlink system is studied, where given the knowledge of the channel and queue states, the optimal centralized allocation policy seeks to minimize the average packet delay. The problem is modeled as a multi-queue multi-server assignment problem with time-varying connectivities. For on-off connectivities and homogeneous users, we show, using a dynamic programming approach, that a simultaneous maximum-throughput and load-balancing policy is delay-optimal. For more general connectivities, we propose heuristic policies that use different degrees of queue and channel state information to provide good delay performance for various traffic loads.

The rest of the dissertation is concerned with cross-layer design of wireless data networks, when there is no channel state information at the transmitter and no retransmission. In Chapter 3, we study how to set up various physical layer parameters, e.g., coding block length and channel transmission rate of point-to-point wireless fading channels, such that the total probability of bit loss is minimized, where bit losses account for both erroneous decoding at the receiver as well as violation of a specified delay constraint. We simplify the problem by considering an asymptotic high signal-to-noise-ratio (SNR) regime and assuming smoothly scaling (with SNR) bit-arrival processes.

Extending this study to multi-user settings, in Chapter 4 we study how to select the optimal channel spatial-diversity in MIMO multiple access channels in order to minimize the asymptotic high-SNR error probability. We also quantify

the amount of the performance improvement that can be achieved from using an optimal queue-aware dynamic rate scheduler. While Chapter 4 answers the above question for sufficiently large delay constraints, Chapter 5 considers a case of finite and small buffer constraints. Finally, in Chapter 5, we propose a large-deviations analysis of the asymptotic buffer overflow probability for a maximum-weight dynamic scheduling policy with simplex rate region, assuming properly-scaled arrival processes.

Extending this study to multi-user settings, in Chapter 4 we study how to select the optimal channel spatial-diversity in MIMO multiple access channels in order to minimize the asymptotic high-SNR error probability. We also quantify the amount of the performance improvement that can be achieved from using an optimal queue-aware dynamic rate scheduler. While Chapter 4 answers the above question for sufficiently large delay constraints, Chapter 5 considers a case of finite and small buffer constraints. Finally, in Chapter 5, we propose a large-deviations analysis of the asymptotic buffer overflow probability for a maximum-weight dynamic scheduling policy with simplex rate region, assuming properly-scaled arrival processes.

Chapter 1

Introduction

In the context of wireless data networks, high demands on the quality of service (QoS), in terms of throughput, delay, and packet loss, have attracted substantial research interest in jointly considering physical layer and higher networking layer issues in an integrated framework. It is well known that this *cross-layer* resource allocation approach, compared to a traditional layered architecture approach, can provide a significant performance gain for wireless fading channels (e.g., see [7, 33, 54]). The motivation for the cross-layer approach lies in the fact that wireless channel is an inherently multi-access channel where the channel capacity is time-varying and susceptible to interference among users. In this way, the knowledge of the channel variability (e.g., in time, frequency, and space) can be exploited by the system to significantly improve performance. This improvement is a result of transmitting more information in good channel states and less in poor conditions. When there are multiple users, the channel quality varies across the users and hence, the system can selectively transmit to the users with good channel conditions.

On the other hand, the knowledge of the current channel state may be unavailable at the transmitters in which case the system cannot dynamically allocate resource according to channel variability. In this case, the channel variability can still be used to improve the reliability of reception (diversity gain) or increase the rate of communication for a fixed reliability level (multiplexing gain) via innovative (space-time) coding schemes. For example, in slow-fading multiple transmit and receive antenna (MIMO) systems, the spatial variability of the channels across the transmit and receive antennas can be exploited to simultaneously provide diversity and multiplexing gains, whose relationship follows a fundamental tradeoff known as *diversity-multiplexing tradeoff* (DMT), first introduced by Zheng and Tse [90]. A fundamental cross-layer question of interest is how to select the optimal operating diversity and multiplexing gains, when the network layer performance such as delay violation probability, as well as the physical layer performance such as channel error performance, are jointly considered.

In this dissertation we study resource allocation problems under both assumptions regarding channel state knowledge. In the first scenario where the channel state is available at the transmitter, we examine the problem of delay-optimal subcarrier allocation in orthogonal frequency-division multiple access (OFDMA) downlink systems. On the other hand, in the second scenario where the channel state is unavailable, we examine the problem of optimal selection of the static operating PHY parameters such that the asymptotic total error performance is optimized. The specific contributions of this work can be characterized as follows.

Chapter 2: Delay-Optimal Server Allocation in Multi-Queue Multi-Server Systems With Time-Varying Connectivities

In the presence of frequency selectivity, code-division, or spatial degrees of freedom, many wireless systems with multiple *orthogonalized sub-channels* and multiple users can be viewed as multi-queue multi-server queuing systems which enable transmission of packets in a parallel manner. Examples of such systems include OFDMA systems, where the total available bandwidth is divided into multiple orthogonal narrow subcarriers to be shared by users [83]. Usually data packets in such systems arrive stochastically to each user and are stored in buffers prior to transmission. In this context, there is a limited number of subcarriers, not allowing simultaneous transmission of all queued packets. This gives rise to a scheduling problem involving the allocation of the orthogonal channels to the different data streams.

In addition, due to selective fading for instance, the "quality" of these channels as perceived by different users varies stochastically with time and users. This introduces the notion of stochastic server quality, also known as connectivity. In the presence of reliable estimates of channel quality at the transmitter, the stochastic variation across users provides opportunities for selectively scheduling the transmission among users to benefit from *multi-user diversity gain* [48].

The focus of Chapter 2 is the optimal delay performance of a multi-server queueing system with stochastic channel state (connectivity) and arrival processes. It is known that, in general, delay-optimal policies must trade off between two competing goals: the desire to get maximum throughput now (which is achieved by *opportunistic scheduling*) and the desire to get the maximum throughput in the future (which is achieved by balancing the remaining load). The second goal, under admissible traffic regimes, accounts for queue occupancies; the intuitive reasoning is that the queued packets should be spread over multiple queues so that the system will have a better chance of avoiding idling any subcarriers in the future. These goals, in general, can be incompatible over short timescales, implying an empty intersection between the two classes of policies.

In Chapter 2 we establish the existence and delay optimality of a policy achieving both goals, when the connectivities between each queue and each server are random but binary (either “on” or “off”). The existence of this optimal policy (*MTLB policy*), which simultaneously maximizes the instantaneous throughput and balances the queues at every timeslot, is shown constructively, while the proof of the optimality of the MTLB policy is based on properties of the dynamic programming value function.

For the systems whose connectivities are more general than the on-off model, we utilize the insights gained from the MTLB policy to propose heuristic policies that use different degrees of channel and queue state information. We illustrate how the significance of queue vs. channel state information varies with the traffic load. This important observation is not only used to devise algorithms with good average delay performance, but also of practical interest when one considers the overhead associated with channel estimation and feedback.

Chapter 3: High-SNR Analysis of Outage-Limited Communications with Bursty and Delay-Limited Information

Communication of delay-sensitive bits over wireless fading channels has been recently addressed under various assumptions and settings. An objective of interest is in the quality of service (QoS), in terms of both packet losses and packet delays in the system due to queueing and transmission delays. Often asymptotic approximations are employed to enable tractable analysis of the problems.

Following this research interest, a significant portion of this dissertation (i.e., Chapters 3 and 4) investigate communication problems of delay-sensitive and bursty data over wireless fading channels. In contrast to Chapter 2, the setting of interest is the case where there is no channel state information available at the transmitter and no feedback from the receiver (hence, no possibility for retransmissions of erroneous packets). Furthermore, in this segment of the dissertation, the notion of delay performance measure is the probability of bits violating a strict delay requirement (*hard* delay performance) as opposed to the average delay (*soft* delay performance) used in Chapter 2. Our performance objective integrates both packet losses and packet delays and defined as the total bit error probability where the bit errors are caused by erroneous decoding at the receiver or violation of the delay constraint.

In this chapter, we study how to set up the coding block length and channel transmission rate (or PHY parameters in general) for point-to-point wireless fading channels such that the total probability of bit loss is minimized. We assume that the values of the PHY parameters can be optimally selected prior to the start of the operation by using the given knowledge of the statistics of the channel and bit-arrival processes.

In general it is difficult to derive the exact relationship between the system parameters and the probabilities of channel decoding error and delay violation. Instead, we simplify the problem by studying an asymptotic approximation when the signal-to-noise ratio (SNR) is asymptotically high. This asymptotic approximation greatly benefits in establishing intuitive understanding of the relation between the system parameters and the performance objective.

The contributions of this chapter are as follow. First, we formulate and express the exponent of the asymptotic high-SNR total error probability for the above setting. Since it is meaningful to ask about the optimal coding block duration only when the optimal value is finite, we propose to study the bit-arrival processes that are smoothly scaling with SNR. This scaling assumption provides the asymptotic high-SNR approximation of the delay violation probability that is valid at finite and small delay constraints. We show that, only at the proper scaling of the source processes with SNR, we can find a non-trivial optimal coding block length and transmission rate that maximize the exponent of the asymptotic total error probability. This optimal exponent reveals a tradeoff that addresses the question of how much of the delay budget and channel capacity should be expended for gaining reliability over the channel and how much for accommodating the burstiness with delay constraints. Finally, we illustrate the applications of the results in different outage-limited communication settings. For examples, we use the results to find the optimal cluster size in cooperative wireless networking and to find the optimal multiplexing gain in quasi-static MIMO communications.

Chapter 4: Optimal Operating Point for MIMO Multiple Access Channel With Bursty Traffic

In this chapter, we extend our study of the optimal selection of PHY parameters in Chapter 3 to the multiple access control (MAC) layer. In particular, we study a multi-user setting of MIMO multiple access channel (MIMO-MAC) with a dynamic longest-delay-first (LDF) rate scheduler that dynamically allocates rates within some rate region to the users, in response to the delays of the head-of-the-line bits.

As discussed earlier, in the point-to-point MIMO channel, the relationship between the spatial diversity gain (channel reliability) and the spatial multiplexing gain (transmission rate) follows the Zheng-Tse diversity-multiplexing tradeoff (DMT) [90]. In a multiple-access situation, multiple receive antennas can also be used to spatially separate signals from different users (multiple-access gain). For a given spatial-diversity gain, the spatial-multiplexing rates can be anywhere within the corresponding rate region. The fundamental relationship between the spatial-diversity gain and the spatial-multiplexing rate region follows the Tse-Viswanath-Zheng DMT tradeoff [78], which shows that the shape of the spatial-multiplexing rate region depends on the spatial-diversity gain.

On the other hand, a dynamic scheduling can provide a statistical multiplexing gain to improve delay performance (e.g., see [9, 72]). From a scheduling perspective, statistical-multiplexing is a key mechanism by which the network resources are used to improve the delay performance for bursty users. In particular, statistical-multiplexing capitalizes on the fact that peaks in traffic of simultaneously ongoing traffic streams rarely coincide.

Combining the multiple-access and statistical-multiplexing gains, we provide bounds on the asymptotic high-SNR total error probability, when the delay bound requirement is sufficiently large. Using these expressions, we find the bounds on the optimal spatial-diversity gain and the corresponding rate region for the MIMO-MAC channel to operate such that the exponent of the asymptotic total error probability is maximized.

An important consequence of our results is that we can quantify the amount of the statistical multiplexing gain that can be achieved and observe the interplay between the optimal exponent of the asymptotic total error probability and the system parameters (traffic load and delay bound). For example, we show that, when the traffic load is sufficiently light, the optimal exponent is equivalent to the case when only one user is transmitting in the system. In this case there is no resource sharing and hence no statistical-multiplexing gain. On the other hand, at higher traffic loads, there is some statistical-multiplexing gain whose amount depends on the values of the traffic load and delay bound. The highest gain is achieved when the traffic load and delay bound are at medium values. In this case, the optimal performance allows each user to perceive the whole resource dedicated to itself.

Chapter 5: Many-Sources Large Deviations for Max-Weight Scheduling

This chapter was originally motivated by our interest in extending the study in Chapter 4 to include the study of optimal selection of the coding block length as in Chapter 3 but in a multi-user setting. Toward this goal, we need to have the asymptotic high-SNR approximation of the delay violation probability that is valid even for finite and small delay constraints in a multi-queue system with a Largest Delay First scheduler. To the best of our knowledge, all existing results in the literature for large-deviations performance analysis of the largest-delay-first (or longest-queue-first) scheduler are concerned about the asymptotic approximations when the delay or buffer constraints are asymptotically large (see [9, 72, 75]).

The only large-deviations analysis of multi-queue systems dealing with finite and small delay constraints in the literature provides a large deviations principle for a Generalized Processor Sharing (GPS) scheduler (e.g., see [68]) and for scheduling policies favoring short jobs (e.g., see [85]), all under a many-sources (or many-flows) asymptotic regime. Instead of assuming scaling of the delay bound (known as *large-buffer* scaling), these results assume a different scaling of the arrival processes, where the arrival process for each user is assumed to be an aggregation of multiple i.i.d. flows or sources (hence, the terms many-flows and many-sources). The many-sources large-deviations analysis in [68,85], for instance, studies the performance of dynamic schedulers with L i.i.d. flows to each queue and a server capacity of Lc (for some constant c) as L grows to infinity.[1]

Following this many-sources framework, in this chapter we are interested in a many-sources large-deviations analysis of the *buffer overflow* probability for the longest-queue-first scheduler (or equivalently, a maximum-weight scheduler when the weights are the queue lengths). In particular, we assume that there is one server of fixed capacity c which is allocated to serve the user with the longest queue when the arrival process to each queue is an average of L i.i.d. flows. The quantity of interest in our study is the buffer overflow probability.

The result on the buffer overflow probability could be used in the multi-user study similar to the single-user study in Chapter 3, in which the delay violation probability is replaced with a requirement on buffer overflows. In addition, we believe that the result in Chapter 5 is a useful and interesting contribution to the large-deviations literature by and of itself. We believe that our result can be taken as a first step in an extension of the large-deviations analysis of the maximum-weight scheduler in the many-sources framework. In fact, the many sources traffic scaling is of practical interest in real applications, where there is a large number of traffic flows passing through each network element. This traffic scaling usually gives a more refined approximation to the probabilistic quantities of interest by incorporating the impact of the statistical-multiplexing gain among

[1] This many-sources limit was first introduced by Weiss in [81] and has been extensively studied in queuing analysis, starting from the works in [11,15,70]. For an excellent introduction of many-sources and large-buffer scalings to queueing applications, see [29,82].

flows (for discussions on the two scaling regimes, see, e.g., [11, 47, 68]).

Chapter 2

Delay-Optimal Server Allocation in Multi-Queue Multi-Server Systems With Time-Varying Connectivities

Abstract

This chapter considers the problem of optimal server allocation in a time-slotted system with N statistically symmetric queues and K servers when the arrivals and channels are stochastic and time-varying. In this setting we identify two classes of "desirable" policies with potentially competing goals of maximizing instantaneous throughput versus balancing the load. Via an example, we show that these goals, in general, can be incompatible, implying an empty intersection

between the two classes of policies. On the other hand, we establish the existence of a policy achieving both goals when the connectivities between each queue and each server are random and either “on” or “off”. We use dynamic programming and properties of the value function to establish the delay-optimality of a policy which, at each time-slot, simultaneously maximizes the instantaneous throughput and balances the queues. For the systems whose connectivities can be general than the on-off model (such as OFDMA wireless systems), we utilize the insights learned from the MTLB policy to create heuristic policies that use different degrees of channel and queue state information. We illustrate how the significance of queue vs. channel state information varies with the traffic load. This is of extreme practical interest when one considers the overhead associated with channel estimation and feedback.

In the presence of frequency selectivity, code-division, or spatial degrees of freedom, many wireless systems with multiple “orthogonalized sub-channels” and multiple users can be viewed as multi-queue multi-server queuing systems which enable transmission of packets in a parallel manner. Examples of such systems include OFDMA systems where the total available bandwidth is divided into multiple orthogonal narrow subcarriers to be shared by users [83].

The focus of this chapter is the optimal delay performance of the system with stochastic channel state and arrival processes. In other words, we are interested in the delay performance of the system under stochastic admissible traffic.

In general, delay-optimal policies must trade off between two competing goals: the desire to get maximum throughput now (which is achieved by opportunistic scheduling) and the desire to get the maximum throughput in the future (which is achieved by balancing the remaining load). These goals, in general, can be incompatible over short timescales, implying an empty intersection between the two classes of policies.

In this chapter, we investigate the delay-optimality of a certain policy in a multi-queue multi-server system with random binary connectivities. Delay optimal policies have been studied in many queuing systems with stochastically varying connectivities under different settings [2,21,30,31,34,56,60,69,76,87]. In this chapter, we consider a statistically symmetric case of arrival and connectivity processes. The intuition behind a symmetric system is that relabeling the queues leads to a statistically identical system. In many instances, as stated in [31], "symmetry sometimes leads to rather simple optimal policies, although their optimality can be hard to establish." The most related models to our work are those introduced in [76] and [31], while our proof technique is closely related to those developed in [20, 21, 34, 40, 49]. In their seminal work, Tassiulas and Ephremides [76] studied a single-server N-queue assignment problem where the connectivity followed an on-off model, i.e., the connectivity was described by a binary vector of dimension N. They showed that a longest-connected-queue (LCQ) policy maximizes the stability region of the system (i.e., throughput-optimal) and is also average-delay optimal when the arrival processes and the channel processes are statistically identical among users, i.e., the users are symmetric. Ganti et al. [30, 31] generalized the problem to a symmetric K-server, N-queue allocation problem with binary connectivity vector of dimension N. However, in their multi-server generalization, no more than one server can be allocated to a queue.[1] The main contribution of our work is a generalization of the model and results in [31,76] to (1) a connectivity model of a K-by-N matrix form, (2) more general arrival processes, not restricted to only Bernoulli (and variation thereof as discussed in Section III of [31]), and (3) a generalization where any queue can be served simultaneously by multiple servers. These generalizations, though, have been established in two restricted cases: (i) with the constraint on the number of users to $N = 2$ and (ii) with the constraint of fluid server allocation.

The chapter is organized as follows. We first model the multi-channel allocation problem as a multi-queue multi-server allocation problem in Section 2.1. We account for stochastically varying channel states, via a general notion of connectivity. In Section 2.2, we discuss two classes of "desirable" policies: the instantaneous throughput maximizing policies and load-balancing policies. Via a simple example, we show that, in general, the intersection of these two classes of policies can be empty. This results in a complicated structure for the optimal policy in the general case. However, in Section 2.3, we show that when connectivity profile has a binary form (on-off connectivity), it is possible to construct a policy which simultaneously maximizes the instantaneous throughput and balances the load. In Section 2.4,

[1] This constraint is relaxed only when a relaxation of integral allocation is also allowed; this is sufficiently different from our scenario of interest.

we show the optimality of this maximum-throughput and load-balancing (MTLB) policy when there are $N = 2$ users, using dynamic programming (DP) arguments and the properties of the DP value function. In Section 2.5, we show that if fractional server allocations are allowed, then the fluid version of the MTLB policy is optimal for general N. For general connectivity model, Section 2.6 provides heuristic policies that use different degrees of channel and queue state information. We illustrate how the significance of queue vs. channel state information varies with the traffic load. Section 2.7 concludes the chapter.

2.1 Problem Formulation and Assumptions

2.1.1 Model and Notations

We consider a multi-queue multi-server system with stochastic connectivities as shown in Figure 2.1. There are N queues (users) and K servers (subcarriers). Let $U = \{1, \ldots, N\}$ be the set of all queues and $V = \{1, \ldots, K\}$ the set of all servers. Fixed-size packets arrive stochastically for each user and are transmitted over a set of allocated servers. Each user has an infinite buffer to store the data packets that cannot be immediately transmitted. The system is time-slotted. The users have the same priority and are symmetric, i.e., they have *statistically* identical arrival and channel connectivity processes. At the beginning of each timeslot, the assignment of servers to users is instantaneous and made by a centralized resource manager. The resource manager has perfect knowledge of the

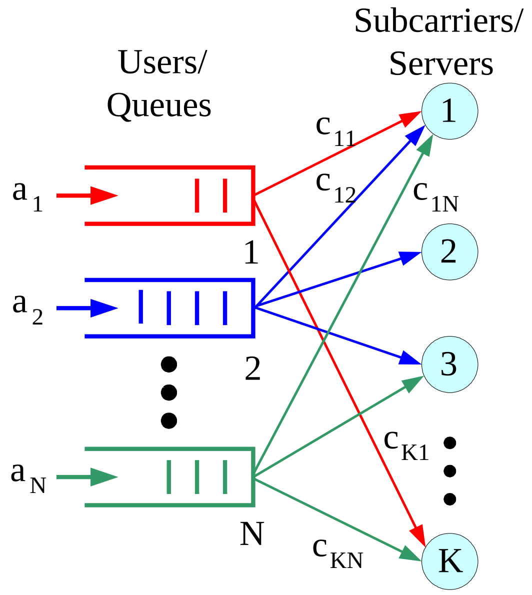


Figure 2.1 Multi-queue-multi-server allocation problem with time-varying connectivities.

current queue backlogs and the connectivities which are assumed constant during a timeslot but varying independently over timeslots (e.g. block fading model). We do not allow sharing of any servers and assume no error in the transmission.

The following notations are used throughout the chapter. Note that we use the following conventions: lower-case letters for scalar, bold-faced lower-case letters for row vectors, upper-case letters for matrices and scripted upper-case letters for space of matrices.

- $\mathbf{b}(t) = (b_1, \ldots, b_N)$: Backlogs (in units of packets) of each queue at the beginning of timeslot t.

- $\mathbf{a}(t) = (a_1, \ldots, a_N)$: Stochastic number of fixed-length packets arrived to each queue during timeslot t. The new packet arrivals at time t can be served only at time $t+1$ or after.

- $C(t) = [c_{ij}]$: the K-by-N stochastic *connectivity matrix* at timeslot t where $c_{ij} \in \{0, 1, \ldots, c_{\max} < \infty\}$ denotes the maximum number of packets subcarrier i can serve from queue j at time t.
- $W(t) = [w_{ij}]$: the K-by-N allocation matrix at the beginning of timeslot t where $w_{ij} \in \{0, 1\}$ and $w_{ij} = 1$ denotes that subcarrier i is assigned to serve queue j during time t.

The dynamics of the queue length vectors under an allocation $W(t)$ is described by the equation

$$\mathbf{b}(t+1) = [\mathbf{b}(t) - \mathbf{1}\big(W(t) \odot C(t)\big)]^{+} + \mathbf{a}(t), \quad t = 1, 2, \ldots \tag{2.1}$$

where an element-wise product $W(t) \odot C(t)$ is a matrix $[w_{ij}c_{ij}]$, $\mathbf{1}$ is a K-dimensional row vector of K ones, and, for a vector $\mathbf{v} \in \mathbb{R}^N$, $[\mathbf{v}]^+ = \left[v_1^+, \ldots, v_N^+\right]$ with $v_j^+ = \max\{0, v_j\}$. For the case of the on-off connectivities, where $c_{\max} = 1$, the above queue dynamics reduces to:

$$\mathbf{b}(t+1) = [\mathbf{b}(t) - \mathbf{1}W(t)]^{+} + \mathbf{a}(t), \quad t = 1, 2, \ldots. \tag{2.2}$$

Definition 2.1. For a row vector $\mathbf{x} = (x_1, \ldots, x_N)$ and a matrix $Y = [\mathbf{y}_1, \ldots, \mathbf{y}_N]$ where $\mathbf{y}_j$ is a column vector, a *column-by-column matrix permutation* Π_π corresponding to a permutation π is defined as, for any j and $k \in \{1, \ldots, N\}$,

$$\pi(x_j) = x_k \Leftrightarrow \Pi_\pi(\mathbf{y}_j) = \mathbf{y}_k$$

Using the above notations and definition, we make the following symmetric assumptions on the arrival and connectivity processes.

(**A1**) The packet arrival processes $[\mathbf{a}(t)]$ are i.i.d. across timeslots and *symmetric* or *exchangeable*, i.e., the joint probability mass function (pmf) is permutation invariant. That is,

$$P[\mathbf{a}(t) = \pi(\mathbf{x})] = P[\mathbf{a}(t) = \mathbf{x}] \tag{2.3}$$

for any t, vector $\mathbf{x}$, and permutation π.

(**A2**) The connectivity profiles $[\mathbf{C}(t)]$ are i.i.d. across timeslots and exchangeable across users, i.e., the joint pmf for $[\mathbf{C}(t)]$ is column-by-column permutation invariant. That is,

$$P[C(t) = \Pi_\pi(Y)] = P[C(t) = Y]$$

for any t, matrix Y, and column-by-column permutation matrix Π_π.

Assumption (**A2**) is valid when the channel and mobility create a homogeneous environment for all users. Note that (**A1**) and (**A2**) imply independence across time but not across users, i.e., at a given time the arrivals to various queues or connectivities among users or servers need not be independent.

2.1.2 Problem Formulation

Problem (P)

Consider the system described above, we wish to determine a Markov server allocation policy σ that minimizes the cost function at the finite horizon T:

$$J_T^\sigma = E[\Lambda_T^\sigma | \mathcal{I}_0] \tag{2.4}$$

where $\mathcal{I}_0$ summarizes all information available at time zero and Λ_T^σ denotes the cost under the Markov policy σ over horizon T:

$$\Lambda_T^\sigma = \sum_{t=0}^{T} \phi(\mathbf{b}(t)) \tag{2.5}$$

where the cost function $\phi(\mathbf{b}) = \sum_{j=1}^{N} \xi(b_j)$ and ξ is a convex and strictly increasing function.

We note that the restriction to Markov policies does not entail any loss of optimality because Problem (P) is a stochastic control problem with perfect observations [51]. Also, note that when ξ is an identity function, Problem (P) reduces to an average total backlog $\left(E[\sum_{t=0}^{T} \sum_{j=1}^{N} b_j(t)]\right)$ minimization problem over horizon T. From Little's Theorem [8], any optimal policy that achieves the minimum average backlog achieves the minimum average packet delay as well. Thus, our study includes the study of the average-delay minimization.

2.2 Instantaneous Throughput Maximizing vs. Load-Balancing Policies

In this section, we consider two classes of server allocation policies: a class comprising of instantaneous throughput maximizing (MT) policies and another class of load-balancing (LB) policies. As discussed in the introduction, each class represents one of the competing goals: an MT policy maximizes the number of packets being served now, while an LB policy maximizes the number of non-empty queues (hence, the multiuser diversity gain and the number of packets served) in the future. We demonstrate by an example that, in general, the intersection of the two classes of policies can be empty. In such cases, the optimal policies for Problem ($\mathbf{P}$) remains and can be, in general, a complicated mixture of policies carefully chosen at different time from one of the above two classes of policies. To be precise, we first define the feasible allocation and non-idling feasible allocation. Then, we describe the two classes of policies mentioned above.

2.2.1 Feasible and Non-Idling Allocations

Assume that at the beginning of time slot n, the state of the system is $(\mathbf{b}, C)$. An allocation $W = [w_{ij}]_{K\times N}$ is a *feasible allocation* for time slot n if

(a) $w_{ij} \in \{0, 1\}$;

(b) $c_{ij} = 0 \Rightarrow w_{ij} = 0$; and

(c) $\sum_{j=1}^{N} w_{ij} \leq 1, \forall\ i = 1, \ldots, K$.

The set of all feasible allocations is denoted by $\mathcal{W}(C)$. In addition, define $\mathcal{W}(\mathbf{b}, C) \subseteq \mathcal{W}(C)$ to denote the set of all *non-idling* feasible allocation W if W also satisfies

(d) $\sum_{i=1}^{K} w_{ij}c_{ij} \leq b_j, \forall\ j = 1, \ldots, N$.

2.2.2 Instantaneous Maximum Throughput Policies (MT)

An MT allocation $W^{\mathrm{MT}} = [w_{ij}^{\mathrm{MT}}] \in \mathcal{W}(\mathbf{b}, C)$ is a non-idling allocation that achieves the maximum throughput at time t if for all $W = [w_{ij}] \in \mathcal{W}(\mathbf{b}, C)$,

$$\sum_{j=1}^{N} \sum_{i=1}^{K} w_{ij}^{\mathrm{MT}} c_{ij} \geq \sum_{j=1}^{N} \sum_{i=1}^{K} w_{ij} c_{ij}. \tag{2.6}$$

2.2.3 Load-Balancing (LB) Policies

It is reasonable to maximize the expected number of packets served in future under stochastic arrival and connectivity processes. For that reason, we consider a load-balancing policy which distributes the *future workload* among the queues as evenly as possible as to minimize the expected future server idling. This roughly ensures a larger set of allocation policies in the future. The future workload is defined as the queue length vector after assignment, i.e., the leftover queue vector. The Longest Connected Queue policy [76] and Most Balanced policy [30] are some examples of the LB policies. Note that the LB policies potentially sacrifice the current throughput (by giving priority to long queues) for the future throughput.

To introduce the LB policy, we need the following definition to compare queue vectors in term of their load distribution:

Definition 2.2. We say $\mathbf{x} \leq_{\mathrm{LQO}} \mathbf{y}$ ($\mathbf{x}$ is more balanced than $\mathbf{y}$) iff $ord(\mathbf{x}) \leq_{lex} ord(\mathbf{y})$ where vector $ord(\mathbf{v})$ has the ordered elements of $\mathbf{v}$ in descending order, and the relation $\leq_{lex}$ on $\mathbb{R}^N$ is the *lexicographic ordering.*

Example 2.3. i) $(5,1,4,2) \leq_{\mathrm{LQO}} (0,3,5,4)$ because $ord(5,1,4,2) = (5,4,2,1) \leq_{lex} (5,4,3,0) = ord(0,3,5,4)$. ii) $(3,3) \leq_{\mathrm{LQO}} (4,1)$, although $(3,3)$ has more total number of packets than $(4,1)$.

Load Balancing: An LB allocation $W^{\mathrm{LB}} \in \mathcal{W}(\mathbf{b}, C)$ is a non-idling allocation that produces the most balanced future (leftover) queue distribution if, for all $W \in \mathcal{W}(\mathbf{b}, C)$,

$$[\mathbf{b} - \mathbf{1}(W^{\mathrm{LB}} \odot C)]^{+} \leq_{\mathrm{LQO}} [\mathbf{b} - \mathbf{1}(W \odot C)]^{+}. \tag{2.7}$$

2.2.4 Example

Here we show that the incompatibility of the MT and LB policies exists even in a single server case.

Example 2.4. If $\mathbf{b} = (6,2)$ and $C = (1,2)$, then the MT allocation $W^{\mathrm{MT}} = (0,1)$ achieves the throughput of 2 and leaves the remaining queue highly unbalanced at $\mathbf{b} - W^{\mathrm{MT}} = (6,0)$. The system with this unbalanced queue state is unlikely to benefit from any multiuser diversity in the next timeslot. In contrast, the load-balancing allocation $W^{\mathrm{LB}} = (1,0)$ sacrifices the throughput with the balancedness of the remaining queues at $(5,2)$.

Note that, in this example, the two goals of throughput maximization and queue balancing cannot be achieved simultaneously.

2.3 Special Case: On-Off Channel

In this section, we consider a special case of the connectivity process where c_{ij} only takes values 0 (OFF) or 1 (ON). Under this on-off connectivity, we show that 1) a policy that simultaneously maximizes the instantaneous throughput and balances the loads always exists, and 2) for the case of two users, this maximum-throughput and load-balancing (MTLB) policy is an optimal policy for Problem (P).

2.3.1 MTLB Policy

Here we define a class of MTLB policies specific for the on-off channel connectivity.

Definition 2.5. Given state $(\mathbf{b}, C)$ at the beginning of time slot t, the MTLB policy chooses a non-idling feasible allocation $W^* = [w^*_{ij}] \in \mathcal{W}(\mathbf{b}, C)$ such that it satisfies the following two conditions:

(C1) Maximum Throughput (MT): W^* achieves the maximum throughput, i.e., for all $W = [w_{ij}] \in \mathcal{W}(\mathbf{b}, C)$,

$$\sum_{j=1}^{N}\sum_{i=1}^{K} w^*_{ij} \geq \sum_{j=1}^{N}\sum_{i=1}^{K} w_{ij}. \tag{2.8}$$

(C2) Load Balancing (LB): W^* produces the most balanced queue configuration, i.e., for all $W = [w_{ij}] \in \mathcal{W}(\mathbf{b}, C)$,

$$\mathbf{b} - \mathbf{1}W^* \leq_{\text{LQO}} \mathbf{b} - \mathbf{1}W. \tag{2.9}$$

Example 2.6. If $C = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 \end{bmatrix}$ and $\mathbf{b} = [3, 3, 2, 2]$, an MTLB allocation is $W^{\text{MTLB}} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}$, resulting in the leftover queues $\mathbf{b} - \mathbf{1}W^{\text{MTLB}} = [2, 3, 1, 2]$. In addition, $\begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$ is another possible MTLB allocation, resulting in the leftover queues $[3, 2, 2, 1]$. For this $(\mathbf{b}, C)$, there are four possible MTLB allocations. Hence,the MTLB policy is not uniquely defined. However, if $\mathbf{b} = [4, 3, 3, 2]$, then there is only one MTLB allocation $\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}$.

2.3.2 Existence of MTLB Policy

Due to the on-off connectivity, the following result shows that there always exists an MTLB allocation satisfying conditions **(C1)** and **(C2)** at every timeslot and for any $(\mathbf{b}, C)$.

Theorem 2.1. *For any given* $(\mathbf{b}, C)$*, an MTLB allocation always exists.*

Proof. See Appendix A.1. □

2.3.3 Construction of MTLB Policy

In this subsection, we specifically propose an algorithm to construct an MTLB assignment. We first convert the original graph of queues and servers (Figure 2.1) into the following *Equivalent Bipartite Graph* with proper weights on the edges.

Equivalent Bipartite Graph Construction

1. Associated with each queue j, construct $m_j = \min(b_j, \sum_{i=1}^{K} c_{ij})$ nodes labeled as $a_{j1}, a_{j2}, \ldots, a_{jm_j}$.

2. Let $U^{eq} = \{a_{11}, a_{12}, \ldots, a_{1m_1}, a_{21}, \ldots, a_{Nm_N}\}$ be the set of all such nodes.

3. Let $V^{eq} = \{v_i\}_{i=1}^{K}$ be the set of servers.

4. Let $E^{eq} = \{(a_{jm}, v_i) : c_{ij} = 1\}$ be the set of edges representing connectivities.

5. Let $\psi : E^{eq} \mapsto \mathbb{Z}_{++}$, $\psi(a_{jt}, *) = b_j - t + 1$ be the positive integer weight of all edges incident to node a_{jt} in E^{eq}.

To arrive at an MTLB allocation, we run a Maximum Weight Matching (MWM) algorithm on the equivalent bipartite graph. In Proposition 2.9, we show that the resulting assignment satisfies conditions **(C1)** and **(C2)**, hence, it is an MTLB allocation. Before we proceed with Proposition 2.9, we provide the following definitions:

Definition 2.7. [58] Consider a bipartite graph (U, V, E) with two vertex sets U and V, an edge set $E \subseteq U \times V$, and a weight function $\psi : E \mapsto \mathbb{R}$. A *matching* M is a subset of E such that no two edges in M share an endpoint. The *weight of a matching* M is $\psi(M) = \sum_{e \in M} \psi(e)$. A matching M is a *maximum weight matching* (MWM) if its weight is no less than the weight of any other matching.

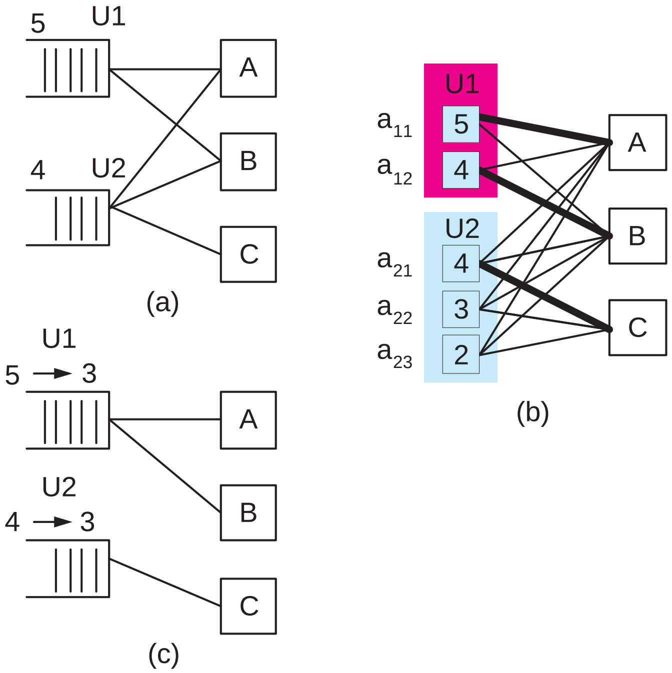


Figure 2.2 Example of MTLB construction (a) queue lengths and connectivities; (b) The equivalent bipartite graph with the weights are shown at each subnode, e.g., the weights of the edges (a_{11}, A) and (a_{11}, B) are five. The thick edges indicate the maximum weight matching; (c) The edges indicate the resulted MTLB assignment.

Definition 2.8. A server allocation $W = [w_{ij}]$ and a matching M are said to be *equivalent* when 1) M is a matching on the equivalent bipartite graph, and 2) $w_{ij} = 1$ if and only if there exists m such that (a_{jm}, v_i) is a matching edge, i.e., $(a_{jm}, v_i) \in M$.

Proposition 2.9. *A maximum weight matching on the equivalent bipartite graph is MTLB, i.e., it satisfies conditions (**C1**) and (**C2**).*

Proof. See Appendix A.1. □

An example of the MTLB assignment based on the proposed algorithm is shown in Figure 2.2. It is intuitive to see that the maximum weight matching on the equivalent bipartite graph achieves an MTLB assignment. This is because the equivalent bipartite graph, in effect, expands the individual packets that can possibly be served into nodes and basically labels each packet with the number of packets waiting behind it (see Figure 2.2(b)). The maximum weight matching selects the matching that serves the packets with the most number of packets waiting behind them. This guarantees the maximum throughput and the load-balancing properties at the same time. Over-assignments are avoided since, in the equivalent bipartite graph, only $\min\left\{b_j, \sum_{i=1}^{K} c_{ij}\right\}$ packets from each node j is expanded. Again, note that since a graph can have multiple maximum weight matchings, MTLB allocation is not unique. The complexity of finding MTLB allocation is equal to the complexity of the existing maximum weight matching algorithm applied to the equivalent bipartite graph which is at most $O(K^2(N +$

$\log K)$) [58].

2.4 Optimality of MTLB Policy for Two Users

With the existence of the MTLB policy, we proceed to establish its optimality:

Theorem 2.2. *Consider Problem (**P**) with on-off connectivity and $N = 2$ users. The MTLB policy is optimal for all choices of g, T, and $\mathcal{I}_0$ as defined in Problem (**P**).*

Remark 2.10. The optimality of the MTLB policy shown in Theorem 2.2 implies that the maximum instantaneous throughput criterion (condition **C1**) is not sufficient to guarantee the delay optimality unless it is complemented by the load-balancing criterion (condition (**C2**)).

To show Theorem 2.2, we use a similar framework as in [20, 21, 34, 49] as follows: we first define a class of functions, $\mathcal{F}$, which contains, for instance, any cost functions $\phi(\mathbf{b})$ of the form $\sum_{j=1}^{N} \xi(b_j)$, where ξ is strictly increasing and convex. We, then, show that if the cost function ϕ belongs to $\mathcal{F}$, the average optimal cost-to-go function (derived via a dynamic programming equation) also belongs to $\mathcal{F}$. The properties of $\mathcal{F}$ are then used to show the optimality of the MTLB policy.

2.4.1 Class of Cost Functions

Here we give the definition of class-$\mathcal{F}$ functions. This is the class of the cost functions for which the solution to Problem (**P**) is proved to be the MTLB policy (Theorem 2.2). But first we define the following for convenience:

Definition 2.11. $R_{ij}(\mathbf{b}) := \mathbf{b} - \mathbf{e}_i + \mathbf{e}_j$, where $\mathbf{e}_m$ is a row vector of zeros except for the m^{th} element which is one, is equivalent to a transfer of a packet from queue i to queue j.

Definition 2.12. A function $f : \mathbb{Z}_+^N \to \mathbb{R}$ belongs to the set $\mathcal{F}$ if, for any $i \neq j \in \{1, \ldots, N\}$, f satisfies the following conditions:

(**B.1**) (monotonicity condition)

$$f(\mathbf{b}) \leq f(\mathbf{b} + \mathbf{e}_i);$$

(**B.2**) (permutation invariance condition)

$f(\mathbf{b}) = f(\pi(\mathbf{b}))$ for any permutation π;

(**B.3**) (supermodularity condition)

$$f(\mathbf{b} + \mathbf{e}_i) - f(\mathbf{b}) \leq f(\mathbf{b} + \mathbf{e}_i + \mathbf{e}_j) - f(\mathbf{b} + \mathbf{e}_j);$$

(**B.4**) (coordinate-wise convexity condition)

$$2f(\mathbf{b}) \leq f(\mathbf{b} + \mathbf{e}_i) + f(\mathbf{b} - \mathbf{e}_i);$$

(**B.5**) (convexity along a constant-sum line condition)

$2f(\mathbf{b}) \leq f(R_{ij}(\mathbf{b})) + f(R_{ji}(\mathbf{b}))$; and

(**B.6**) (balancing advantage condition)

$f(R_{ij}(\mathbf{b})) \leq f(\mathbf{b})$ if and only if $b_i \geq b_j + 1$.

The terminologies in (**B.3**)-(**B.5**) follow that used in [49]. Conditions (**B.3**)-(**B.5**) are second-order relations related to convexity over lattice spaces.[2] Condition (**B.6**) establishes the optimality of the MTLB policy. It can be easily shown that:

Fact 2.13. *Any function of the form* $\phi(\mathbf{b}) = \sum_{j=1}^{N} \xi(b_j)$*, where* ξ *is strictly increasing and convex, belongs to* $\mathcal{F}$*.*

2.4.2 Dynamic Programming Formulation

Next we use a dynamic programming approach to relate the cost function in (2.4) to the expected cost-to-go $V_n^\sigma(\mathbf{b}, C)$ at time $t = T - n$ (i.e., at horizon n) under a Markovian policy σ. Let allocation $W^\sigma(\mathbf{b}, C) \in \mathcal{W}(\mathbf{b}, C)$ denote the allocation at state $(\mathbf{b}, C)$ prescribed by policy σ. Note that, in general, the action $W^\sigma(\mathbf{b}, C)$ depends on horizon n and is assumed implicitly. It is clear that the following recursion:

$$V_0^\sigma(\mathbf{b}, C) = \phi(\mathbf{b}),$$

and for $n \geq 1$,

$$V_n^\sigma(\mathbf{b}, C) = \phi(\mathbf{b}) + E_{\mathbf{a},\tilde{\mathbf{C}}}[V_{n-1}^\sigma(\mathbf{b} + \mathbf{a} - \mathbf{1}W^\sigma(\mathbf{b}, C), \tilde{C})] \tag{2.10}$$

is related to the cost function in (2.4) as

$$J_T^\sigma = E_{\mathbf{C}}[V_T^\sigma(\mathbf{b}, C)] \tag{2.11}$$

[2] Due to the symmetric assumptions (i.e., condition (**B.2**)) in our model, conditions (**B.3**)-(**B.5**) are special cases of the multimodularity condition in [1].

when $\mathcal{I}_0 = \mathbf{b}$. This is due to the validity of the dynamic programming theorem for a finite horizon Markov Decision Process (MDP) [51]. Define

$$V_n^*(\mathbf{b}, C) := \min_{\sigma \in \mathcal{U}_n} V_n^\sigma(\mathbf{b}, C) \tag{2.12}$$

to be the minimum cost-to-go over the set $\mathcal{U}_n$ of all Markovian policies at horizon n. Furthermore, we define the *average optimal cost-to-go* function as

$$v_n(\mathbf{b}) := E_{\mathbf{a},\mathbf{C}}\left[V_n^*(\mathbf{b}+\mathbf{a}, C)\right]. \tag{2.13}$$

In the following Proposition we show the recursive structure of v_n.

Proposition 2.14. *Given a horizon n, the average optimal cost-to-go at n, $v_n(\mathbf{b})$, satisfies the following recursions:*

$$\begin{aligned} v_0(\mathbf{b}) &= \bar{\phi}(\mathbf{b}) := E_{\mathbf{a}}\left[\phi(\mathbf{a}+\mathbf{b})\right] \\ v_n(\mathbf{b}) &= \bar{\phi}(\mathbf{b}) + E_{\mathbf{a},\mathbf{C}}\left[\min_{W \in \mathcal{W}(C)} v_{n-1}\left([\mathbf{b}+\mathbf{a}-\mathbf{1}W]^+\right)\right]. \end{aligned} \tag{2.14}$$

Proof. From the recursion in (2.10), we have the following recursion for the optimal cost-to-go $V_n^*(\mathbf{b}, C)$:

$$\begin{aligned} V_0^*(\mathbf{b}, C) &= \phi(\mathbf{b}) \\ V_n^*(\mathbf{b}, C) &= \phi(\mathbf{b}) + \min_{W \in \mathcal{W}(\mathbf{b},C)} E_{\mathbf{a},\tilde{\mathbf{C}}}[V_{n-1}^*(\mathbf{b}+\mathbf{a}-\mathbf{1}W, \tilde{C})] \\ &= \phi(\mathbf{b}) + E_{\mathbf{a},\tilde{\mathbf{C}}}\left[V_{n-1}^*(\mathbf{b}+\mathbf{a}-\mathbf{1}W^*(\mathbf{b}, C), \tilde{C})\right] \\ &= \phi(\mathbf{b}) + v_{n-1}\left(\mathbf{b}-\mathbf{1}W^*(\mathbf{b}, C)\right), \end{aligned} \tag{2.15}$$

where $W^*(\mathbf{b}, C)$ denotes an optimal allocation at horizon n when the state of the queue backlogs is equal to the vector $\mathbf{b}$ and the connectivity profile is C. In other words, $W^*(\mathbf{b}, C) \in \arg\min_{W \in \mathcal{W}(\mathbf{b},C)} v_{n-1}(\mathbf{b} - \mathbf{1}W)$. Now taking the expectation of both sides we have:

$$\begin{aligned} v_n(\mathbf{b}) &= E_{\mathbf{a},\mathbf{C}}\left[V_n^*(\mathbf{b}+\mathbf{a}, C)\right] \\ &= \bar{\phi}(\mathbf{b}) + E_{\mathbf{a},\mathbf{C}}\left[\min_{W \in \mathcal{W}(\mathbf{b}+\mathbf{a},C)} v_{n-1}(\mathbf{b}+\mathbf{a}-\mathbf{1}W)\right] \\ &= \bar{\phi}(\mathbf{b}) + E_{\mathbf{a},\mathbf{C}}\left[\min_{W \in \mathcal{W}(C)} v_{n-1}([\mathbf{b}+\mathbf{a}-\mathbf{1}W]^+)\right], \end{aligned}$$

where the last equality holds because, for any allocation $W \in \mathcal{W}(C)$, there exists $W' \in \mathcal{W}(\mathbf{b}, C)$ such that $\mathbf{b} - \mathbf{1}W' = [\mathbf{b} - \mathbf{1}W]^+$. Finally, $v_0(\mathbf{b}) = E_{\mathbf{a},\mathbf{C}}\left[V_0^*(\mathbf{b}+\mathbf{a}, C)\right]$ $= \bar{\phi}(\mathbf{b})$. □

2.4.3 Proof of the Optimality

Using the above class of functions $\mathcal{F}$ and the recursive structure of v_n in Proposition 2.14, we are ready to prove Theorem 2.2. Note that the lemmas used here are proved in Appendix A.2.

Proof of Theorem 2.2. We first show the strict monotonicity of v_n for all horizon n, using the strict monotonicity of the cost function. This is shown in the following lemma:

Lemma 2.15. *$v_n(\mathbf{b})$ is strictly increasing on $\mathbf{b}$ for all $n = 0, \ldots, T$, i.e., $\mathbf{b}' > \mathbf{b} \Rightarrow v_n(\mathbf{b}') > v_n(\mathbf{b})$.*

Next, we show that, for any horizon n, the MTLB policy is optimal at time $n + 1$ whenever $v_n \in \mathcal{F}$. This is shown in the following lemma:

Lemma 2.16. *For any horizon n, if $v_n \in \mathcal{F}$, then the MTLB policy is optimal at horizon $n+1$.*

The above lemma immediately establishes Theorem 2.2 if we can show that $v_n \in \mathcal{F}$ for all $n = 0, \ldots, T$. To show that $v_n \in \mathcal{F}$ for all $n = 0, \ldots, T$, we use the following induction:

Induction basis: From Proposition 2.14, $v_0(\mathbf{b}) = \bar{\phi}(\mathbf{b}) = \sum_{\mathbf{a}} P_{\mathbf{a}}(\mathbf{a})\phi(\mathbf{b} + \mathbf{a})$, where $P_{\mathbf{a}}(\mathbf{a})$ is the probability of the arrival vector is $\mathbf{a}$. Using Fact 2.20 below, $v_0 \in \mathcal{F}$ since $\phi \in \mathcal{F}$.

Induction step: Suppose $v_n \in \mathcal{F}$. To show that $v_{n+1} \in \mathcal{F}$, we recall from Proposition 2.14 that

$$v_{n+1}(\mathbf{b}) = \bar{\phi}(\mathbf{b}) + E_{\mathbf{a},C}\left[\min_{W \in \mathcal{W}(C)} v_n([\mathbf{b}+\mathbf{a}-\mathbf{1}W]^+)\right]. \tag{2.16}$$

We find that it is more convenient to work with a relaxed version of v_n which allows the queue vector to be negative. That is, we work with $\hat{v}_n$ where $\hat{v}_n(\mathbf{b}) = v_n([\mathbf{b}]^+)$ for $\mathbf{b} \in \mathbb{Z}^N$. This relaxation technique (used in [20, 21, 34, 49]) removes the need for the separate treatment of various boundary cases. To facilitate this relaxation, we define an extended class of functions $\hat{\mathcal{F}}$ as follows:

Definition 2.17. Consider $f : \mathbb{Z}_+^N \to \mathbb{R}$. We denote $\hat{f} : \mathbb{Z}^N \to \mathbb{R}$ as an extension of f on $\mathbb{Z}^N$ such that $\hat{f}(\mathbf{b}) = f([\mathbf{b}]^+)$. Furthermore, we define an extension $\hat{\mathcal{F}}$ of $\mathcal{F}$:

$$\hat{\mathcal{F}} := \left\{\hat{f} : \mathbb{Z}^N \to \mathbb{R} : \ \hat{f} \text{ meets } (\mathbf{B.1}) \text{ to } (\mathbf{B.6})\right\} \tag{2.17}$$

With this extension, it is clear that v_{n+1} in (2.16) is the restriction of $\hat{v}_{n+1}$ to the non-negative domains, where for $\mathbf{b} \in \mathbb{Z}^N$,

$$\hat{v}_{n+1}(\mathbf{b}) = \hat{\bar{\phi}}(\mathbf{b}) + E_{\mathbf{a},C}\left[\min_{W \in \mathcal{W}(C)} \hat{v}_n(\mathbf{b}+\mathbf{a}-\mathbf{1}W)\right]. \tag{2.18}$$

Now, we show that $v_{n+1} \in \mathcal{F}$ via the following four steps:

Assuming $v_n \in \mathcal{F}$

$\xrightarrow{\text{Step 1}} \hat{v}_n \in \hat{\mathcal{F}}$ (Fact 2.18)

$\xrightarrow{\text{Step 2}} E_{\mathbf{a},\mathbf{C}}\left[\min_{W \in \mathcal{W}(C)} \hat{v}_n(\mathbf{b}+\mathbf{a}-\mathbf{1}W)\right]$ satisfies (**B.3**) to (**B.6**)

(Lemmas 2.23 and 2.24)

$\xrightarrow{\text{Step 3}} \hat{v}_{n+1} \in \hat{\mathcal{F}}$ (Lemmas 2.15, 2.22 and Facts 2.18, 2.19)

$\xrightarrow{\text{Step 4}} v_{n+1} \in \mathcal{F}$ (Fact 2.21)

where the facts and lemmas in the parentheses indicate how to establish the above steps. For completeness, the statements of these facts and lemmas are listed below after the proof. All steps except Step 3 are immediate from the listed facts and lemmas. In Step 3, we first show that $\hat{v}_{n+1}$ satisfies (**B.3**) to (**B.6**) (note that $\bar{\phi} \in \mathcal{F}$ and Fact 2.18 imply that $\hat{\bar{\phi}}(\cdot) \in \hat{\mathcal{F}}$). Then, using Lemmas 2.15 and 2.22 and Fact 2.18, $\hat{v}_{n+1}$ satisfies (**B.1**) and (**B.2**) as well. Hence, $\hat{v}_{n+1} \in \hat{\mathcal{F}}$. Note that Lemmas 2.23 and 2.24, which establish Step 2, also use the fact (Lemma 2.16) that the optimal allocation W at horizon $n+1$ is MTLB if $\hat{v}_n \in \hat{\mathcal{F}}$.[3] □

Here we list the facts and the (remaining) lemmas used in the above proof.

[3] Although Lemma 2.16 assumes $v_n \in \mathcal{F}$, it is easy to show that the lemma works with $\hat{v}_n \in \hat{\mathcal{F}}$ as well.

All the facts are taken from [34], [21] and can be easily verified. The lemmas are proved in Appendix A.2.

Fact 2.18. *If $f \in \mathcal{F}$, then the function $\hat{f} : \mathbb{Z}^N \to \mathbb{R}$ defined as $\hat{f}(\mathbf{b}) = f([\mathbf{b}]^+)$ is in $\hat{\mathcal{F}}$.*

Fact 2.19. *If $\hat{f}_1, \hat{f}_2, \ldots$ are functions that belong to $\hat{\mathcal{F}}$, then $\hat{h}(\mathbf{b}) = \sum_l p_l \hat{f}_l(\mathbf{b})$ also belongs to $\hat{\mathcal{F}}$, where p_l are non-negative constants.*

Fact 2.20. *If $f_1, f_2, \ldots$ are functions that belong to $\mathcal{F}$, then $h(\mathbf{b}) = \sum_l p_l f_l(\mathbf{b})$ also belongs to $\mathcal{F}$, where p_l are non-negative constants.*

Fact 2.21. *If $\hat{f} \in \hat{\mathcal{F}}$, then the restriction of $\hat{f}$ to non-negative domain is in $\mathcal{F}$.*

Lemma 2.22. *$v_n(\mathbf{b})$ is permutation invariant on $\mathbf{b}$ for all $n = 0, \ldots, T$.*

Lemma 2.23. *Assuming $N = 2$ and $\hat{v}_n \in \hat{\mathcal{F}}$. For any state $\mathbf{b}$,*

$$E_{\mathbf{a},\mathbf{C}} \left[\min_{W \in \mathcal{W}(C)} \hat{v}_n(\mathbf{b} + \mathbf{a} - \mathbf{1}W) \right]$$

*satisfies (**B.3**), (**B.4**), and (**B.5**).*

Lemma 2.24. *Assuming $N = 2$ and $\hat{v}_n \in \hat{\mathcal{F}}$. For any state $\mathbf{b}$ such that $b_1 \geq b_2 + 1$, $E_{\mathbf{a},\mathbf{C}} \left[\min_{W \in \mathcal{W}(C)} \hat{v}_n(\mathbf{b} + \mathbf{a} - \mathbf{1}W) \right]$ satisfies condition (**B.6**).*

Remark 2.25. All the above lemmas and facts, except Lemmas 2.23 and 2.24, hold for general N. Lemmas 2.23 and 2.24 are proved for $N = 2$. We detail the difficulty in extending these lemmas to general N after Theorem 2.3 in Section 2.5 and also after the proof of Lemma 2.24 in Remark A.11 in Appendix A.2.

Remark 2.26. Theorem 2.2, in addition, can be extended to the optimality of the MTLB policy in an expected average cost sense for an infinite horizon problem.

Corollary 2.27. *Consider an infinite horizon version of Problem (**P**), where the cost is modified to be the average expected cost at each horizon. Then the MTLB policy is optimal for any initial state $\mathcal{I}_0 = \mathbf{b}$.*

Proof. Theorem 2.2 proves that there exists a stationary MTLB policy which is optimal for Problem (P) for any finite horizon T. Hence, our MTLB policy achieves the minimization of the average expected cost Λ_T^π/T for any finite horizon T. Since the policy is independent of the horizon T, it is optimal with respect to an average expected cost criterion for the infinite horizon version of the problem. □

2.5 Optimality of MTLB Policy under Fluid Relaxation

In the previous section, we established the optimality of the MTLB policy for a very restricted case of $N = 2$. As we will see later, the major difficulty in extending the proof to general N is due to the *integral* server allocation constraint. In this section, we study a relaxed system where we allow each server to serve a fractional (*fluid*) number of packets from queues as long as the total number of packets served per server is no greater than one. In other words, we consider a real allocation $W = [w_{ij}] \in \mathbb{R}^{K \times N}$, with $w_{ij} \in [0, 1]$. We call this relaxed constraint the *fluid server allocation relaxation* or *fluid relaxation*. Under this fluid relaxation, we can show the optimality of the MTLB policy for general N and the on-off channel model. Before we proceed, we provide a modification of $\mathcal{W}(\mathbf{b}, C)$ to include fluid allocations and a definition of the fluid version of the MTLB policy:

Definition 2.28. A class of *fluid* non-idling feasible allocations $\mathcal{W}^f(\mathbf{b}, C)$ is equivalent to $\mathcal{W}(\mathbf{b}, C)$ with the fluid server allocation condition, i.e., $W = [w_{ij}] \in \mathcal{W}^f(\mathbf{b}, C)$ if

(**a'**) $0 \leq w_{ij} \leq 1$;

(**b**) $c_{ij} = 0 \Rightarrow w_{ij} = 0$;

(**c**) $\sum_{j=1}^{N} w_{ij} \leq 1, \ \forall \ i = 1, \ldots, K$; and

(**d**) $\sum_{i=1}^{K} w_{ij} \leq b_j, \ \forall \ j = 1, \ldots, N$.

Definition 2.29. The MTLB-F policy is a fluid version of the MTLB policy defined in Section 2.3.1, i.e., the MTLB-F policy chooses $W^* \in \mathcal{W}^f(\mathbf{b}, C)$ such that (**C1**) and (**C2**) are satisfied.

Example 2.30. It is clear that the MTLB-F policy is not uniquely defined although the leftover queue vector is. For example, for the $C = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 \end{bmatrix}$ in Example 2.6, if $\mathbf{b} = [3, 3, 3.3, 3.1]$ then all allocations of the form

$$\begin{bmatrix} 0.4 & 0.4 & x & 0.2 - x \\ 0 & 0 & y & 1 - y \end{bmatrix},$$

where $x \in [0, 0.2]$ and $y = 0.7 - x$, are MTLB-F allocations, resulting in the unique leftover queues $[2.6, 2.6, 2.6, 2.6]$.

Assume that the cost function $\phi(\mathbf{b})$ is convex on $\mathbf{b} \in \mathbb{R}_+^N$, we have the following result:

Theorem 2.3. *For the problem (**P**) with the fluid server allocation relaxation, the MTLB-F policy is optimal.*

Proof. See Appendix A.3. □

The key element in the proof of Theorem 2.3 is the convexity of the cost-to-go function v_n in the fluid relaxation, for all $n = 0, \ldots, T$. This convexity property directly establishes the optimality of the MTLB-F policy. However, under the integral server allocation constraint, it is hard to establish a similar convexity property over lattices. In fact, one can interpret the difficulty in establishing Lemma 2.23 and 2.24 as an indication that the properties (**B3**) - (**B6**) of the set $\mathcal{F}$ provide a sufficient set of properties for convex functions over two dimensional lattices, while they fail to sufficiently capture the convexity in higher dimensional lattices.

2.6 Application of MTLB Policy in OFDMA Wireless Systems

This section considers the issue of delay optimal subcarrier allocation in OFDMA wireless networks. In previous studies, we have shown that the optimal policy is complicated and unknown when the channel connectivities are more general than the on-off channel model. However, based on the insights learned from the MTLB policy, this section provides heuristic policies that use different degrees of channel and queue state information. More importantly, these examples show how the significance of queue vs. channel state information varies with the traffic load. This is of extreme practical interest when one considers the overhead associated with channel estimation and feedback.

2.6.1 Heuristic Policies and Value of QSI versus CSI

1) *Algo-I* (full QSI, On-Off CSI): The subcarrier assignment uses full information about the queue lengths (full QSI) and binary (ON-OFF) information about the channel: a subcarrier is considered ON if $c_{ij} \geq c_{\text{threshold}}$[4]. Then, MTLB policy [43] is used for subcarrier allocation.

Algo-I:

- $\bar{c}_{ij} = \begin{cases} 1 & \text{if } c_{ij} \geq c_{\text{threshold}}, \\ 0 & \text{otherwise.} \end{cases}$
- For state $(\mathbf{b}/c_{\text{threshold}}, \bar{C})$ compute an MTLB allocation W^*.

2) *Algo-II* (full CSI, On-Off QSI): The subcarrier allocation uses full information about the channel (full CSI) and minimal information about the queue lengths (On-Off QSI). The subcarrier allocation considers only the queues which have some data to transmit.

[4] The threshold is arbitrarily chosen. It should be adjusted depending on the load of the system and the number of subcarrier and users. We assume here that the threshold is fixed in our simulation.

Algo-II:

- Assign $W^* = \left[w^*_{ij}\right]$ such that $w^*_{ij^*(i)} = 1$ where

$$j^*(i) = \arg \max_{j \in \{1,\ldots,N\}} c_{ij} I_{\{b_j > 0\}}.$$

3) *Algo-III* (full CSI and full QSI): Algo-III is the Maximum-Weight policy proposed in [39] where each subcarrier is assigned to the queues to achieve the highest value of $c_{ij} b_i$.

Algo-III:

- Assign $W^* = \left[w^*_{ij}\right]$ such that $w^*_{ij^*(i)} = 1$ where

$$j^*(i) = \arg \max_{j \in \{1,\ldots,N\}} b_j c_{ij}.;$$

4) *Algo-IV* (full CSI and full QSI): Since Algo-III may over assign subcarriers to some users, and with an insight into the significance of load-balancing, we modify the known Algo-III to avoid imbalanced queues.

Algo-IV:

```
• $X = \{1, \ldots, K\}$;
• Loop (until stop):
    – If $X = \emptyset$, then stop;
    – $(i^*, j^*) = \arg\max_{i \in X, j \in \{1,\ldots,N\}} b_j c_{ij}$;
    – If $b_{j^*} c_{i^* j^*} > 0$, then $w^*_{i^* j^*} = 1$ else stop;
    – $b_{j^*} = b_{j^*} - c_{i^* j^*}$ and $X = X - \{i^*\}$;
• Assign $W^* = \left[w^*_{ij}\right]$.
```

2.6.2 Numerical Comparisons and Simulation

We consider a downlink OFDMA system in a single cell with one base station composed of $N = 32$ statistically independent and identical users and $K = 128$ subcarriers. We generate a frequency-selective channel by using 26-tap multipath with exponential intensity profile and use adaptive QAM modulation. We use the fact that there is only a very small loss of channel capacity if a white power spectrum is used (i.e. each subcarrier receives equal power) instead of the optimal power spectrum [16]. The channel gain h_{ij} can be mapped to the number of packets per time slot, c_{ij}, that subcarrier i can potentially transmit for user j as [16]:

$$c_{ij} = \frac{D}{\beta} \max \left\{ 0, \left\lfloor 0.31(10 \log_{10}(\frac{P|h_{ij}|^2}{KN_o}) - 6.7) \right\rfloor \right\} \tag{2.19}$$

where D the number of QAM symbols per channel in a timeslot, β the fixed packet length (in bits) and N_o is the noise power in the subcarrier. The parameters P, D, β and N_o are chosen such that the allocation of subcarriers over a block is equivalent to the server scheduling problem where the connectivity $c_{ij} \in \{0, 1, 2, 3\}$. All simulations are conducted over 6,000 timeslots. We consider arrivals of fixed-size packets where the number of arrivals per timeslot for each queue is a random variable having one of the two distributions: bounded pareto ($\alpha = 2$ and $x_{min}/E(x) = 1/2$) [79] and Poisson.

Figure 2.3 and Figure 2.4 provide comparisons of the performance of the proposed algorithms under different traffic models in terms of the average total queue backlog (equivalently, in terms of the average delay by Little's Theorem). For all traffic types, Algo-IV, as expected, outperform the others, since it mimics IMT and LB policies as closely as possible. This observation is consistent with those studies in literature which take advantage of backlog information (e.g. [53] and [71]). However, the more interesting and important observation is the performance

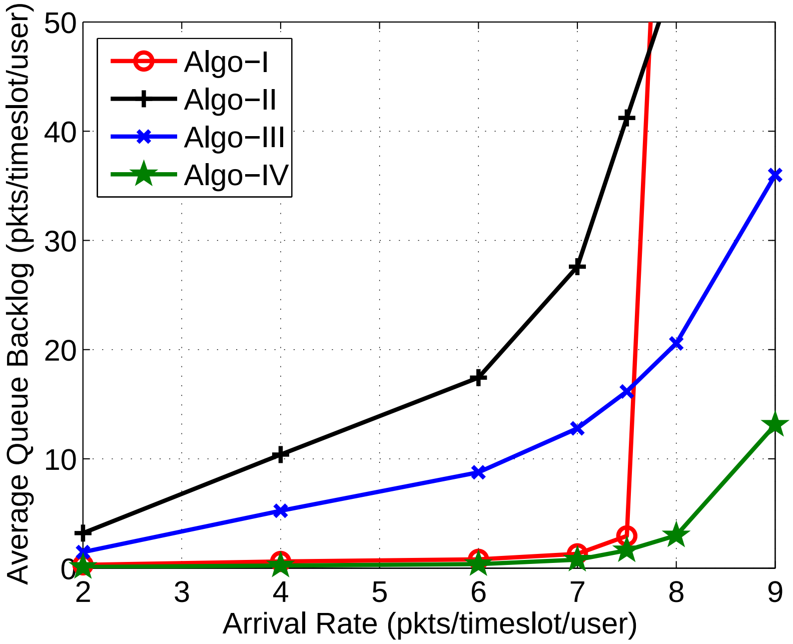


Figure 2.3 Average queue backlog for bounded pareto distribution.

of Algo-I in the light-to-moderate traffic regime (below 6 packets/user/timeslot). Before the performance of Algo-I sees a sharp degradation reflecting the policy's low throughput, it outperforms Algo-II and Algo-III in light-to-moderate traffic, even though it does ignore much of CSI available to the other algorithms. This insight sheds light on nature of delay performance versus throughput considerations and the benefit of using queue information. When considering light-to-moderate traffic intensity (resulting in reasonable delays), the value of QSI outplays that of CSI. This is of extreme practical interest when one considers the overhead associated with CSI estimation and feedback in an OFDMA system with large number of subcarriers.

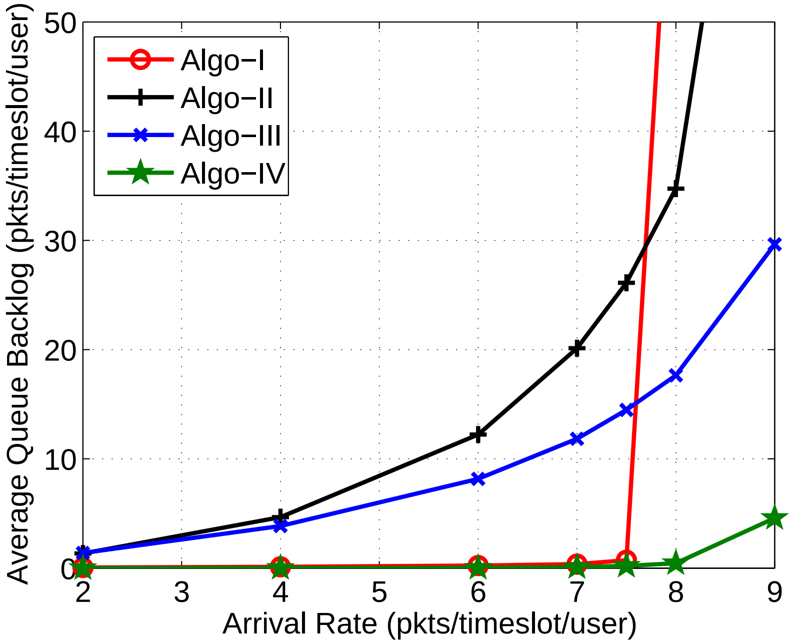


Figure 2.4 Average queue backlog for the Poisson distribution.

2.7 Summary

In this chapter, we have considered the problem of optimal server allocation in a time-slotted system with N symmetric queues and K servers when the arrivals and channels are stochastic and time-varying. Focusing on a long-term average-delay objective, we identified the MTLB policy that achieves the instantaneous maximum throughput as well as balancing the queue lengths. Such a policy always exists when the channel connectivity follows an on/off model. In such a case, we proved that the MTLB policy achieves the minimum average delay (mean response time) at any time when there are only $N = 2$ users.

Although the on-off connectivity model may seem restricted for practical wireless systems, the optimality of the MTLB policy in such on-off model can lead to useful insight for the general connectivity setting. We have obtained excellent performance for the general connectivity using some MTLB-based heuristic policies. We showed by simulation that the value of CSI and QSI in optimizing the performance heavily depends on the arrival statistics. We showed that in low-to-moderate traffic regime and from a delay optimality perspective, balancing the queues is more critical than opportunistically taking advantage of CSI. The opposite becomes true in the heavy traffic regime.

Acknowledgment

This chapter, in part, appears in the following publications. The dissertation author was the primary investigator and author of these papers.

- S. Kittipiyakul and T. Javidi, “Delay-optimal server allocation in multi-queue multi-server systems with time-varying connectivities,” *submitted to IEEE Trans. Inf. Theory.*
- S. Kittipiyakul and T. Javidi, “Resource allocation in OFDMA with time-varying channel and bursty arrivals,” *IEEE Commun. Lett.*, vol. 11, no. 9, pp. 708–710, Sept. 2007.

I would like to thank my coauthor Prof. Tara Javidi.

Chapter 3

High-SNR Analysis of Outage-Limited Communications with Bursty and Delay-Limited Information

Abstract

This chapter analyzes the high-SNR asymptotic error performance of outage-limited communications with fading, where the number of bits that arrive at the transmitter during any time slot is random but the delivery of bits at the receiver must adhere to a strict delay limitation. Specifically, bit errors are caused by erroneous decoding at the receiver or violation of the strict delay constraint. Under certain scaling of the statistics of the bit-arrival process with SNR,

this paper shows that the optimal decay behavior of the asymptotic total probability of bit error depends on how fast the burstiness of the source scales down with SNR. If the source burstiness scales down too slowly, the total probability of error is asymptotically dominated by delay-violation events. On the other hand, if the source burstiness scales down too quickly, the total probability of error is asymptotically dominated by channel-error events. However, at the proper scaling, where the burstiness scales linearly with $\frac{1}{\sqrt{\log \mathrm{SNR}}}$ and at the optimal coding duration and transmission rate, the occurrences of channel errors and delay-violation errors are asymptotically balanced. In this latter case, the optimal exponent of the total probability of error reveals a tradeoff that addresses the question of how much of the allowable time and rate should be used for gaining reliability over the channel and how much for accommodating the burstiness with delay constraints.

This chapter analyzes the high signal-to-noise-ratio (SNR) performance of outage-limited communications where the information to be communicated is delay-limited and where the information arrives at the transmitter in a stochastic manner. We consider the following setting (Figure 3.1) in our study:

- A random number of bits arrive at the transmitter during any given timeslot. Bits accumulate in an infinite buffer while waiting for their turn to be bunched into codewords and transmitted under a first-come, first-transmit policy.

- There is no feedback to the transmitter; retransmission of the bits in error

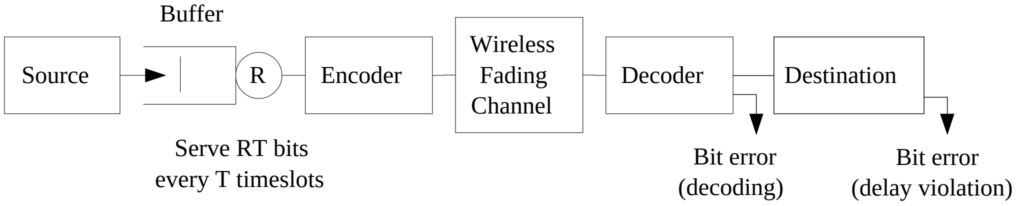


Figure 3.1 System model for point-to-point communication

is not considered.

- Communication over the fading channel is outage-limited ([62, 90]), where the transmitter is unaware of the instantaneous channel state and, as a consequence, operates at a fixed transmission rate, R. During a deep fade (also known as an outage), the channel seen by the decoder is too weak to allow recovery of the data content from the transmitted signal. Characteristic settings are those of MIMO and cooperative outage-limited communications.

- Coding takes place in blocks where each codeword spans over a fixed and finite integral number, T, of timeslots. Each codeword has an information content of RT bits. In addition, coding is “fully-diverse,” i.e., the decoding at the receiver takes place only at the end of the coding block.

- The delay bound, D, is a maximum allowable time duration from the moment a bit arrives at the transmitter until the moment it is decoded at the receiver. The delay experienced by a bit is the sum of the time spent waiting in the buffer and the time spent in the block decoding process. Note that the waiting time in the buffer is random due to the stochastic arrival process.

- A bit is declared in error either when it is decoded incorrectly at the decoder, or when it violates the delay bound.

For the above setting, we are interested in the high-SNR asymptotic total probability of bit error. Note that for a given transmission rate, R, and a coding block duration, T, there exists a tradeoff between the probabilities of decoding error versus the delay violation. We expect that longer coding blocks allow the encoded bits to be transmitted over more fading realizations and hence, achieve higher diversity and fewer decoding errors. However, longer coding blocks cause more bits to violate the delay requirement. In other words, one intuitively expects that there is an optimal choice of the fixed transmission rate, R, and the fixed coding block duration, T, for which the total probability of bit error is minimized. The goal of this chapter is to analytically identify these optimal quantities.

3.0.1 Prior Work and Our Contribution

High demands on the quality of service (QoS), in terms of both packet losses and packet delays, have fueled substantial research interest in jointly considering channels and queues. Communication of delay-sensitive bits over wireless channels has been addressed under various assumptions and settings in works such as [5, 6, 10, 55, 61, 63]. Often, asymptotic approximations are employed to enable tractable analysis of the problem. Below we detail the existing work with their corresponding settings and the relationships to this chapter.

The first group we discuss, [5, 6, 61, 63], consists of scenarios where the current channel state information (CSI) is assumed to be known at both the transmitter and receiver. For example, in [6] and [5], Berry and Gallager address the tradeoff between the minimum average power consumption and the average delay (the power-delay tradeoff) over a Markovian fading channel with CSI both at the transmitter and the receiver. In such a setting, the transmitter dynamically varies power (i.e., the rate) in response to the current queue length and channel state. In [63], Rajan et al. derive optimal delay-bounded schedulers for transmission of constant-rate traffic over finite-state fading channels. In [61], Negi and Goel apply the effective capacity [84] and error exponent techniques to find the code-rate allocation that maximizes the decay rate of the asymptotic probability of error for a given asymptotically-large delay requirement. Similar to [6] and [5], the proposed dynamic code-rate allocation in [61] is in response to the current channel fading and is possible by assuming CSI knowledge at the transmitter.

A second group of work (e.g., [10, 55]) focuses on scenarios where CSI is unknown to the transmitter but there is a mechanism for retransmission of codewords when the channel is in outage. As a tradeoff to protection against channel outage, this retransmission incurs extra delays to the bits in the buffer. In [10], for example, Bettesh and Shamai (Shitz) address the problem of minimizing the average delay, under average power constraints and fixed transmission rate. They provide asymptotic analysis, under heavy load condition and asymptotically large queue length, for the optimal adaptive policies that adjust the transmission rate and/or transmission power in response to the current queue length at the transmitter. In another example, Liu et al. in [55] study the problem of optimal (fixed) transmission rate to maximize the decay rate of the probability of buffer overflow for on-off channels and Markov-modulated arrivals. The channel is considered “off” when outage occurs.

Although our work uses a similar performance measure to [61], namely the decay rate of the asymptotic probability of error, it covers the scenarios in which CSI is not available to the transmitter (no CSIT) and there is no retransmission. In such a setting, the variation of the fading channel is combatted via a coding over multiple independent fading realizations.[1] While this approach improves the transmission reliability, its longer coding duration increases the end-to-end delay any bit faces, and can potentially increase the probability of delay violation. In other words, in the absence of CSIT and retransmission, the transmission reliability, as well as the delay violation probability, are functions of the coding rate and duration. Consequently, our work compliments this previous research as it considers the effect of a delay violation requirement, in the absence of CSI at the transmitter and retransmission, on the operation of the physical layer. We consider a fixed transmission rate and code duration, as opposed to dynamic policies.

Since it is difficult to derive the exact relationship between the system parameters and the probabilities of channel decoding error and the delay violation, we choose to study an asymptotic approximation when the signal-to-noise ratio (SNR)

[1] For example, the multiple independent fading realizations can be a result of fading in multiple channel coherence time intervals (known as time diversity), or fading in multiple independent spatial channels, as in MIMO channel (spatial diversity), or cooperative relay channel (cooperative diversity).

is asymptotically high. The first advantage of this choice is the availability of an asymptotic high-SNR analysis for the channel decoding error probability. This high-SNR analysis is known as the *diversity-multiplexing-tradeoff* (DMT) analysis [90] and has received a great deal of attention during the past few years. Another advantage of the high-SNR analysis is that, for the class of arrival processes we consider in this chapter, we can derive an asymptotic approximation of the delay violation probability that is valid even when the delay requirement D is finite and small. This derivation (Lemma 3.10) is based on a large-deviations result known as the Gärtner-Ellis theorem (see e.g., [18]) and extends the large deviations exponent for a queue with asymptotic number of flows (as provided in [11, 15, 29, 81]) to a queue with batch service discipline. Given that the asymptotic expression of the total probability of bit error is valid without requiring asymptotically large D, it is then meaningful to ask about the optimal coding block duration, a question which is not answered in studies with asymptotic D (e.g., [5, 6, 26, 44, 45, 61]).

We also would like to point out that our work was motivated by the work of Holliday and Goldsmith [36] where, under a high-SNR asymptotic approximation, the optimal operating channel transmission rate for a concatenated source/channel system is studied. Following the approach in [36], we study a concatenated queue/channel system under a high-SNR approximation.

3.0.2 Overview of the Results

This chapter focuses on the notion of SNR error exponent as a measure of performance. Specifically, we are interested in finding how the asymptotic total probability of error decays with SNR. To keep the problem meaningful, we consider a scenario under which the overall traffic loading of the system (the ratio between the mean arrival rate and the ergodic capacity of the channel) is kept independent of SNR. That is, we consider a case where the arrival rate scales with $\log \mathrm{SNR}$. Note that this scaling of arrival process is necessary to ensure a fixed loading and hence a comparable cross-layer interaction as SNR scales.

From the DMT result, we already know that, if the channel operates below the channel ergodic capacity, the asymptotic probability of channel decoding error decays with SNR. The best one can hope for is that the asymptotic total probability of error decays exponentially with SNR. For that, the asymptotic probability of delay violation needs to decay with SNR. Specifically, we consider a class of i.i.d.[2] arrival processes with light tail (i.e., the processes have all moments finite) whose burstiness (defined as the ratio of the standard deviation over the mean of the number of bits arrived at a timeslot) monotonically goes to zero as SNR goes to infinity. We show that for all such processes (called smoothly-scaling processes), the total probability of error decays.

[2] Note here that the channel is not necessarily i.i.d. in time. Since the adopted channel model is not assumed to be i.i.d., assuming an i.i.d. arrival process, intuitively, is not consequential: think of our chosen time slot as an upperbound for the "coherence time" of the arrival process. The i.i.d. source assumption mostly serves to simplify the exposition and presentation of results, and does not fundamentally limit the setting.

The main result of the chapter shows that the optimal decay behavior of the asymptotic total probability of bit error depends on how fast the burstiness of the source scales down with SNR. If the source burstiness scales down too slowly (too quickly), the majority of the errors are due to delay violation (channel error), i.e., the total probability of error is asymptotically dominated by delay-violation (channel-error) events. However, at the proper scaling where the burstiness scales linearly with $\frac{1}{\sqrt{\log \mathrm{SNR}}}$ and with the optimal coding duration and transmission rate, the occurrences of channel errors and delay-violation errors are asymptotically balanced. Equivalently, one can interpret our result, the optimal choice of block coding duration and transmission rate, as that which balances the channel atypicality (deep fading or outage events) and the arrival atypicality (large bursts of arrivals).

We apply this result to several examples of outage-limited communication systems to find the optimal setting of the operating parameters.

3.0.3 Outline of the Chapter

The precise models for the coding and channel process and the bit-arrival and queue process are described in Section 3.1. We precisely define the scaling of the source process with SNR and give a simple example of such source processes. Section 3.2 provides the asymptotic probability of delay violation. The main result of the chapter is found in Theorem 3.1 of Section 3.3. This theorem provides the optimal asymptotic decay rate of the total error probability as well as the optimal coding duration and transmission rate. Section 3.4 gives some examples to illustrate the utility of Theorem 3.1. These examples consider the question of optimally communicating delay sensitive packet stream with a compound Poisson traffic profile over the following outage-limited channels: SISO Rayleigh fast-fading channel, quasi-static cooperative relay channel, and quasi-static MIMO channel. Section 3.5 concludes the chapter.

3.0.4 Notations

We use the following symbols and notations. We use ρ to denote SNR. The notation $\overset{g}{=}$ for a strictly increasing and positive-valued function g represents the equivalence between $y(\rho) \overset{g}{=} z(\rho)$ and $\lim\limits_{\rho\to\infty} \frac{\log y(\rho)}{g(\log \rho)} = \lim\limits_{\rho\to\infty} \frac{\log z(\rho)}{g(\log \rho)}$. We define $\overset{g}{\geq}$ and $\overset{g}{\leq}$ in a similar manner. Note that when g is an identity function, then $\overset{g}{=}$ is equivalent to the familiar $\doteq$ notation in the DMT analysis [90].

We denote the high-SNR approximation of the ergodic capacity of AWGN channel by $N := \log \rho$ and use N and $\log \rho$ interchangeably. The sets $\mathbb{Z}$, $\mathbb{N}$, and $\mathbb{Z}^+$ represent the set of all, positive, and non-negative integers, respectively. In addition, the set $\mathbb{T}$ represents the set $\{1, 2, \ldots, \lfloor \frac{D}{2} \rfloor\}$. Flooring and ceiling functions are denoted by $\lfloor\cdot\rfloor$ and $\lceil\cdot\rceil$, respectively. For all $a \leq b$, $[x]_a^b = \max(a, \min(b, x))$ and $[x]^+ = \max(x, 0)$. We write $g(x) = \Theta(h(x))$ to denote that the function g scales linearly with the function h, i.e., $\lim\limits_{x\to\infty} \frac{g(x)}{h(x)} < \infty$ and $\lim\limits_{x\to\infty} \frac{h(x)}{g(x)} < \infty$. Finally, for any function f, we denote its convex conjugate or the Fenchel-Legendre transform, f^*, by

$$f^*(x) = \sup_{\theta \in \mathbb{R}} \ \theta x - f(\theta). \tag{3.1}$$

3.1 System Model

As discussed in the introduction, we consider a system composed of a bursty and delay-limited information source, concatenated with an infinite buffer and a fading channel, as shown in Figure 3.1. We assume the queue follows a first-come-first-serve (FCFS) discipline. The departures out of the queue occur according to a block channel coding scheme, while the arrivals to the queue follow a stochastic model. If the transmission rate is above the instantaneous capacity of the channel, an outage event is said to occur where the received signal is erroneously decoded. The delay requirement asks that each bit of information be decoded at the destination within a *maximum allowable delay* of D time-slots from the time it arrives at the buffer. Otherwise, the bit will be obsolete, discarded, and counted as erroneous.We assume no retransmission of unsuccessful transmissions or those bits which violate the delay bound.[3] In the next three subsections, we describe in detail the models for the channel, the arrival process, and the system performance measure.

[3] Note that due to the constant service rate of the queue and the FCFS service discipline, any bits arriving at the queue know immediately whether they will exceed their delay constraints, using the knowledge of the current queue length. It seems wise to drop these bits immediately after their arrivals to improve the system performance. However, we do not need to consider such method because it has been established (see [29, Theorem 7.10]) that, in the asymptotic regime of interest, such method does not improve the exponent of the delay violation probability.

3.1.1 Channel and Coding Model

We consider a general fading-channel model,

$$\underline{y} = H\underline{x} + \underline{w},$$

where $\underline{x}$ is the transmitted vector, H is the channel matrix, $\underline{y}$ is the received signal, and $\underline{w}$ is the noise vector. The average SNR is defined as [90]

$$\rho := \frac{\mathbb{E}[\|H\underline{x}\|^2]}{\mathbb{E}[\|\underline{w}\|^2]},$$

and in the asymptotic scale of interest, it is equivalent to

$$\rho \doteq \mathbb{E}[\|\underline{x}\|^2].$$

Coding takes place over T timeslots, using rate-R, length-T codes that meet the DMT tradeoff $d_{\text{ch}}(r, T)$ [90], defined as

$$d_{\text{ch}}(r, T) := -\lim_{\rho\to\infty} \frac{\log P_{\text{ch}}(r, T, \rho)}{\log \rho}, \tag{3.2}$$

where $P_{\text{ch}}(r, T, \rho)$ is the codeword error probability induced by the channel, given an optimal code of *multiplexing gain* r, coding block size T timeslots[4], and average SNR ρ. The channel multiplexing gain r is related to the transmission rate R as (refer to [90])

$$r := \lim_{\rho\to\infty} \frac{R}{\log \rho}. \tag{3.3}$$

That is, the transmission rate R is assumed to scale linearly as $r \log \rho$. We denote by $r_{\max}$ the maximum value of r, i.e., $0 \le r \le r_{\max}$. This $r_{\max}$ relates to the ergodic

[4] For most settings, there exist codes that meet the entire DMT tradeoff in minimum delay, independent of channel dimensionality and fading statistics [27, 77, 86].

capacity as

$$r_{\max} = \mathbb{E}_H \frac{\max_{p_x} I(\underline{x};\underline{y})}{\log \rho}$$

and is the smallest r such that $d_{\text{ch}}(r,T) = 0$.

The DMT tradeoffs have been extensively studied for various finite duration communication schemes (for example, see [13, 23, 27, 28, 77] for MIMO point-to-point communications, [78] for multiple access communications, [4, 52] for cooperative communications, and [24, 86] for cooperative communications with small delay).

Remark 3.1. The condition that each bit be transmitted over all timeslots in the coding block[5], together with the first-come first-transmit service discipline, makes it so that every T timeslots, the RT oldest bits are instantaneously removed[6] from the queue and are transmitted over the next T timeslots. We assume that it is only at the end of the T timeslots that all the RT bits are decoded by the decoder.

Example 3.2 (Rayleigh Fast-Fading SISO Channel)**.** Consider the single-input single-output (SISO) time-selective channel with Rayleigh fading coefficients (correlated or uncorrelated) and with additive white Gaussian noise at the receiver. The corresponding channel model over T timeslots is given by

$$\underline{y} = \text{diag}(\underline{h})\ \underline{x} + \underline{w},$$

where $\underline{y}, \underline{h}, \underline{x}$, and $\underline{w}$ are $T \times 1$ vectors and $H = \text{diag}(\underline{h})$ is a $T \times T$ diagonal fading

[5] Currently, all known minimum-delay DMT optimal codes over any fading channel with non-zero coefficients ask that each bit be transmitted over each timeslot.

[6] If an insufficient number of bits exists in the buffer, null bits are used and the rate is maintained. It is easy to show that, in the asymptotic scale of interest, the use of null-bits does not incur any change in the SNR exponent of the probability of error.

matrix with the fading at the tth timeslot, h_t, as its (t,t) element. The optimal DMT, given optimal signaling, takes the form

$$\begin{aligned} d_{\text{ch}}(r,T) \quad &:= \quad -\lim_{\rho\to\infty} \frac{\log \Pr\bigl(I(\underline{x};\underline{y}|\underline{h}) < 2^{RT}\bigr)}{\log \rho} \\ &= \quad -\lim_{\rho\to\infty} \frac{\log \Pr\bigl(\prod_{t=1}^{T}(1+\rho|h_t|^2) < \rho^{rT}\bigr)}{\log \rho}. \end{aligned}$$

For the fast-fading case where the coherence time is equal to one timeslot and the elements of $\underline{h}$ are Rayleigh i.i.d. random variables, the tradeoff takes the form

$$d_{\text{ch}}(r,T) = T(1-r), \tag{3.4}$$

and it can be met entirely in T timeslots (see [1]). This SISO channel allows for $r_{\max} = 1$.

Other examples which will be discussed later in Section 3.4 are quasi-static MIMO and cooperative-relay channels. In this chapter, for simplicity we assume that $d_{\text{ch}}(r,T)$ is continuous on r, decreasing on r, and increasing on T.

3.1.2 Smoothly-Scaling Bit-Arrival Process

In this subsection, we describe the SNR-scaling of a family of arrival processes of interest. The specific choice of SNR-scaling for the statistics of the bit-arrival process is such that the average traffic load of the system (defined as the ratio of the average arrival rate over the ergodic capacity) is kept constant, independent of SNR.[7] This means that scaling in the ergodic capacity $r_{\max} \log \rho$

[7] It can be seen that unless the traffic load (average bit arrival rate over the channel rate) scales as log(SNR), i.e., $\lim_{\rho\to\infty} \frac{\mathbb{E}[A_t]}{\log \rho} = \ell$ for some fixed $0 < \ell < \infty$, the problem is void of cross-layer

($=r_{\max}N$) is matched by scaling the average bit-arrival rate as $\lambda \log \rho$ ($=\lambda N$) as well, for some $\lambda > 0$. Now we are ready to introduce the arrival process of interest: The sequence of asymptotically *smoothly-scaling* bit-arrival processes, in which the process becomes "smoother" for increasing N.

Definition 3.3. Let $\mathcal{G}$ denote a class of functions which contains any function $g : \mathbb{R}^+ \mapsto \mathbb{R}^+$ (called *scaling function*) which is continuous and strictly increasing and whose tail behavior is such that

$$\lim_{x\to\infty} \frac{g(x)}{\log x} = \infty. \tag{3.5}$$

Definition 3.4. (*g-smoothly-scaling source*) Consider a scaling function $g \in \mathcal{G}$ and a family of bit-arrival processes $(A^{(N)}, N \in \mathbb{N})$, where $A^{(N)} = (A_t^{(N)}, t \in \mathbb{Z})$ denotes an i.i.d. sequence of the random numbers $A_t^{(N)}$ of bits that arrive at time t with $E[A_t^{(N)}] = \lambda N$, for all t. The family of bit-arrival processes is said to be *g-smoothly-scaling* if the *limiting g-scaled logarithmic moment generating function*, defined for each $\theta \in \mathbb{R}$ as

$$\Lambda(\theta) = \lim_{N\to\infty} \frac{\log E[\exp(\frac{\theta g(N)}{N} A_1^{(N)})]}{g(N)}, \tag{3.6}$$

exists as an extended real number in $\mathbb{R}^* := \mathbb{R}\cup\{\infty\}$ and is finite in a neighborhood of the origin, *essentially smooth*, and *lower-semicontinuous*.[8]

interactions. Otherwise if $\ell = 0$, corresponds to the case where too few bits arrive and effectively there is no queuing delay. On the other hand, when the traffic load scales much faster than log(SNR), i.e., $\ell = \infty$, the overall performance is dominated by queueing delay, independently of the channel characteristics.

[8] [29] A function $f : \mathbb{R} \mapsto \mathbb{R}^*$ is *essentially smooth* if the interior of its effective domain $\mathcal{D} = \{x : f(x) < \infty\}$ is non-empty, if it is differentiable in the interior of $\mathcal{D}$ and if f is steep, which means that for any sequence θ_n which converges to a boundary point of $\mathcal{D}$, then $\lim_{n\to\infty} |f'(\theta_n)| = +\infty$. f is *lower-semicontinuous* if its level sets $\{x : f(x) \le \alpha\}$ are closed for $\alpha \in \mathbb{R}$.

Remark 3.5. It is straight forward to show that Λ is convex and $\Lambda'(0) = \lambda$ (see [29, Lemma 1.11]).

Note that λ describes how close the average bit-arrival rate is to the asymptotic approximation of the ergodic capacity of the channel. For stability purpose and to ensure the existence of a stationary distribution, we require that $\lambda < r_{\max}$. Also, note that we abuse the notation and denote the arrival process by $A_t^{(N)}$, despite its possible dependency on the scaling function g.

Motivation for Smoothly-Scaling Assumption

The assumption of g-smoothly-scaling arrival processes allows us to find the decay rate of the tail probability of the sequence of process $(S_t^{(N)}, N \in \mathbb{N})$, which is a sum process defined as

$$S_t^{(N)} = \sum_{j=1}^{t} A_j^{(N)}, \quad t \in \mathbb{N};$$

since $(A_j^{(N)}, j \in \mathbb{Z})$ are i.i.d., $S_t^{(N)}$ is also a g-smoothly scaling process with the limiting g-scaled log moment generating function Λ_{S_t} given as

$$\Lambda_{S_t}(\theta) := \lim_{N \to \infty} \frac{\log E[\exp(\frac{\theta g(N)}{N} S_t^{(N)})]}{g(N)} = t\Lambda(\theta). \tag{3.7}$$

Now, given that the sequence $(S_t^{(N)}, N \in \mathbb{N})$ is g-smoothly-scaling, we can use the Gärtner-Ellis theorem (see e.g., [18] and [29]) to give the following result on the decay rate of the tail probability of the sequence. The following proposition provides an important basis for the analysis of the asymptotic probability of delay violation in Section 3.2.

Proposition 3.6. (Gärtner-Ellis theorem for g-smoothly-scaling process) *Consider $g \in \mathcal{G}$ and a family of g-smoothly-scaling processes $(A^{(N)}, N \in \mathbb{N})$ with the limiting g-scaled log moment generation function Λ. Let $S_t^{(N)} = \sum_{i=1}^{t} A_i^{(N)}$, for $t \in \mathbb{N}$. Then, for $a > \lambda t$, we have*

$$\lim_{N\to\infty} \frac{1}{g(N)} \log Pr\left(\frac{S_t^{(N)}}{N} > a\right) = -t\Lambda^*(a/t), \tag{3.8}$$

where Λ^* is the convex conjugate of Λ.

Proof. See Appendix B.1. □

Asymptotic Characteristic of Smoothly-Scaling Processes

Intuitively, the g-smoothly-scaling arrival processes become smoother as SNR increases. This intuition follows from (3.6), which implies that for $\theta \in \mathbb{R}$ such that $\Lambda(\theta) < \infty$ and $\epsilon > 0$, there exists N_0 such that for $N > N_0$,

$$\begin{aligned}\exp\left(g(N)\Lambda(\theta) - g(N)\epsilon\right) < E\left[\exp\left(\frac{\theta g(N)}{N} A_1^{(N)}\right)\right] \\ < \exp\left(g(N)\Lambda(\theta) + g(N)\epsilon\right).\end{aligned}$$

Then, if we let $Y_{g(N)}$ be a sum of $g(N)$ i.i.d. random variables (i.e., $Y_{g(N)} := X_1 + \cdots + X_{g(N)}$ with $E[e^{\theta X_1}] = e^{\Lambda(\theta)}$), we have $E[e^{\theta(Y_{g(N)})}] = e^{\Lambda(\theta)g(N)}$. Therefore, at sufficiently large N, $\frac{g(N)}{N} A_t^{(N)}$ and $Y_{g(N)}$ have the same moment generating function and hence the same distribution. If we define the burstiness of the random variable $A_1^{(N)}$ as the (dimensionless) ratio of its standard deviation over its mean,[9] then,

[9] Note that the burstiness definition here is basically the normalized variation of the random variable around its typical value (its mean). A more familiar definition of traffic burstiness would involve how the traffic are correlated with time, i.e., a bursty source tends to have large bursts of arrivals in a short period of time. However, since we only consider the source which is i.i.d. over time, we use this definition of burstiness.

using the above intuition, the burstiness $\frac{\mathrm{std}(A_1^{(N)})}{E[A_1^{(N)}]}$ for large N is approximately equal to $\frac{\mathrm{std}\left(\frac{N}{g(N)}\sum_{i=1}^{g(N)} X_i\right)}{\lambda N}$, which is reduced to $\frac{\mathrm{std}(X_1)}{\lambda\sqrt{g(N)}}$. Hence, the burstiness of $A_1^{(N)}$ decays to zero approximately as $\frac{1}{\sqrt{g(\log\rho)}}$. In other words, the g-smoothly-scaling arrival processes become smoother as SNR increases.

Examples of Smoothly-Scaling Processes

One of the common arrival processes used for traffic modeling is a compound Poisson process with exponential packet size, denoted as CPE. For this source, the random number of bits, $A_t^{(N)}$, arrived at timeslot t, is i.i.d. across time t and is in the form of

$$A_t^{(N)} = \sum_{i=1}^{M_t^{(N)}} Y_{i,t}^{(N)}, \tag{3.9}$$

where $M_t^{(N)}$ is the random variable corresponding to the number of packets that have arrived at the t^{th} timeslot, and where $Y_{i,t}^{(N)}$ corresponds to the random number of bits in the i^{th} packet. $M_t^{(N)}$ are independently drawn from a Poisson distribution with mean $\nu(N)$; and $Y_{i,t}^{(N)}$, $i = 1, \ldots, M_t^{(N)}$, are independently drawn from an exponential distribution with mean $\frac{1}{\mu(N)}$ (nats per packet). Note that the assumption that $E[A_t^{(N)}] = \lambda N$ forces that $\frac{\nu(N)}{\mu(N)} = \lambda N$. In addition, a larger average packet size $\frac{1}{\mu(N)}$ implies a more bursty arrival process.[10] It is known (see [44]) that

[10] It can be easily shown that the burstiness of this CPE process, as defined in Section 3.1.2, is $\sqrt{\frac{2}{\lambda N \mu(N)}}$.

the log moment generating function of this CPE random variable $A_t^{(N)}$ is

$$\log E[e^{\theta A_t^{(N)}}] = \begin{cases} \frac{\theta \nu(N)}{\mu(N)-\theta}, & \theta < \mu(N), \\ \\ \infty, & \text{otherwise.} \end{cases} \tag{3.10}$$

The following examples illustrate that, depending on the scaling of the average packet arrival rate and the average packet size, some CPE processes may or may not be g-smoothly-scaling.

Example 3.7 (g-smoothly-scaling CPE process)**.** For $g \in \mathcal{G}$ and $\mu > 0$, consider a CPE process $A_t^{(N)}$ with packet arrival rate $\mu\lambda g(N)$ and average packet size $\frac{N}{\mu g(N)}$. This family of processes is g-smoothly-scaling because, using (3.10), we have

$$\Lambda(\theta) := \lim_{N\to\infty} \frac{\log E[e^{\frac{\theta g(N)}{N} A_1^{(N)}}]}{g(N)} = \begin{cases} \frac{\mu\lambda\theta}{\mu-\theta}, & \theta < \mu, \\ \\ \infty, & \text{otherwise,} \end{cases} \tag{3.11}$$

which satisfies the conditions in the definition of g-smoothly-scaling. Since we will use this particular g-smoothly-scaling CPE process for examples in the chapter, we denote it as CPE(λ, μ, g, N). It is useful to note a particular case when $g(N)$ grows linearly with N. Using a property of the Poisson process [79], this particular scaling case can be viewed as aggregating an increasing number of Poisson traffic streams (this number grows linearly with N), with each stream having the same packet length distribution.

To complete our discussion on smoothly-scaling processes, we give an example below of a family of CPE arrival processes which is not g-smoothly-scaling.

Example 3.8. A family of CPE processes where $A_t^{(N)}$ has packet arrival rate $\mu\lambda$ and average packet size N/μ (note the dependence on N only in the average packet size) is not g-smoothly-scaling for any $g \in \mathcal{G}$. This is because, using (3.10), we have

$$\lim_{N\to\infty} \frac{\log E[e^{\frac{\theta g(N)}{N} A_t^{(N)}}]}{g(N)} = \begin{cases} 0, & \theta \leq 0, \\ \infty & \text{otherwise}, \end{cases}$$

which is not finite in the (open) neighborhood of $\theta = 0$. Hence, this family of processes is not g-smoothly-scaling.

Remark 3.9. The scaling function, g, describes the way the source statistics scale with SNR. Example 3.7 describes the case of the compound Poisson process, where g can be identified as the function that specifies how the average packet arrival rate ($\mu\lambda g(\log \text{SNR})$) and the average packet size ($\frac{\log \text{SNR}}{\mu g(\log \text{SNR})}$) scale with SNR.

3.1.3 Performance Measure and System Objective

The overall performance measure is the total probability of bit loss, $P_{\text{tot}}(r, T)$, where loss can occur due to channel decoding error or the end-to-end delay violation. Specifically,

$$P_{\text{tot}}(r, T) := P_{\text{ch}}(r, T) + (1 - P_{\text{ch}}(r, T))\, P_{\text{delay}}(r, T), \tag{3.12}$$

where $P_{\text{ch}}(r, T)$ denotes the probability of decoding error due to channel outage and $P_{\text{delay}}(r, T)$ denotes the probability of delay violation. We are interested in finding the high-SNR asymptotic approximation of $P_{\text{tot}}(r, T)$ as a function of r, T, SNR, D, as well as the source and channel statistics (including λ and the source scaling function g). In the interest of brevity, we denote P_{tot} as a function of only r and T, the two parameters over which the performance will later be optimized.

Since the high-SNR asymptotic expression of $P_{\text{ch}}(r, T)$ is already given by the DMT in (3.2), what remains is to find the asymptotic expression for $P_{\text{delay}}(r, T)$, which is shown in the next section.

3.2 Asymptotic Analysis of Probability of Delay Violation

In this section, we derive the asymptotic probability of delay violation $P_{\text{delay}}(r, T)$ for the channel multiplexing rate r and coding block size T. We observe that the adopted block coding forces the queue to have a *batch service* that occurs every T timeslots with the instantaneous removal of the oldest rNT bits. The decay rate of the asymptotic tail probability of the sum arrival process, given in Proposition 3.6, in conjunction with an asymptotic analysis of a queue with deterministic batch service, gives the following result:

Lemma 3.10. *Given $g \in \mathcal{G}$, $T \in \mathbb{T}$, $r > \lambda$, a batch service of rNT every T timeslots, and a g-smoothly-scaling bit-arrival process characterized by the limiting g-scaled log moment generation function Λ, the decay rate of $P_{delay}(r, T)$ is given* *by the function I, i.e.,*

$$\lim_{N\to\infty}\frac{1}{g(N)}\log P_{delay}(r,T) = -I(r,T), \tag{3.13}$$

where

$$I(r,T) = \min_{\substack{t\in\mathbb{Z}^+:\\ tT+T-1-k>0}} (tT+T-1-k)\Lambda^*\left(r+\frac{(D+1-2T)r}{tT+T-1-k}\right), \tag{3.14}$$

for $k = D(\bmod T)$. In addition, $I(r,T)$ is lower-semicontinuous and increasing on r.

Proof. See Appendix B.2. □

Approximation 3.11. *Relaxing the integer constraint in (3.14) gives the lower bound of I as*

$$I(r,T) \geq \delta_r r(D+1-2T) =: I_{ir}(r,T), \tag{3.15}$$

where

$$\delta_r = \sup\{\theta > 0 : \Lambda(\theta) < \theta r\}. \tag{3.16}$$

We use this lower bound as an approximation to I as well, i.e.,

$$I(r,T) \approx I_{ir}(r,T) = \delta_r r(D+1-2T). \tag{3.17}$$

Proof. See Appendix B.2. □

Example 3.12. For a g-smoothly-scaling CPE(λ,μ,g,N) bit-arrival process, the function I in (3.14) can be calculated exactly with the following Λ^*:

$$\Lambda^*(x) = \mu\left(\sqrt{x}-\sqrt{\lambda}\right)^2, \quad x\in\mathbb{R}. \tag{3.18}$$

However, an approximation of I in (3.17) is simpler to work with and given as

$$I(r,T) \approx I_{ir}(r,T) = \mu(r-\lambda)(D+1-2T), \tag{3.19}$$

where, using (3.16) and (3.11), δ_r is given as

$$\delta_r = \mu\left(1 - \frac{\lambda}{r}\right). \tag{3.20}$$

We will see via numerical examples in Section 3.4.1 that the approximation in (3.19) is sufficient for our purpose.

3.3 Main Result: Optimal Asymptotic Total Probability of Error

In this section, we present the main result of the chapter which states the optimal decay rate of the high-SNR asymptotic total probability of bit error. Recall the definition of P_{tot} from (3.12):

$$P_{\text{tot}}(r,T) := P_{\text{ch}}(r,T) + (1 - P_{\text{ch}}(r,T))P_{\text{delay}}(r,T),$$

where we now know that

$$P_{\text{ch}}(r,T) \doteq \rho^{-d_{\text{ch}}(r,T)}$$

and

$$P_{\text{delay}}(r,T) \overset{g}{=} e^{-I(r,T)g(\log\rho)}.$$

Hence, the asymptotic optimal decay behavior of P_{tot} depends on the function g. The following theorem gives the main result of the chapter.

Theorem 3.1. *Consider $g \in \mathcal{G}$ and a g-smoothly-scaling bit-arrival process. The optimal rate of decay of the asymptotic probability of total bit error, maximized over all $r \in (\lambda, r_{max})$ and $T \in \mathbb{T}$, and the optimizing r^* and T^* are given, depending on the tail behavior of the function g, as follows:*

Case 1: If $\lim_{N\to\infty} \frac{g(N)}{N} = \gamma \in (0,\infty)$, then

$$d^* := \sup_{\substack{r\in(\lambda,r_{max})\\ T\in\mathbb{T}}} \lim_{\rho\to\infty} \frac{-\log P_{tot}(r,T)}{\log\rho} = d_{ch}(r^*,T^*) = \gamma I(r^*,T^*), \quad (3.21)$$

where

$$\begin{aligned} r^*(T) &:= \inf\{r \in (\lambda, r_{max}) : \gamma I(r,T) = d_{ch}(r,T)\} && (3.22)\\ T^* &= \arg\max_{T\in\mathbb{T}} I(r^*(T),T) && (3.23)\\ r^* &= r^*(T^*). && (3.24)\end{aligned}$$

Case 2: If $\lim_{N\to\infty} \frac{g(N)}{N} = 0$ and $\lim_{N\to\infty} \frac{g(N)}{\log N} = \infty$, then

$$\sup_{\substack{r\in(\lambda,r_{max}),\\ T\in\mathbb{T}}} \lim_{\rho\to\infty} \frac{-\log P_{tot}(r,T)}{g(\log\rho)} \le \max_{T\in\mathbb{T}} I(r_{max},T). \quad (3.25)$$

Case 3: If $\lim_{N\to\infty} \frac{g(N)}{N} = \infty$, then

$$\sup_{\substack{r\in(\lambda,r_{max}),\\ T\in\mathbb{T}}} \lim_{\rho\to\infty} \frac{-\log P_{tot}(r,T)}{\log\rho} \le d_{ch}\left(\lambda, \left\lfloor \frac{D}{2} \right\rfloor\right). \quad (3.26)$$

Proof. See Appendix B.3. □

Theorem 3.1 shows that the optimal decay behavior of the asymptotic total probability of error depends on the tail behavior of the function g. As discussed earlier, the burstiness of the g-smoothly-scaling arrival process scales down as $\Theta(\frac{1}{\sqrt{g(\log\rho)}})$. Below, we discuss each case of Theorem 3.1, with respect to the scaling of the source burstiness:

In Case 1, where the source burstiness scales down with $\Theta(\frac{1}{\sqrt{\log\rho}})$, both components of the probability of error decay exponentially with $\log\rho$. In this setting, one can optimize the choices of r and T to arrive at a non-trivial optimal decay rate d^*. The optimal r^* and T^* balance and minimize the decay rate in P_{ch}(r,T) and $P_{\mathrm{delay}}(r,T)$. Hence, for Case 1, the optimal asymptotic total probability of error decays as follows:

$$P_{\mathrm{tot}}(r^*,T^*) \doteq P_{\mathrm{delay}}(r^*,T^*) \doteq P_{\mathrm{ch}}(r^*,T^*) \doteq \rho^{-d^*}.$$

Note that d^* is nothing but the optimal *negative SNR exponent*.

In Case 2, where the source burstiness scales down slower than $\Theta(\frac{1}{\sqrt{\log\rho}})$ but faster than $\Theta(\frac{1}{\sqrt{\log\log\rho}})$, we have that $P_{\mathrm{tot}}(r,T)$ is asymptotically equal to $P_{\mathrm{delay}}(r,T)$ for all $r \in (\lambda, r_{\max})$ and $T \in \mathbb{T}$. In this case, the decay rate of $P_{\mathrm{tot}}(r,T)$ is equal to $I(r,T)$. In other words, the channel error (outage) probability is dominated by the delay violation probability and, hence, can be ignored.

Finally, in Case 3, when the source burstiness scales down faster than $\Theta(\frac{1}{\sqrt{\log\rho}})$, we have the opposite of Case 2. In Case 3, the delay violation probability is dominated by the channel error probability and, hence, can be ignored.

3.3.1 Approximation of the Optimal Negative SNR Exponent

For Case 1 in Theorem 3.1, we use the following approximation which is an immediate result of relaxing the integer-constrained optimizations of I and T^* to obtain approximated expressions with much simpler forms. These approximations become especially useful in Section 3.4.

Approximation 3.13. *Relaxing the integer constraints in the calculation of I (as in Approximation 3.11) and T^* in (3.23) gives the following "integer-relaxed" approximations for $d^*, r^*,$ and T^*:*

$$d^* \approx d^*_{ir} := d_{ch}(r^*_{ir}, T^*_{ir}), \tag{3.27}$$

$$T^* \approx T^*_{ir}, \ \textit{and } r^* \approx r^*_{ir},$$

where, for δ_r given in (3.16) and any $T \in \mathbb{T}$,

$$r^*_{ir}(T) := \min\{r \in (\lambda, r_{max}) : d_{ch}(r,T) = \gamma\delta_r r(D-2T+1)\}, \tag{3.28}$$

and

$$T^*_{ir} = \left[\min\left\{T \in \mathbb{R}^+ : \frac{d}{dT}\big(d_{ch}(r^*_{ir}(T),T)\big) = 0\right\}\right]_1^{\lfloor \frac{D}{2} \rfloor}, \tag{3.29}$$

$$r^*_{ir} = r^*_{ir}(T^*_{ir}). \tag{3.30}$$

3.4 Applications of the Result

In this section, we apply the result of Case 1 in Theorem 3.1 to analyze and optimize the end-to-end error probability of systems communicating delay-sensitive and bursty traffic over three outage-limited channels: SISO Rayleigh fast-fading channel, quasi-static cooperative relay channel, and quasi-static MIMO channel.

To illustrate the methodology, we restrict our attention to the case of CPE$(\lambda, \mu, g, \log \rho)$ arrival process where $g(\log \rho) = \log \rho$, for simplicity. Note that to better gain insights, we use the integer-relaxed approximations obtained in Approximation 3.13.

3.4.1 SISO Rayleigh Fast-Fading Channel

Our first example considers an example of SISO Rayleigh fast-fading channel, whose $d_{\text{ch}}(r, T) = T(1 - r)$ (see (3.4)). Combining this with (3.20) and (3.28) gives the optimal choice of multiplexing gain when the coding duration is fixed at T as

$$r^*_{ir}(T) = \lambda + \frac{1 - \lambda}{1 + \frac{\mu(D+1-2T)}{T}}. \tag{3.31}$$

In addition, using (3.29), the integer-relaxed approximated optimal coding duration can be expressed as

$$T^*_{ir} = \left[\frac{1}{1 + \frac{1}{\sqrt{2\mu}}} \frac{D+1}{2} \right]_1^{\lfloor \frac{D}{2} \rfloor}. \tag{3.32}$$

Inserting T^*_{ir} into (3.31), we get the approximated optimal channel multiplexing gain as

$$r^*_{ir} = r^*_{ir}(T^*_{ir}) = \left[\lambda + \frac{1 - \lambda}{1 + \sqrt{2\mu}} \right]_{r^*_{ir}(1)}^{r^*_{ir}(\lfloor \frac{D}{2} \rfloor)}. \tag{3.33}$$

Also, from (3.27), the approximated optimal negative SNR exponent is given as:

$$d^*_{ir} = T^*_{ir}(1 - r^*_{ir}) = \left[\frac{1}{(1+\frac{1}{\sqrt{2\mu}})^2} \frac{(D+1)}{2}(1-\lambda) \right]_{1-r^*_{ir}(1)}^{\lfloor \frac{D}{2} \rfloor (1-r^*_{ir}(\lfloor \frac{D}{2} \rfloor))} . \qquad (3.34)$$

Below, we provide some observations of the above results:

- The above result on d^* can also be interpreted as a tradeoff which describes the relation between the normalized average arrival rate,

$$\lambda := \lim_{N\to\infty} \big(\text{average bit-arrival rate}\big)/N = \lim_{N\to\infty} \frac{\mathbb{E}[A_t^N]}{N},$$

and the corresponding optimal negative SNR exponent $d^*_{ir}(\lambda)$ as a function of the delay bound D, and the average packet size $1/\mu$. For constant bit arrivals (CBR) at rate $\lambda \log \rho$, i.e., mathematically when $1/\mu \to 0$, any coding durations less than half[11] of D (or more precisely $\lfloor \frac{D}{2} \rfloor$) and any channel multiplexing rates greater than λ result in zero probability of delay violation. Hence, the optimal negative SNR exponent of the total error probability, denoted by d^*_{CBR}, is equal to the corresponding channel diversity when the optimal coding duration is at its maximum value, $\lfloor \frac{D}{2} \rfloor$, and the channel multiplexing gain is at its minimum, λ. That is,

$$d^*_{CBR}(\lambda) = \left\lfloor \frac{D}{2} \right\rfloor (1-\lambda).$$

It is not surprising that this coincides with the classical DMT. With traffic burstiness, however, the optimal negative SNR exponent $d^*_{ir}(\lambda)$ given in (3.34) is smaller than $d^*_{CBR}(\lambda)$. The ratio

$$\frac{d^*_{ir}(\lambda)}{d^*_{CBR}(\lambda)} \approx \frac{1}{(1+\frac{1}{\sqrt{2\mu}})^2} \leq 1$$

[11] The first half of D is spent waiting for the next coding block and the other half waiting to be decoded at the end of the block.

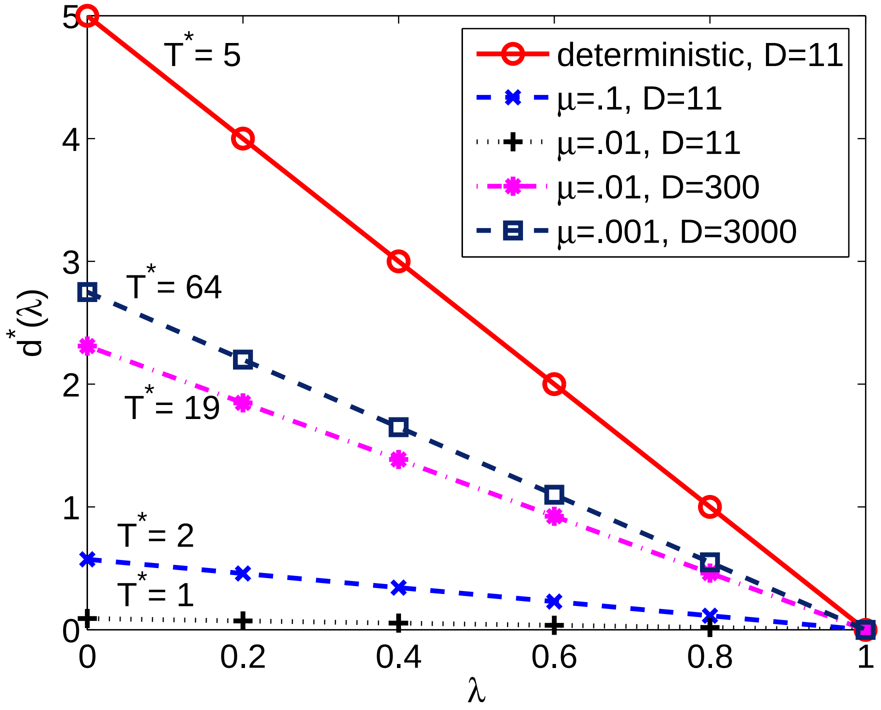


Figure 3.2 Optimal negative SNR exponent for SISO, Rayleigh fast-fading channel. The solid line describes the DMT ($r = \lambda$). The dashed and dotted lines describe $d^*_{ir}(\lambda)$ for various μ and D.

can be interpreted as the reduction factor on the SNR exponent in the presence of burstiness. Figure 3.2 shows the impact of traffic burstiness (which is parameterized by μ) on $d^*_{ir}(\lambda)$.

- From a coding point of view, T^*_{ir} is independent of the average bit-arrival rate λ. This implies that for a fixed value of the average packet size $1/\mu$, the optimal negative SNR exponent is achieved by a fixed-duration $1 \times T^*_{ir}$ code. Optimal codes for this setting exist for all values of r and T ([24, 77, 86]). On the other hand, if T is already given, the performance is optimized when the coding multiplexing gain is chosen as in (3.31), i.e.,

$$r_{ir}^{*}(T) = \lambda + \frac{1-\lambda}{1+\frac{\mu(D+1-2T)}{T}}.$$

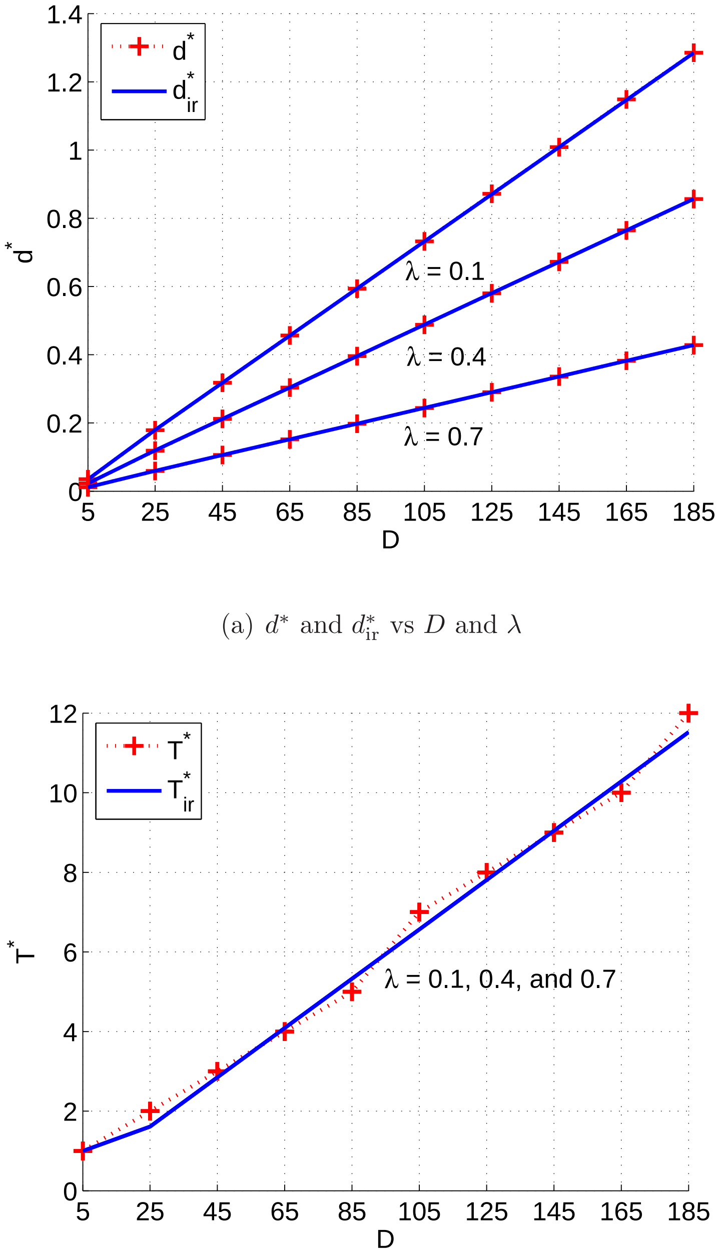
- Since $r_{\max} = 1$ for this SISO channel, we can verify that $r_{ir}^{*} \nearrow r_{\max}$ for very bursty traffic (i.e., $1/\mu \to \infty$). That is for very bursty traffic the channel should operate close to its highest possible rate, which is the channel ergodic capacity.

Numerical Comparison of the Approximation

Before we move to the next example, we illustrate numerically that the approximations in (3.32)-(3.34) well approximate their actual values in Theorem 3.1. In Figure 3.3, we show an example of a comparison at $1/\mu = 100$ and various values of D and λ. We observe that the approximated values match well with the exact values if D is sufficiently large. The matching is remarkably good for d^{*} and d_{ir}^{*}. Note that r_{ir}^{*} is independent of D, except when D is so small that $T_{ir}^{*} = 1$.

3.4.2 Cooperative Wireless Networking with Optimal Clustering

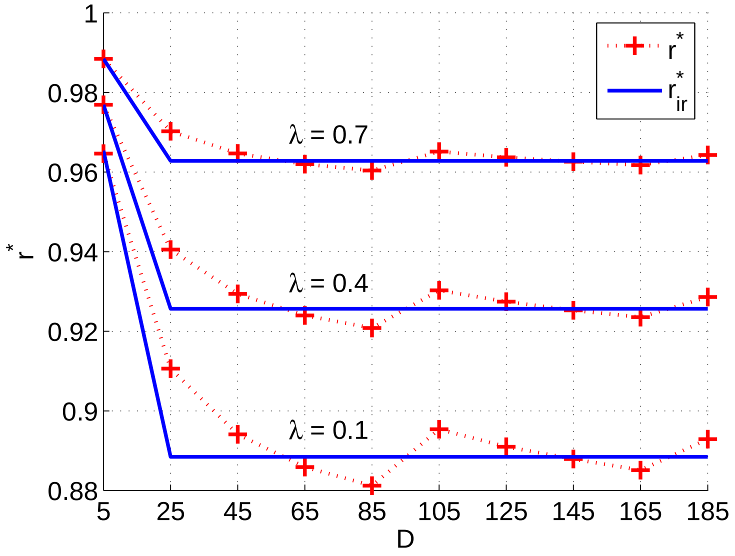
As studied in [25, 46], we consider communicating bursty and delay-limited information from an information source in a cooperative wireless relay network, shown in Figure 3.4, where the diversity benefit of user cooperation is due to encoding across space and time [52, 66]. In the absence of delay limitation,



(a) d^* and d^*_{ir} vs D and λ

(b) T^* and T^*_{ir} vs D and λ

Figure 3.3 Comparisons of the exact values d^*, T^*, and r^* and the integer-relaxed approximations d^*_{ir}, T^*_{ir}, and r^*_{ir}, at various D and λ. The dotted lines with markers correspond to the exact solutions while the solid lines represent the approximated solutions.



(c) r^* and r_{ir}^* vs D and λ

Figure 3.3 Comparisons of the exact values, continued.

having more cooperative users almost always improves performance. This is not the case, though, when one considers burstiness and delay QoS requirement. Take for example a network where the information-source node cooperates with v relays, under an orthogonal amplify-and-forward (OAF) cooperative diversity scheme and half-duplex constraint. This cooperation scheme gives the DMT:

$$d_{\text{ch}}^{\text{coop}}(r) = (v+1)(1-2r).$$

Note that $r_{\max} = 1/2$ under this protocol. To realize this amount of diversity, the coding duration T is required to be at least $2(v+1)$ channel uses or timeslots. This means that, in spite of the increase in the negative SNR exponent of the probability of decoding error with the number of cooperative relays, relaying over all nodes in the network might not be desirable as it increases the delay violations. Applying the result of Approximation 3.13 to CPE source and the above $d_{\text{ch}}^{\text{coop}}(r)$ with $T = 2(v+1)$, the optimal performance is achieved when the nodes cooperate in clusters with

$$v^* \approx v_{ir}^* = \left[\frac{D+1}{4(1+\frac{1}{\sqrt{2\mu}})} - 1 \right]_1^v$$

relays and transmit at multiplexing rate,

$$r^* \approx r_{ir}^* = \frac{1}{2} - \frac{\frac{1}{2} - \lambda}{1 + \frac{1}{\sqrt{2\mu}}}.$$

Note that v_{ir}^* is independent of the traffic average arrival rate λ. This means that meeting the optimal tradeoff for various values of λ does not require modifying the cluster sizes, unless the traffic burstiness (parameterized by the average packet size $1/\mu$) changes.

3.4.3 MIMO Quasi-Static Communications

In the case of the MIMO Rayleigh fading channel with n_t transmit and n_r receive antennas, and with complete channel state information at the receiver (CSIR) and no CSI at the transmitter, the channel diversity gain $d_{\text{ch}}(r)$ is shown (see [90]) to be a piecewise linear function that connects points

$$(k, (n_t - k)(n_r - k)), \quad k = 0, 1, \ldots, \min(n_t, n_r). \tag{3.35}$$

The entire tradeoff is met if $T \geq n_t$ [27]. An example of the effect of burstiness is shown in Figure 3.5, for the case of the 2×2 Rayleigh fading channel ($n_t = n_r = 2$). By assuming that T is given (not an optimizing parameter) and equal to 2, the

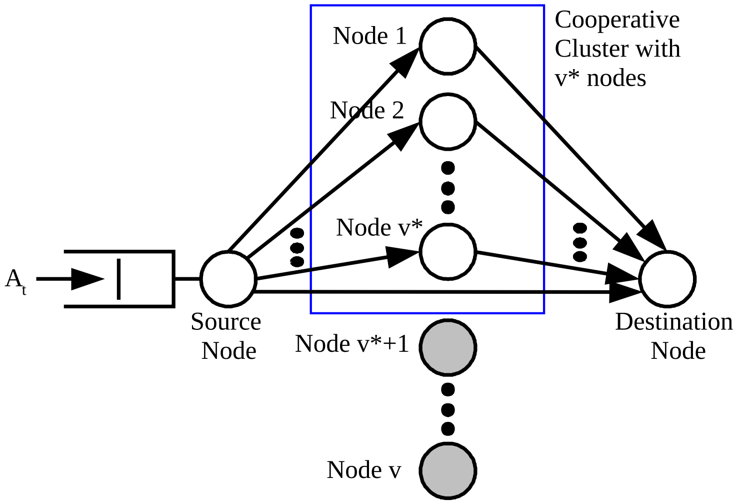


Figure 3.4 Snapshot of a cooperative relay wireless network, where the source node utilizes a subset of its peers (nodes $1, 2, \ldots, v^*$) as relays for communicating with the destination.

optimal multiplexing gain r^*, which balances the SNR exponents of the probabilities of delay violation and decoding error, is the solution to $d_{\text{ch}}(r^*) = I(r^*, T = 2)$. Using the approximation (3.19) of I for CPE source, the approximation r^*_{ir} is the solution to

$$d_{\text{ch}}(r^*_{ir}) - \mu(r^*_{ir} - \lambda)(D - 3) = 0,$$

where d_{ch} is the piecewise linear function connecting points in (3.35). In other words, r^*_{ir} is given as

$$r^*_{ir} = \begin{cases} \lambda + \frac{2-\lambda}{1+\mu(D-3)}, & \text{if } \lambda \in [1 - \frac{1}{\mu(D-3)}, 2), \\ \lambda + \frac{4-3\lambda}{3+\mu(D-3)}, & \text{if } \lambda \in (0, 1 - \frac{1}{\mu(D-3)}]. \end{cases}$$

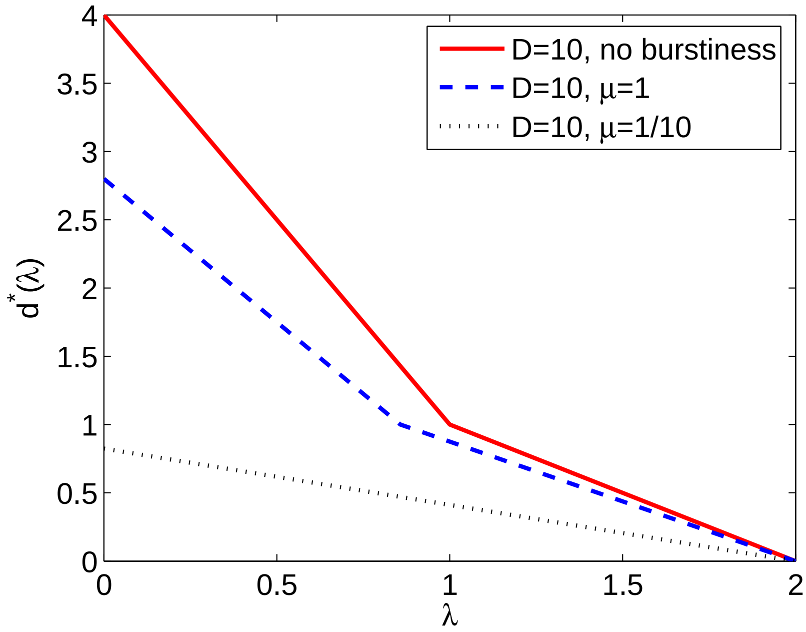


Figure 3.5 MIMO, quasi-static, coherent 2x2 channel. d^* v.s. λ for different values of D and μ.

Figure 3.5 shows the resulting $d^*(\lambda) = d_{\text{ch}}(r^*_{ir})$ for various values of burstiness μ and D.

3.5 Summary

This chapter offers a high-SNR asymptotic error performance analysis for communications of delay-limited and bursty information over an outage-limited channel, where errors occur either due to delay or due to erroneous decoding. The analysis focuses on the case where there is no CSIT and no feedback, and on the static case of fixed rate and fixed length of coding blocks. This joint queue-channel analysis is performed in the asymptotic regime of high-SNR and in the assumption of smoothly scaling (with SNR) bit-arrival processes. The analysis provides closed-form expressions for the error performance, as a function of the channel and source statistics. These expressions identify the scaling regime of the source and channel statistics in which either delay or decoding errors are the dominant cause of errors, and the scaling regime in which a prudent choice of the coding duration and rate manages to balance and minimize these errors. That is, in this latter regime, such optimal choice manages to balance the effect of channel atypicality and burstiness atypicality. To illustrate the results, we provide different examples that apply the results in different communication settings. We emphasize that the results hold for any coding duration and delay bound.

Acknowledgment

This chapter, in full, appears in the following publication. The dissertation author was the primary investigator and author of this paper.

- S. Kittipiyakul, P. Elia, and T. Javidi, “High-SNR analysis of outage-limited communications with bursty and delay-limited information,” is accepted for publication in *IEEE Trans. Inf. Theory.*.

I would like to thank my coauthors Prof. Petros Elia and Prof. Tara Javidi.

Chapter 4

Optimal Operating Point for MIMO Multiple Access Channel With Bursty Traffic

Abstract

Multiple antennas at the transmitters and receivers in a multiple access channel (MAC) can provide simultaneous diversity, spatial multiplexing, and space-division multiple access gains. The fundamental tradeoff in the asymptotically large SNR regime is shown by Tse et al. [78]. On the other hand, MAC scheduling can provide a statistical-multiplexing gain to improve the delay performance as shown by Bertsimas et al. [9] and Stolyar and Ramanan [72]. In this chapter, we formulate and analytically derive bounds on the optimal operating point and the asymptotic (high-SNR and large delay bound D) error performance

of MIMO-MAC channel for bursty sources with delay constraints. Our system model brings together the four types of gains: diversity, spatial multiplexing, space-division multiple-access, and statistical-multiplexing gains. As in Chapter 3, our objective is to minimize the end-to-end performance as defined by the delay bound violation probability as well as the channel decoding error probability. We find the optimal diversity gain and rate region in which the system should operate. As an example, we illustrate our technique and the optimal operating point for the case of a compound Poisson source. In addition, we note an interesting interplay between the intensity of the traffic and resource pooling with regard to both multiple-access and statistical-multiplexing gains.

Multiple antennas can be used to enhance the performance of wireless systems. The multiple antennas can be used to simultaneously boost the reliability (providing *diversity* gain) and the data rate (providing *spatial multiplexing* gain). In addition, in multiple access scenarios where multiple users are transmitting to a common receiver, multiple receive antennas also provide *multiple-access* gain by allowing for spatial separation of the signals of different users. Tse, Viswanath, and Zheng [78] have characterized the fundamental tradeoff between these three types of gains at high SNR.

In this chapter, we consider a system where each user has a bursty source concatenated with an infinite buffer and a MIMO multiple access channel (MIMO-MAC). The end-to-end performance metric of interest is the total bit error probability, where bit errors can be due to either delay violation or decoding errors in the MIMO-MAC channel. From a user's perspective, we face the following tradeoff: the higher the multiplexing gain the better the delay performance, but the inevitable decrease in diversity results in an increase in MIMO channel errors. At the same time, the statistical variation in the traffic patterns among users provides us with flexibility in allocating the resources.

Hence, in this chapter, we study the similar cross-layer queue-channel optimization problem studied in Chapter 3, but in a multi-user context where the dynamic scheduling of the resources can improve the delay violation probability. Like Chapter 3, the analysis is performed in the asymptotic regime of high-SNR. However, unlike Chapter 3, we assume large delay bound D in the analysis so that we can utilize existing results in [9, 72].

The main contribution of this chapter is the formulation of a cross-layer optimal operating point for a MIMO-MAC channel with bursty sources and delay constraints. In particular, we provide a methodology for characterizing the optimal diversity gain and rate region in which the system should operate in a MIMO-MAC channel with a given high SNR and description of the bursty traffic sources. To achieve this, we assume an optimal scheduler design which dynamically controls users' transmission rates (or equivalently, the multiplexing gains) as a function of queue backlogs. This dynamic adaptation of multiplexing gains accounts for *statistical-multiplexing* while leveraging the known tradeoff between diversity, spatial multiplexing, and multiple-access gains given in [78]. From a scheduling perspective, statistical-multiplexing is a key mechanism by which the network resources are used to improve the delay performance for bursty users. In particular, statistical-multiplexing capitalizes on the fact that peaks in traffic of simultaneously ongoing traffic streams rarely coincide. We believe that our result can be viewed as a first step in integrating the known spatial diversity and multiplexing and multiple-access gains with that of the statistical-multiplexing. In other words, for the first time, our model brings together the four types of gains offered at a MIMO MAC.

The remainder of the chapter is organized as follows. In Section 4.1, we provide a detailed description of the system model as well as the problem formulation for general number of users. In Section 4.2, we provide the main analytical results and bounds on the optimal channel diversity gain. We also discuss the notion of *statistical-multiplexing* and its benefits. In Section 4.3, we find the optimal operating diversity gain d^* or its bounds for a particular class of compound Poisson sources. Section 4.4 summarizes the chapter.

4.1 System Model

We consider the architecture shown in Figure 4.1. The system is time-slotted and consists of three main components, each shown with a different number. The first component consists of K homogeneous users, each of which has an identical but independent bursty source: each source generates information bits according to a stochastic process. When appropriate, the bits are buffered prior to

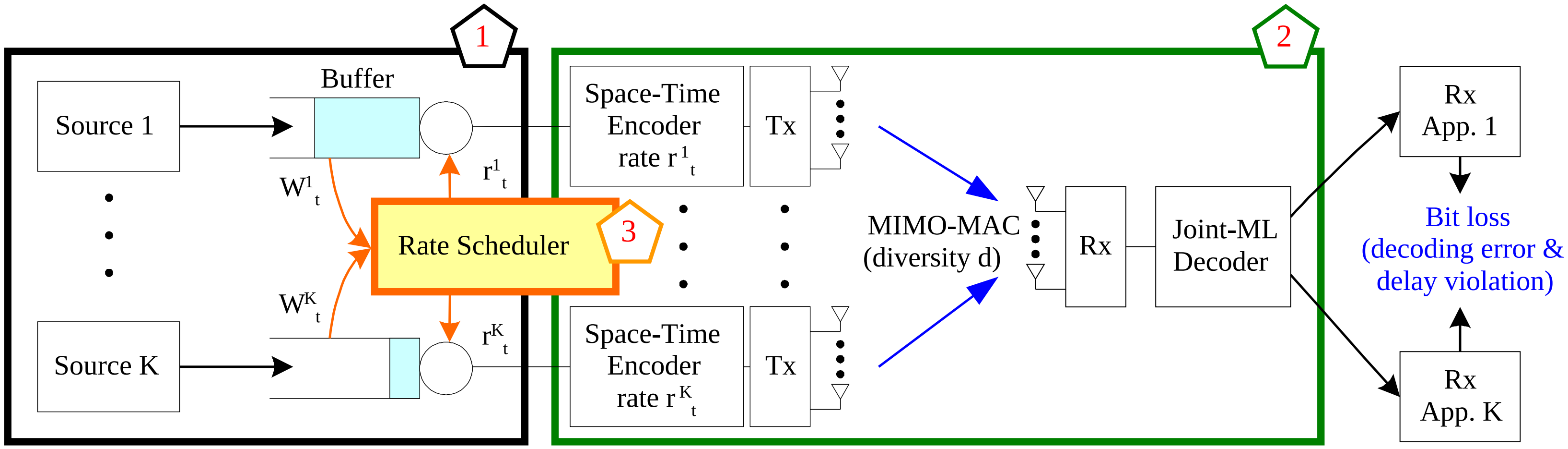


Figure 4.1 System model and the two causes of bit loss: delay violation and channel decoding error.

transmission over the channel. The second component of interest is a MIMO multiple access channel without channel state information (CSI) at the transmitters but with perfect CSI at the receiver. The receiver consists of a joint maximum-likelihood decoder. In the absence of CSI at the transmitters, we assume that the MIMO-MAC operates at a common diversity gain, which in turn specifies the corresponding capacity region of the MIMO MAC channel as given in [78]. However, the individual rate of each user is determined dynamically by the rate scheduler which is the third component in our system. This is a centralized rate scheduler that dynamically determines the transmission rates of the individual users given queue state information (QSI) of each user.

We assume no retransmission of the bits in error and map our objective to the sum of the probability of delay violation and the probability of channel error. In order to mathematically define this problem, we now model each of the above components precisely.

4.1.1 Source Model

We assume that the total number of information bits generated by user i ($i = 1, \ldots, K$) is given by a sequence $S^i = \{S^i_t, t = 1, 2, \ldots\}$, where S^i_t is the total number of bits of user i generated up to timeslot t and $S^i_0 \equiv 0$. In addition, we assume that the arrival processes S^i, $i = 1, \ldots, K$, are identical and mutually independent. We also assume that each arrival process S^i has stationary increments and satisfies a *Large Deviations Principle* (LDP) defined in Definition 4.1. In the appendix, we discuss an additional sample path LDP assumption (Assumption B) on the arrival processes. Here, to keep the flow of the chapter, in this section we only provide the LDP assumption and the consequent characterization of the sources which is based on LDP.

In general, consider a source process S generating a sequence $(S_t, t \in \mathbb{Z}^+)$ of random variables, where S_t is the total number of bits generated up to timeslot t. The following definition gives a rough definition of the LDP, suitable for our purpose in this chapter.

Definition 4.1. A source S is said to satisfy an LDP with *rate function*[1] $\Lambda^* : \mathbb{R} \to [0, \infty]$ if, for large enough t and for small $\epsilon > 0$,

$$\Pr\left[\frac{S_t}{t} \in (a - \epsilon, a + \epsilon)\right] \approx e^{-t\Lambda^*(a)} \tag{4.1}$$

where Λ^* is a lower semicontinuous function and has compact level sets[2] (see [9] and [18] for more discussions in the LDP).

Fact 4.2 (Gärtner-Ellis theorem). *Suppose a source S satisfies the following:*

Assumption 4.2.A*:*

1. *The limiting log-moment generating function*

$$\Lambda(\theta) := \lim_{t\to\infty} \frac{1}{t} \log \mathbb{E}[e^{\theta S_t}] \tag{4.2}$$

exists for all θ, where $\pm\infty$ are allowed both as elements of the sequence and as limit points.

[1] In large deviations literature, the Λ^* function is typically called a "rate" function. Here we use the same name, but caution the reader not to confuse with the transmission rate.

[2] The level set $\{x : \Lambda^*(x) \le a\}$ is compact for every real a

2. *The origin is in the interior of the domain* $D_\Lambda := \{\theta | \Lambda(\theta) < \infty\}$ *of* $\Lambda(\theta)$.

3. Λ *is* essentially smooth, *i.e.,* $\Lambda(\theta)$ *is differentiable in the interior of* D_Λ *and the derivative tends to infinity as* θ *approaches the boundary of* D_Λ.

4. $\Lambda(\theta)$ *is lower semicontinuous, i.e.* $\liminf_{\theta_n \to \theta} \Lambda(\theta_n) \geq \Lambda(\theta)$ *for all* θ.

Then the source S *satisfies an LDP and its decay function described by (4.1) is given as*

$$\Lambda^*(a) = \sup_{\theta} \, [\theta a - \Lambda(\theta)]. \tag{4.3}$$

Remark 4.3. It can be shown that Λ^* is a convex function taking values in $[0, \infty]$ such that $\Lambda^*(E[S_1]) = 0$ where $E[S_1]$ is the average arrival rate of process S (see [18]).

Remark 4.4. Many source models commonly used to model bursty traffic in communication networks satisfy Assumptions A and B. Such source models include renewal processes, Markov-modulated processes, and more generally stationary processes with mild mixing conditions [9].

4.1.2 MIMO-MAC Channel Model and PHY Model

We use the same symmetric MIMO multiple access channel model as described in [78] which assumes symmetric transmitters seeing i.i.d. fading channels, perfect symbol synchronization and perfect CSI at the receiver but no CSI at any transmitters. Each transmitter has n_t transmit antennas, while the receiver has

n_r receive antennas. Space-time coding happens over a channel coherence time which is assumed to contain T symbols[3]. We assume the duration of a timeslot is equal to the channel coherence period, i.e., a timeslot contains T symbols. Since the transmitters are assumed to know only the channel statistics, including the average received SNR, they always transmit at the maximum powers which are equal for all transmitters. The channel fading processes of the transmitters are assumed to be stationary over time, mutually independent, and identical. For each transmitter, the channel fadings for different antenna paths are assumed to be slow block-fading with i.i.d. Rayleigh fading where the fadings stay constant during a timeslot and change independently and simultaneously over timeslots. We denote by ρ the average received signal-to-noise ratio (SNR) at each receive antenna.

From a system perspective, at each SNR level ρ, the PHY layer for the MIMO-MAC channel provides a tradeoff between the reliability of the transmissions and the transmission rates. Equivalently, we can say that the PHY layer provides a tradeoff between a common diversity d and the multiplexing gain region, denoted by $\mathcal{R}(d)$, where d and $\mathcal{R}(d)$ are as defined in [90] and [78]. We state these definitions below.

Definition 4.5. (Theorem 2 in [78]) Let r_t^i be the multiplexing gain of user i, $i = 1, \ldots, K$, at time t. Given a common diversity requirement d for all users, i.e.,

$$P_{\text{ch}}^i \dot{\leq} SNR^{-d}, \quad i = 1, \ldots, K, \tag{4.4}$$

[3] We assume either a sufficiently large symbol rate or a sufficiently small number of antennas such that $T \geq Kn_t + n_r - 1$.

where P_{ch}^i is the average error probability for user i. Then the spatial multiplexing gains $(r_t^1, \ldots, r_t^K)$ at any timeslot t must be within the (time-independent) multiplexing gain region

$$\mathcal{R}(d) = \left\{ (r^1, \ldots, r^K) : \sum_{s \in S} r^s \leq r^*_{|S|n_t, n_r}(d), \forall S \subseteq \{1, \ldots, K\} \right\}. \tag{4.5}$$

where $r^*_{m,n}(d)$ for any integers m and n is the largest multiplexing gain achieved for an $m \times n$ point-to-point MIMO link for a given diversity d and is defined as a piecewise linear function joining the points $((m-k)(n-k), k)$ for $k = 0, \ldots, \min(m, n)$.

In this chapter, we consider a system which always operates at a common diversity gain d at any time t. This d directly determines the multiplexing region $\mathcal{R}(d)$ and its shape. In particular, d determines the sum of all the rates at all time t, i.e. $\sum_{i=1}^K r_t^i \leq r^*_{Kn_t, n_r}(d)$, which is independent of time. However, the individual rate at time t, $r_t^i, i = 1, \ldots, K$, is determined dynamically by the rate scheduler discussed later.

In Figure 4.2, we illustrate the dependence of the shape of $\mathcal{R}(d)$ and d for a simple case of $K = 2$ users and $n_t = n_r = 2$. As seen in this figure, there exists a diversity gain d_0 (in this example, $d_0 = 2$) such that, for large d ($d > d_0$), the shape of $\mathcal{R}(d)$ follows a rectangular shape (*single-user performance* regime), while, for small d ($d < d_0$), $\mathcal{R}(d)$ is a polymatroid shape (*antenna-pooling* regime). Furthermore, [78] shows that d_0 is the unique solution to

$$r^*_{Kn_t, n_r}(d_0) = K r^*_{n_t, n_r}(d_0). \tag{4.6}$$

Later we will see the impact of this change of shape on the working of the scheduler

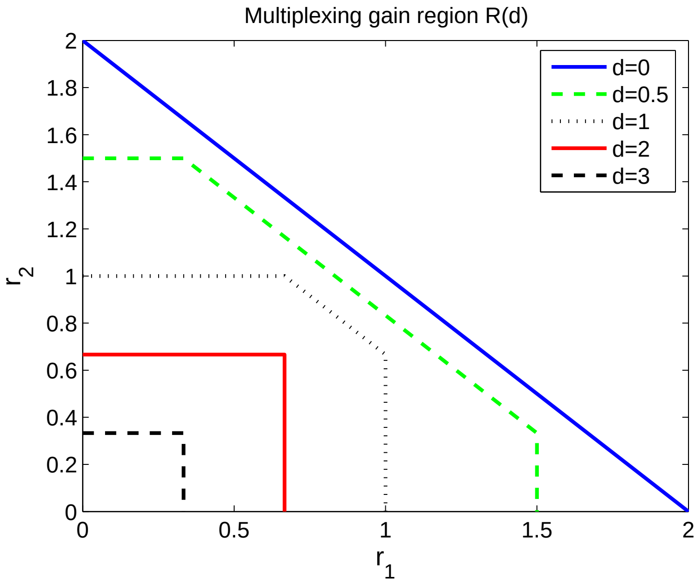


Figure 4.2 Example of the multiplexing gain region $\mathcal{R}(d)$ for $n_t = n_r = 2$ case.

block.

4.1.3 Rate Scheduler

Given that each user operates at a fixed and common diversity gain d and given an average SNR of ρ in the MIMO-MAC subsystem, the function of scheduler $H_d : \mathbb{R}_+^K \mapsto \mathcal{R}(d)$ is to allocate, at the beginning of every timeslot t, the set of feasible multiplexing gains to the users. This is done, equivalently, by selecting a vector of spatial multiplexing gains $(r_t^1, \ldots, r_t^K)$ from the multiplexing gain region $\mathcal{R}(d)$. The decision is based on the delay of the head-of-the-line bit in queue i, denoted by D_t^i, $i = 1, \ldots, K$, at the beginning of timeslot t. Specifically, we assume that

$$(r_t^1, \ldots, r_t^K) = H_d(D_t^1, \ldots, D_t^K).$$

Without loss of optimality, one can assume that the rate scheduler always assigns the highest possible sum rate. At timeslot t, an amount of $r_t^i T \log \rho$ bits are taken out from head-of-the-line of the buffer of user i. We assume that if any buffers do not have enough data to transmit, the null data is used to fulfill the rates.

Remark 4.6. Recall that the shape of the multiplexing gain $\mathcal{R}(d)$ depends on d (e.g. see Figure 4.2). As a result, the choice of diversity gain d determines the class of feasible dynamic schedulers. In the single-user performance regime ($d \geq d_0$), the users are decoupled and independent from one another, hence, reducing the scheduler to a static (and decoupled) choice of multiplexing gain $r_t^i = r^*_{n_t,n_r}(d)$ for all $i = 1, \ldots, K$ and all time $t \in \mathbb{Z}$. For the antenna-pooling regime ($d < d_0$), on the other hand, $\mathcal{R}(d)$ is a polymatroid. In other words, in this regime, the multiplexing gains of the users are dependent on one another and must be jointly allocated.

Remark 4.7. The model in this chapter assumes that there is no CSI available at the transmitters and the central scheduler. However, the scheduler has perfect knowledge of the queue state information (QSI). This is not unrealistic given the fact that it is less bandwidth consuming and more accurate to send the QSI of each buffer (an observable scalar number) to the centralized scheduler than to *estimate* CSI for MIMO channels (K matrices, each of dimension $n_t \times n_r$) at the receiver and feed back these matrices to the transmitters.

Remark 4.8. Due to lack of CSI, the role of the rate scheduler in this chapter is not to minimize the channel error performance; instead, the scheduler improves the delay violation probability by taking advantage of the statistical-multiplexing gain provided by the multiple bursty sources sharing the multiple access channel.

4.1.4 Arrival Rate Scaling and Stability Condition

Since the rates of transmission in the MIMO-MAC channel are scaled as $\log \rho$, we scale the arrival rates with $\log \rho$ as well. In other words, we assume that the average bit arrival rate $\tilde{\lambda}$ of each user is

$$\tilde{\lambda} = \lambda T \log \rho \tag{4.7}$$

bits per timeslot for a given constant positive λ.

In addition, to guarantee system stability, we require that the total average arrival rate to be no greater than the (sum) capacity of the MIMO-MAC channel [72]. In particular, we assume that

$$K\tilde{\lambda} < \min(Kn_t, n_r) T \log \rho,$$

or equivalently

$$\lambda < \min(n_t, n_r/K). \tag{4.8}$$

Since the system is stable, it reaches a steady state. We let D^i and Q^i denote the steady-state delay and queue length, respectively, for queue i, $i = 1, \ldots, K$.

For the rest of the chapter, we denote $r_{av}(d)$ as the average multiplexing gain at the common diversity d, defined as

$$r_{av}(d) := \begin{cases} r^*_{n_t,n_r}(d) & \text{if } d \geq d_0, \\ \frac{1}{K} r^*_{Kn_t,n_r}(d) & \text{if } d < d_0. \end{cases} \tag{4.9}$$

and denote

$$C_{av}(d) := r_{av}(d) T \log \rho \quad \text{(bits per timeslot)} \tag{4.10}$$

as the average per-queue channel capacity at diversity d.

4.1.5 Objective

Our system objective is to find the optimal operating channel diversity gain d^* and the corresponding multiplexing gain region $\mathcal{R}(d^*)$ in which the system should operate. This diversity d^* minimizes the end-to-end total error probability caused by two phenomenas: 1) delay violation of the delay bound D and 2) channel decoding error. [4]

In particular, we define the following probabilities:

$$\begin{aligned} P^i_{\text{ch}}(d) &:= \text{Pr[decoding error for user } i] \\ P_{\text{ch}}(d) &:= \text{Pr[decoding error for any user]} & (4.11) \\ P^i_{\text{delay}}(d) &:= \text{Pr[delay violation of user } i] = \text{Pr}[D^i > D] \\ P_{\text{delay}}(d) &:= \text{Pr[delay violation for any user]} = \text{Pr}[\max_{i=1,\ldots,K} D^i > D]. & (4.12) \end{aligned}$$

With the above definitions, the total error probability $P_{\text{tot}}(d)$ is expressed

[4] We assume no retransmission for the lost bits due to channel decoding errors or delay violation. Furthermore, the source processes are not effected by the lost bits.

as

$$\begin{aligned} P_{\text{tot}}(d) &:= \max_{i=1,\ldots,K} \Pr[\text{bit loss for user } i] \\ &= \max_{i=1,\ldots,K} \left\{ P_{\text{ch}}^{i}(d) + (1 - P_{\text{ch}}^{i}(d)) P_{\text{delay}}^{i}(d) \right\}, \end{aligned} \tag{4.13}$$

where $P_{\text{ch}}^{i}(d) + (1 - P_{\text{ch}}^{i}(d))P_{\text{delay}}^{i}(d)$ is the total error probability of user i due to channel and delay violation. We will later show that $P_{\text{ch}}^{i}(d) \doteq P_{\text{ch}}(d) \doteq \rho^{-d}$ and $P_{\text{delay}}^{i}(d) \doteq P_{\text{delay}}(d) \doteq \rho^{-f(d)}$ where f is some functional taking positive values. Hence, the asymptotic large-SNR expression of $P_{\text{tot}}(d)$ is given as

$$P_{\text{tot}}(d) \doteq P_{\text{ch}}(d) + P_{\text{delay}}(d) \doteq \rho^{-d} + \rho^{-f(d)}. \tag{4.14}$$

We note that both probabilities $P_{\text{ch}}(d)$ and $P_{\text{delay}}(d)$ are functions of the diversity gain d as well as the average SNR ρ. However, there is a tradeoff between the two probabilities as a function of d: Intuitively, for a fixed ρ, we expect that a high diversity gain, which translates into a smaller transmission rate region, results in faster queue build-up and larger delays. On the other hand, this higher diversity gain yields better channel performance.

In the remainder of the chapter, we will derive analytically large SNR approximations for $P_{\text{delay}}(d)$ and $P_{\text{ch}}(d)$ and show that given a fixed and high ρ, $P_{\text{delay}}(d)$ is increasing on d while $P_{\text{ch}}(d)$ is decreasing on d (confirming the above intuition). Furthermore, we find the best PHY layer operating point, i.e. diversity gain d^{*}, so as to minimize the total error probability P_{tot} in the high SNR regime. In other words, we will find d^{*} that balances the exponents of the two probabilities.

4.2 Problem Analysis

In this section, we analytically derive the two loss probabilities, $P_{\text{ch}}(d)$ and $P_{\text{delay}}(d)$, for asymptotically large SNR. As we will see, the two probabilities decay exponentially with SNR. For the channel, the definition of diversity gain [78] gives a direct asymptotic approximation of $P_{\text{ch}}(d)$ for large SNR. Obtaining the asymptotic $P_{\text{delay}}(d)$, however, requires more work. Depending on the value of d, we either directly compute the asymptotic $P_{\text{delay}}(d)$ or provide lower and upper bounds of the asymptotic $P_{\text{delay}}(d)$.

4.2.1 Asymptotic $P_{\text{ch}}(d)$

The asymptotic expression of $P_{\text{ch}}(d)$ for large SNR comes directly from the definition of diversity gain. By the union bound and the symmetry among users, we have the following bounds:

$$P^1_{\text{ch}}(d) \leq P_{\text{ch}}(d) \leq K P^1_{\text{ch}}(d)$$

where $P^1_{\text{ch}}(d)$ is the probability of decoding error for user 1. Using $P^1_{\text{ch}}(d) \doteq \rho^{-d}$ in Definition 2 and the fact that K is a constant independent of ρ, we have

$$P_{\text{ch}}(d) \doteq \rho^{-d}. \tag{4.15}$$

4.2.2 Asymptotic $P_{\textbf{delay}}(d)$

Similarly, by the union bound and the symmetry among users, we have the following bounds:

$$\Pr[D^1 > D] \leq P_{\text{delay}}(d) \leq K \Pr[D^1 > D], \tag{4.16}$$

where $\Pr[D^1 > D]$ implicitly depends on d.

Now, let us first focus on $\Pr[D^1 > D]$. To get an analytical expression for the asymptotic $\Pr[D^1 > D]$, we consider two cases depending on the value of d.

Case 1: Single-user performance regime ($d_0 \leq d < n_t n_r$)

As discussed in Section II-C, the multiplexing gain region $\mathcal{R}(d)$ in this regime is a square and the scheduler assign decoupled rates to the queues. Hence, the optimal scheduler simply assigns a fixed transmission rate of $C_{av}(d)$ given in (4.10) to each user (i.e. we call this the *symmetric static scheduler*). Therefore, the asymptotic approximation (when D is sufficiently large) of the delay violation probability is given as follows.

Lemma 4.9. *For $d_0 \leq d \leq n_t n_r$ and sufficiently large D, the asymptotic large-SNR approximation of* $\Pr[D^1 > D]$ *is such that*

$$\lim_{\rho\to\infty} \frac{\log \Pr\left[D^1 > D\right]}{\log \rho} = -\sigma_s(d) D T r_{av}(d) \tag{4.17}$$

where $\sigma_s(d)$ is defined such that

$$\Lambda(\sigma_s(d)) = \sigma_s(d) C_{av}(d). \tag{4.18}$$

Equivalently, we can write (4.17) as

$$\Pr[D^1 > D] \doteq \rho^{-\sigma_s(d) D T r_{av}(d)}, \tag{4.19}$$

where the proof of this Lemma is given in Appendix C.2.

Case 2: Antenna-pooling regime ($0 < d < d_0$)

In this case, the multiplexing gain region $\mathcal{R}(d)$ is polymatroid and the transmission rates of the users must be jointly allocated by a scheduler. Since the optimal policy (with respect to the delay violation probability objective) is unknown, we provide the following lower bound and upper bound to $\Pr[D^1 > D]$.

Upper Bound on $\Pr[D^1 > D]$

An upper bound on $\Pr[D^1 > D]$ is easily found since any feasible scheduling policy can provide an upper bound. In particular, to arrive at the upper bound $P^u(d)$, we consider the same symmetric static scheduler as described in Case 1: the symmetric static scheduler always assigns the symmetric rate of $C_{av}(d)$ to each user at all time. By Lemma 4.9, the asymptotic approximation of $P^u(d)$ for large D is given as

$$P^u(d) \doteq \rho^{-\sigma_s(d) D T r_{av}(d)}. \tag{4.20}$$

We note that this upper bound becomes tighter as d increases to d_0 since $\mathcal{R}(d)$ approaches a K-dimensional hypercube.

Lower Bound on $\Pr[D^1 > D]$

The lower bound $P^l(d)$ on $\Pr[D^1 > D]$ is obtained from Fact 4.10, the construction of a fictitious system, and Fact 4.11, as follow.

Fact 4.10. *Consider two systems whose multiplexing gain regions are given by* $\mathcal{R}_1$ *and* $\mathcal{R}_2$*, respectively, where* $\mathcal{R}_1 \subseteq \mathcal{R}_2$*. The delay violation probability associated with the second system is no greater than that of the first system.*

For a given d, consider a fictitious system whose multiplexing gain region is given by

$$\mathcal{R}_{\text{fic}}(d) := \left\{ (r_1, \dots, r_K) : \sum_{i=1}^{K} r_i \leq r^*_{Kn_t,n_r}(d) \right\}. \tag{4.21}$$

Since $\mathcal{R}(d) \subseteq \mathcal{R}_{\text{fic}}(d)$, Fact 4.10 states that the delay violation probability for this system is a lower bound for $\Pr[D^1 > D]$. Stolyar and Ramanan [72] have shown that the largest-delay-first (LDF) policy achieves the minimum asymptotic delay violation probability for this fictitious system.

Fact 4.11. *(Theorem 2.2 in [72]) Consider a single-server queuing model with* K *users (illustrated in Figure 4.4(iii) for* $K = 2$*). For the sources considered in this chapter, the largest-delay-first (LDF) policy achieves the minimum delay violation probability* $\Pr[\max_{i=1,\dots,K} D^i > D]$ *when* D *is large.* [5]

Hence, we compute the minimum asymptotic delay violation probability

[5] *See more details of this fact in Fact C.1. The result in [72] is much more general than this. It works with any weighted delays, i.e.* $\Pr[\max_{i=1,\dots,K} D^i/\alpha_i > D]$*, where* α_i *is the weight for user* i*.*

for this fictitious system to arrive at a lower bound, $P^l(d)$, for $\Pr[D^1 > D]$, as given in the following Lemma:

Lemma 4.12. *For $0 < d < d_0$ and sufficiently large D, the asymptotic large-SNR approximation of $P^l(d)$ is given by*

$$P^l(d) \doteq \rho^{-K\sigma_s(d)DTr_{av}(d)}. \tag{4.22}$$

Proof. See Appendix II. □

We also note that $P^l(d)$ becomes a tighter bound as $d \to 0$.

Remark 4.13. Comparing the exponents in (4.20) and (4.22), we see that the LDF scheduler improves the exponent of the delay violation probability by K times of that of the symmetric static scheduler. Talking in the language of channel diversity, the LDF scheduler improves the diversity gain by K folds by taking advantage of statistical-multiplexing of the sources. However, we want to emphasize that the lower bound in $P^l(d)$ derived from the fictitious system with the multiplexing gain region $\mathcal{R}_{\text{fic}}(d)$ becomes more loose as the number of users K grows. This is expected because the actual multiplexing gain region $\mathcal{R}(d)$ in (4.5) is a polymatroid (see an example of $K = 3$ users in Figure 4.3) while that of the fictitious system is just the K-dimensional simplex given by the constraint $\sum_{i=1}^{K} r_i \leq r^*_{Kn_t,n_r}(d)$. Thus, the lower-bound becomes more optimistic as the number of users increases.

Remark 4.14. For an example of $K = 2$ users, Figure 4.4 summarizes the two bounds with the queuing models in mind. The upper bound $P^l(d)$ is the tail probability of system (i) which always serves each queue with multiplexing gain

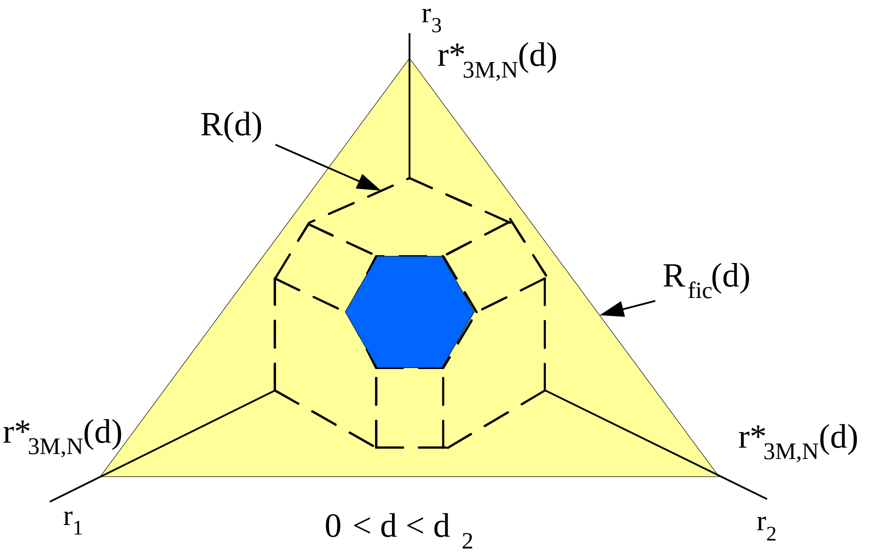


Figure 4.3 The MIMO-MAC multiplexing gain region $\mathcal{R}(d)$ and the multiplexing gain region of the fictitious system $\mathcal{R}_{\text{fic}}(d)$ for $K = 3$ users.

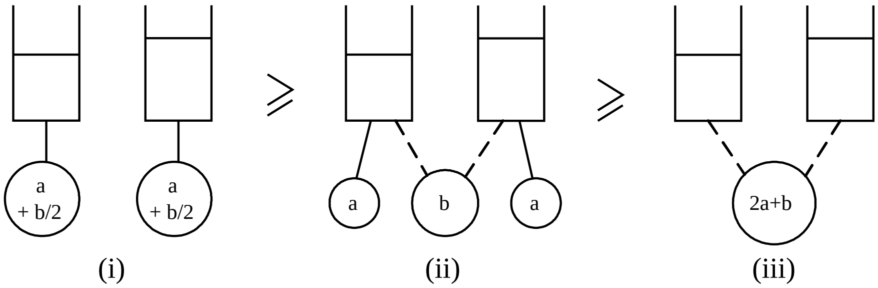


Figure 4.4 Queuing models of the upper and lower bounds of $\Pr[D^1 > D]$ for the case of $K = 2$ users and the antenna-pooling regime $(r^*_{2n_t,n_r}(d) \leq 2r^*_{n_t,n_r}(d))$, where a and b are defined such that $a + b = r^*_{n_t,n_r}(d)$ and $2a + b = r^*_{2n_t,n_r}(d)$.

$r^*_{2n_t,n_r}(d)/2$. The lower bound $P^l(d)$ is the tail probability of system (iii) which assigns the single server of multiplexing gain $r^*_{2n_t,n_r}(d)$ based on LDF scheduling. System (ii) is the queuing model given by the multiplexing gain region $\mathcal{R}(d)$.

Now, using the above two cases and the bounds in (4.16), we arrive at an asymptotic characterization of $P_{\text{delay}}(d)$ as follows

$$P_{\text{delay}}(d) \doteq \Pr[D^1 > D], \tag{4.23}$$

and, in particular, when $d_0 \leq d < n_t n_r$,

$$P_{\text{delay}}(d) \doteq \rho^{-\sigma_s(d)DTr_{av}(d)} \tag{4.24}$$

and, when $0 < d \leq d_0$,

$$\rho^{-K\sigma_s(d)DTr_{av}(d)} \dot{\leq} P_{\text{delay}}(d) \dot{\leq} \rho^{-\sigma_s(d)DTr_{av}(d)}. \tag{4.25}$$

In summary, so far, we have seen that $P_{\text{delay}}(d)$ and $P_{\text{ch}}(d)$ exponentially decay with ρ. The rate of decay of $P_{\text{ch}}(d)$ is known. When $d > d_0$, the rate of decay of $P_{\text{delay}}(d)$ is known via (4.24). The rate of decay of $P_{\text{delay}}(d)$ when $d < d_0$ is, however, unknown but is bounded as in (4.25).

Next, with $P_{\text{delay}}(d)$ and $P_{\text{ch}}(d)$ at hand, we proceed with the minimization of the total error probability.

4.2.3 Minimizing Asymptotic Total Error Probability

From the asymptotic expressions of $P_{\text{ch}}(d)$ given in (4.15) and $P_{\text{delay}}(d)$ in (4.24) and (4.25), the asymptotic characterization of the total error probability $P_{\text{tot}}(d) \doteq P_{\text{delay}}(d) + P_{\text{ch}}(d)$ is immediate:

For $d_0 \leq d \leq n_t n_r$,

$$P_{\text{tot}}(d) \doteq \rho^{-\sigma_s(d)DTr_{av}(d)} + \rho^{-d}. \tag{4.26}$$

For $0 < d < d_0$,

$$\rho^{-K\sigma_s(d)DTr_{av}(d)} + \rho^{-d} \dot{\leq} \ P_{\text{tot}}(d) \dot{\leq} \rho^{-\sigma_s(d)DTr_{av}(d)} + \rho^{-d}. \tag{4.27}$$

Since the term $\sigma_s(d)DTr_{av}(d)$ is decreasing in d while the term d is increasing on d, the minimum of $P_{\text{tot}}(d)$ in (4.26) or its bounds in (4.27) happen when the value of d makes the exponents of the two terms are within $o(1)$ of each other (note that if the exponents were not in the same order, one term would dominate in the sum as $\rho \to \infty$). We now introduce an algorithm which guarantees such choices of d:

Algorithm 1:

1. Solve for d which is a solution of

   $$\sigma_s(d)DTr_{av}(d) = d. \tag{4.28}$$

   If $d \geq d_0$, then $d^* = d_u^* = d_l^* = d$ and stop. Otherwise, set $d_l^* = d$. Go to Step 2.

2. Solve for d which is a solution of

   $$K\sigma_s(d)DTr_{av}(d) = d \tag{4.29}$$

   and set $d_u^* = \min(d, d_0)$.

Theorem 4.1. *Algorithm 1 results in a closed interval $[d_l^*, d_u^*]$ in which the optimal common diversity gain d^* lies.*

Proof. See Appendix C.2. □

Notice that the optimal diversity d^* and its bounds depend on the statistical characteristics of the symmetric sources (Λ, μ, λ), the parameters of the MIMO-MAC channel (e.g. T, n_t, n_r), and the delay bound D.

4.2.4 Statistical-Multiplexing and Optimal Diversity Gain

From the above analysis, we obtain the following critical observation. Given a delay constraint, the statistical property of the source has a significant impact on the level of diversity a well-designed system can enjoy. In other words, the optimal scheduler which statistically multiplexes the MIMO resources allows the combined bursty sources to perceive as smaller aggregate traffic and hence a higher degree of diversity. Rigorously, this performance improvement can be attributed to a *statistical-multiplexing gain* as follows:

Definition 4.15. An optimal dynamic scheduler with the total error probability P^*_{tot} provides *statistical-multiplexing gain* of s over the static rate scheduler with P^f_{tot}, where

$$s := -\lim_{\rho\to\infty} \frac{\log P^*_{\text{tot}} - \log P^f_{\text{tot}}}{\log \rho}. \tag{4.30}$$

From this definition and the fact that $P^f_{\text{tot}} \doteq \rho^{-d_l^*}$, the following lemma is immediate.

Lemma 4.16. *Consider the system model in Section II. The optimal statistical-multiplexing gain s^* is given by*

$$s^* = d^* - d_l^*. \tag{4.31}$$

Furthermore, it is bounded above by $d_u^* - d_l^*$.

Remark 4.17. The two concepts of "statistical-multiplexing gain" and "multi-user diversity gain" are related conceptually. The former takes advantage of the troughs (due to burstiness) of the traffic of different users while the latter takes advantage of the peaks (due to fadings) of the channels of different users. But their impacts on the design are sufficiently different, as multi-user diversity gain requires channel CSI at the transmitters while statistical-multiplexing requires QSI.

4.2.5 Resource Pooling and Statistical-Multiplexing

Here we discuss the effect of the arrival rate λ and the average delay bound D to the performance region of the MIMO-MAC. The relationship between (λ, D) and the system performance is summarized in Figure 4.5. The system performance is divided into three main regions: the single-user performance region, the antenna-pooling with *significant* statistical-multiplexing region, and the antenna-pooling with *insignificant* statistical-multiplexing region. In the single-user performance region, the achieved optimal diversity gain d^* is equivalent to the case when only one user is in the system, i.e. the case $d^* \geq d_0$. Specifically, this case happens when

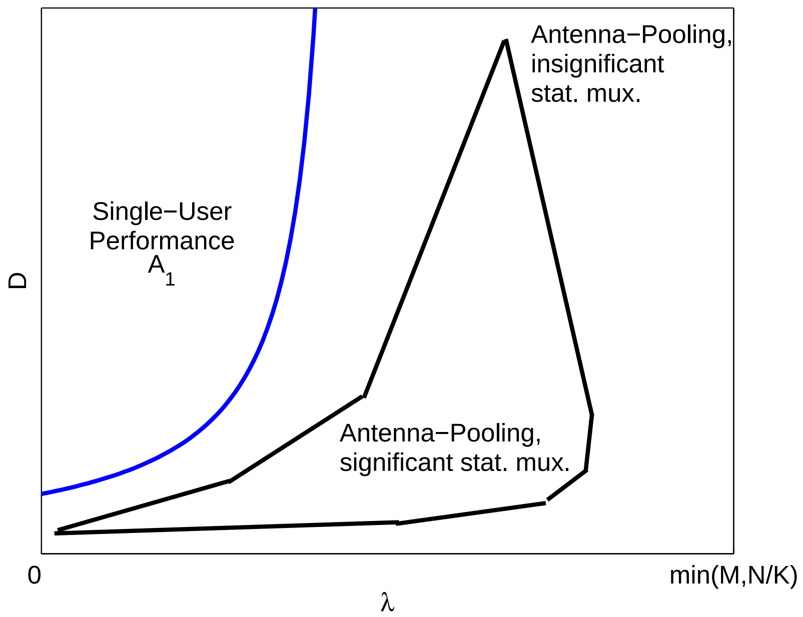


Figure 4.5 The relation between (λ, D) to the system performance.

λ is sufficiently small and D is sufficiently large. We denote this region as $\mathcal{A}_1$.

$$\mathcal{A}_1 := \left\{ (\lambda, D) : \lambda \leq r_0, D \geq \frac{d_0}{\sigma_s(d_0)Tr_0} \right\}$$

where $r_0 = r^*_{n_t,n_r}(d_0)$.[6] Since in this region the transmission rate of each user is independent, there is no resource sharing and hence no statistical-multiplexing gain.

On the other hand, the significance of statistical-multiplexing gain outside $\mathcal{A}_1$ is impacted by the rate of arrivals λ as well as the average delay bound D. In particular, for (λ, D) in the neighborhood of $\mathcal{A}_1$, the statistical-multiplexing gain is not significant as the queues still behave in a roughly independent manner. Similarly, as λ increases to an overload situation or the delay bound D becomes

[6] Note that $\frac{d_0}{\sigma_s(d_0)Tr_0}$ is an increasing function on λ since $\sigma_s(d_0)$ which is the delay violation exponent is itself decreasing on the arrival rate λ.

very tight, the benefits of juggling resources diminishes. In contrast, for medium values of λ and D and under the optimal dynamic scheduler, each queue perceives the whole (pooled) resource to itself, compared to $1/K$ of the resource as in case of a symmetric static scheduler.

To illustrate the approach shown in this chapter and the corresponding calculation, we look at a simple example of a compound Poisson source with $K = 2$ users in the next section.

4.3 Example: Compound Poisson Sources and $K = 2$

In this section, we illustrate the proposed approach via an example. We consider two independent but identical source processes. For each source i, arrivals are independent across timeslots. The number of bits that arrive in a timeslot t, A_t^i, is an aggregation of a random number of packets whose sizes are also random, i.e. $A_t^i = \sum_{n=1}^{N} Y_n$. Furthermore, we assume that the number of packets at each slot, N, is an independent Poisson random variable with rate ν packets per timeslot, while the length of the packets, Y_i, $i = 1, 2, \ldots$, are i.i.d. random variables with exponential distribution of mean $1/\mu$. The average bit arrival rate $\tilde{\lambda}$ for each source is equal to ν/μ and scales with $\log \rho$ as in (4.7), i.e.

$$\tilde{\lambda} = \nu/\mu = \lambda T \log \rho. \tag{4.32}$$

Proposition 4.18. *For the compound Poisson source with exponential packet length, the $\sigma_s(d)$ defined in Lemmas 4.9 is given as*

$$\sigma_s(d) = \mu(1 - \frac{\lambda}{r_{av}(d)}). \tag{4.33}$$

Proof. See Appendix C.2. □

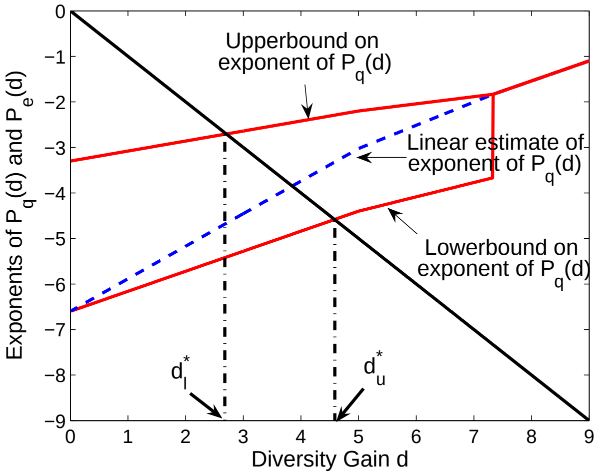
Note that the ratio of the per-queue average bit arrival rate over the average service rate, $\frac{\lambda}{r_{av}(d)}$, can be called the traffic load per queue. It is important to note that the delay violation exponent in (4.33) is a decreasing function of the average packet size $1/\mu$, for a fixed packet arrival rate ν. A larger packet size in effectively creates more burstiness in the arrivals, hence a higher delay violation probability.

With Proposition 4.18 in hand, we are now ready to use Algorithm 1 to obtain d^* (or its bounds). Figure 4.6 shows the optimal d^* and its bounds when $n_t = n_r = 4$, the average packet size $1/\mu$ is 100 nats, and the symbol rate such that there are $T = 2n_t + n_r - 1 = 11$ symbols per timeslot. In these figures, we plot the exponents of $P_{\text{delay}}(d)$ or its bounds, i.e. $-\sigma_s(d)DTr_{av}(d)$ and $-2\sigma_s(d)DTr_{av}(d)$, and the exponent of $P_{\text{ch}}(d)$, i.e. $-d$. To better illustrate the procedure followed by Algorithm 1, we plot the exponents of $P_{\text{delay}}(d)$ and $P_{\text{ch}}(d)$ separately. Note that when $d \leq d_0$ ($d_0 = 7.3$ in this example), we only have lower and upper bounds for the exponents of $P_{\text{delay}}(d)$. In this case, in addition to the bounds, we plot a linear approximation (dotted line) to emphasize the tightness of the lower bound around d_0 and the upper bound around 0. The optimal diversity gain d^* or its bounds (d_l^*

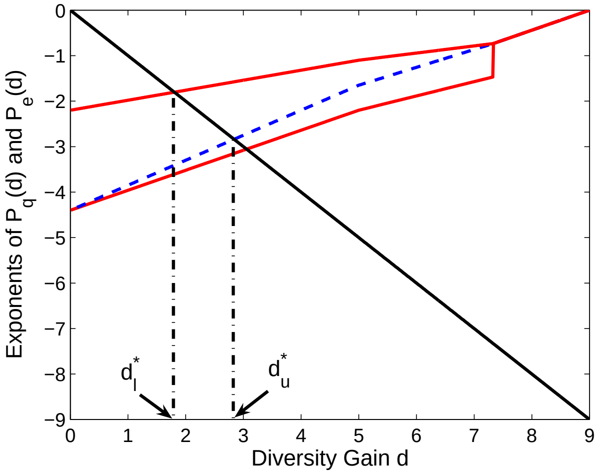
and d_u^*) are shown in each plot as the crossing of the exponents.

As we discussed in Section 4.2.5, the statistical-multiplexing gain s^* defined in (4.31) depends on the optimal choice of d^* which itself is a function of the arrival rate λ and the average delay bound D. Depending on λ and D, we may or may not have statistical-multiplexing gain. For example, Figure 4.7(a) shows that, in the case of sufficiently low arrival rates and a large delay bound, the optimal P_{tot}^* happens when the users operate in the single-user performance region. Hence, in this case, there is no statistical-multiplexing gain to be achieved by dynamic scheduler. Here, the dominant form of loss occurs on the channels. On the other hand, Figure 4.6(b) corresponds to the case of large arrival rate and small delay bound, where the loss probability due to delay violation dominates that of the channel. In this case, the optimal diversity d^* necessitates resource sharing in form of antenna pooling. As a result, the impact of n optimal dynamic scheduler, in a form of statistical-multiplexing gain, becomes more significant.

Figure 4.8 illustrates the performance region discussed in Section 4.2.5. In particular, the figure gives a characterization of the region shown in Figure 4.5 for the compound Poisson case. We approximate the statistical-multiplexing gain $s^* = d^* - d_l^*$ achieved with an optimal dynamic scheduler with that of a simple linear approximation $d_a^* - d_l^*$, where d_a^* is the optimal diversity gain derived from the dotted line in Figure 4.6 and Figure 4.7.

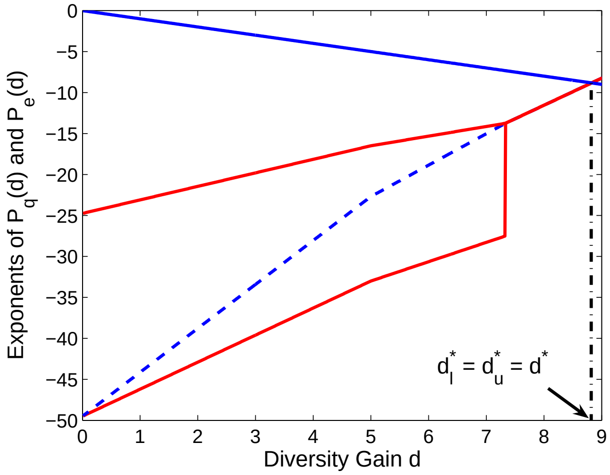


(a) $\lambda = 0.5, D = 20$

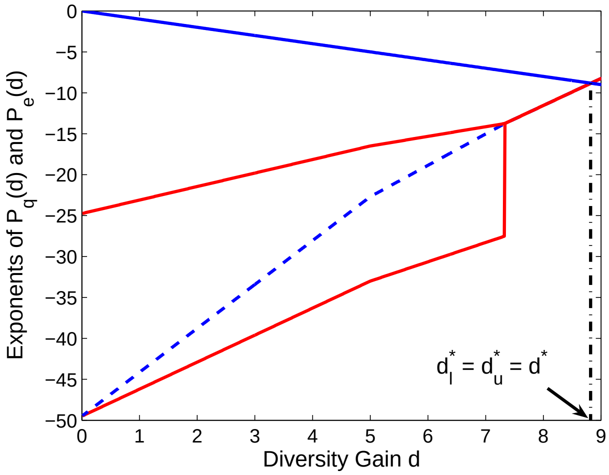


(b) $\lambda = 1, D = 20$

Figure 4.6 Plots of the exponents of $P_{\text{delay}}(d)$ and $P_{\text{ch}}(d)$ for two different average arrival rates ($\lambda = 0.5$ and 1) and delay bound $D = 20$. For $d < d_0 = 7.3$, the upper and lower bounds of the exponent of $P_{\text{delay}}(d)$ are shown. A simple linear estimate (dotted line) of the exponent of $P_{\text{delay}}(d)$ between the two bounds is also drawn.

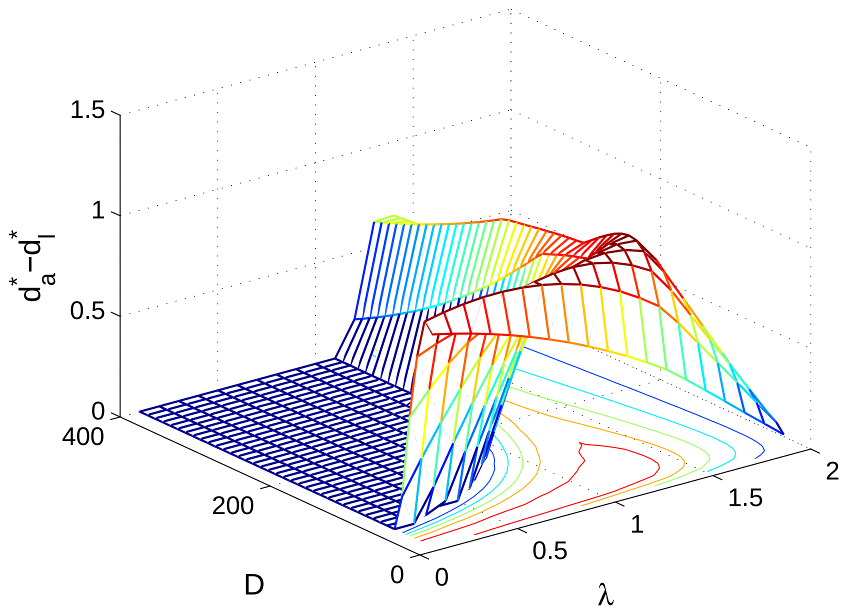


(a) $\lambda = 0.5, D = 150$

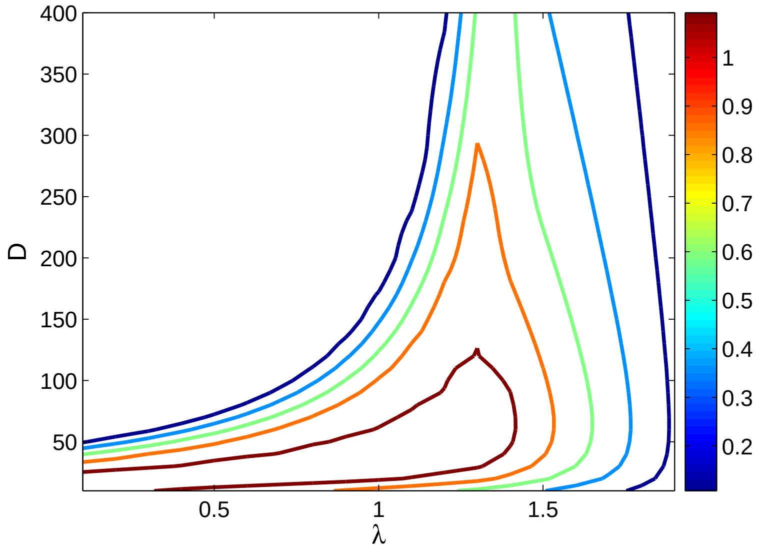


(b) $\lambda = 1, D = 150$

Figure 4.7 Similar plots as in Figure 4.6 but with delay bound $D = 150$.



(a) $d_a^* - d_l^*$ vs. (λ, D)



(b) Contour Plot $d_a^* - d_l^*$

Figure 4.8 3D and contour plots characterizing the statistical-multiplexing gain, approximated by $d_a^* - d_l^*$, v.s. delay bound D and arrival rate λ.

4.4 Summary

In this chapter, we considered a system of bursty and delay-sensitive symmetric sources concatenated with a symmetric MIMO-MAC channel. We assumed no CSI information available to the transmitters and a block fading model with a block coding whose block lengths are matched to the coherence time of the channel. Furthermore, we assumed a fixed and equal high transmission power at each transmitter, i.e., high SNR regime. We addressed the optimal choice of the spatial diversity gain d^* such that it minimizes an end-to-end loss performance where loss can occur due to delay violation as well as channel decoding error. We showed how an optimal choice of diversity gain d^* depends on a queue-based scheduler module whose job is to statistically multiplex the resources of the MIMO-MAC. In doing so, we integrated the notion of statistical-multiplexing gain with those of spatial diversity, multiplexing, and multi-access gains provided by the MIMO-MAC.

Acknowledgment

This chapter, in full, appears in the following publication. The dissertation author was the primary investigator and author of this paper.

- S. Kittipiyakul and T. Javidi, "Optimal operating point for MIMO multiple access channel with bursty traffic," *IEEE Trans. Wireless Commun.*, vol. 6, no. 12, pp. 4464–4474, Dec. 2007.

I would like to thank my coauthor Prof. Tara Javidi.

Chapter 5

Many-Sources Large Deviations for Max-Weight Scheduling

Abstract

In this chapter, we establish a many-sources large deviations principle (LDP) for the stationary workload of a multi-queue single-server system with simplex capacity, operated under a stabilizing and non-idling maximum-weight scheduling policy. Assuming a many-sources sample path LDP for the arrival processes, we establish an LDP for the workload process by employing Garcia's extended contraction principle that is applicable to quasi-continuous mappings. The LDP result can be used to calculate asymptotic buffer overflow probabilities accounting for the multiplexing gain, when the arrival process is an average of i.i.d. processes. We express the rate function for the stationary workloads in term of the rate functions of the finite-horizon workloads when the arrival processes have

i.i.d. increments.

In this chapter, we consider a single-server multi-class discrete-time queueing system where the server is allocated to queues according to a maximum weight scheduler, which is known to be stabilizing [2]. We provide a refined analysis of the statistical performance of this policy under stochastic arrivals. In particular, with K independent queues we seek to derive the probability of buffer overflow. Specifically, for a given finite value B, we consider the transient behavior, i.e., quantities such $\Pr(Q_{0,T} \geq B\mathbf{1}_K)$ where $Q_{0,T} \in \mathbb{R}_+^K$ is the workload (to be formally defined later) at time 0 with "zero" initial workload at time $-T$ and $\mathbf{1}_K \in \mathbb{R}_+^K$ is the vector of all 1s, as well as the stationary behavior, i.e., the similar probabilistic quantities as before for the limiting workload vector as $T \to \infty$. Like many recent papers on analysis of scheduling algorithms [9, 67, 68, 72, 75, 85, 88], our work considers logarithmic asymptotics to the probabilities by analyzing a large-deviation approximation to the problem. The present chapter is closely related to [75], where the buffer overflow probability for the workload processes of a single-server multi-queue queueing system under max-weight policies and general compact and convex capacity regions was established. While [75] addresses the large-buffer scaling regime, this chapter establishes similar results for a classical multi-class single-server (simplex capacity region) system under a "many-sources" asymptotic regime (see [11, 15, 29, 68, 70, 81, 82, 85]).

In a many-sources asymptotic regime, one considers a sequence of queueing systems indexed by the number of the (independent) sources multiplexed (or averaged) over a particular queue, i.e., the arrival process to each queue is the average of L processes. The analysis focuses on the asymptotic behavior of the systems when $L \to \infty$. The motivation to consider many-sources scaling includes the following considerations: 1) practical interest in real applications when there are large number of flows to each user or node. This asymptote usually gives a more refined approximation to the probabilistic quantities of interest by incorporating the impact of the multiplexing gain [11, 12, 14, 15, 19, 70, 81]; and 2) a cross-layer optimization for the optimal duration of the finite code blocks when the transmission channel is operated at high-SNR regime (see [42]).

Given a sample path large deviation principle for the arrival processes (in the space of real-valued sequences with the scaled uniform topology), we derive a large deviations principle for the workload. In particular, we first show that the workload is a quasi-continuous map of the arrival process. The first contribution of the chapter is, thus, obtained based on a recent extension of the contraction principle by J. Garcia [32]. More precisely, we use Garcia's extended contraction principle together with an assumed sample path large deviations principle (LDP) (see Definition 5.2) for the arrival process to establish an LDP for the workload at any given time t as well as the stationary workload. The LDP results (Theorems 5.1 and 5.2) directly imply that the probability of buffer overflow has an exponential tail whose decay rate is dictated by a good rate function whose form is determined by the statistics of the arrival process. This rate function can be expressed as a solution to a finite-dimensional optimization problem which has the same flavor of a deterministic optimal control problem. When the arrival process has i.i.d. increments, we provide a simplified form for the rate function.

The outline of the chapter is as follows. The problem formulation is given in Section 5.1. Section 5.2 provides background and preliminary results on the large deviations principle. The main results of the chapter, which are the LDPs of the workloads, are given in Section 5.3. Section 5.4 gives simplified expressions of the rate functions. We conclude in Section 5.5.

5.1 Problem Formulation

We consider a discrete-time queueing system with K independent queues and one server with capacity c (bits per timeslot). For every queue $k \in \mathcal{K} := \{1, \ldots, K\}$ we assume that work (in bits) arrives into the queue given by a sequence $(A_t^k, t \in \mathbb{N})$ where $A_t^k \in \mathbb{R}_+$ is the work brought in at time $-t$. For $0 \leq m_1 \leq m_2$ integers, we define $A^k(m_1, m_2] := \sum_{t=m_1+1}^{m_2} A_t^k$ as the total amount of work to arrive for user k from timeslot $-m_2$ and until timeslot $-m_1 - 1$. We also write $A^k_{(m_1,m_2]}$ to denote the finite sequence of arrivals A^k restricted to time $\{-m_2, \ldots, -m_1 - 1\}$, i.e., $A^k_{(m_1,m_2]}$ is the vector $(A^k_{m_1+1}, A^k_{m_1+2}, \ldots, A^k_{m_2}) \in \mathbb{R}_+^{m_2-m_1}$.

We assume a maximum-weight server allocation policy where the weights are functions of the unfinished workloads, and under which we are interested in the statistical properties of the unfinished workload in queue k at time t. Let $Q_t^k \in \mathbb{R}_+$

be the unfinished workload (queue length) of queue k at the beginning of time $-t$ and R_t^k be the amount of service allocated to queue k during time $(-t, -t+1]$. Let $Q_t := (Q_t^k, k \in \mathcal{K})$ be the corresponding workload vector and $R_t := (R_t^k, k \in \mathcal{K})$ be the rate vector. One can define a simplex rate region $\mathcal{R}$,

$$\mathcal{R} := \left\{ \mathbf{r} = (r^1, \ldots, r^K) \in \mathbb{R}_+^K : \sum_{k=1}^{K} r^k \leq c \right\}, \tag{5.1}$$

as the set of server's operating points, i.e., $R_t \in \mathcal{R}$. At the beginning of timeslot $-t$, the rate vector $R_t \in \mathcal{R}$ is selected by a work-conserving max-weight scheduler H in response to the current workload Q_t; that is, $R_t = H(Q_t)$ where the scheduler H serves c bits from the queue k^* which has the largest workload Q_t^k when the workload of the longest queue is at least c. In case of a tie, the scheduler chooses the queue with the lowest index. To make the scheduler non-idling, we assume the scheduler splits the service when the unfinished workload in each queue is less than c. That is, we assume that the scheduler assigns $H(\mathbf{x}) = \mathrm{Proj}_{\mathcal{R}}(\mathbf{x})$ when $\mathbf{x} \in [0, c)^K$, where $\mathrm{Proj}_B(\mathbf{b})$ is the projection of vector $\mathbf{b}$ on the set B. Specifically, for $\mathbf{x} \in \mathbb{R}_+^K$ we consider $H(\mathbf{x})$ to be given by

$$H(\mathbf{x}) := \begin{cases} \mathbf{e}(\mathbf{x}) & \text{if } \mathbf{x} \notin [0, c)^K; \\ \mathrm{Proj}_{\mathcal{R}}(\mathbf{x}) & \text{if } \mathbf{x} \in [0, c)^K. \end{cases} \tag{5.2}$$

Above $\mathbf{e}(\mathbf{x})$ is defined as the K-dimensional vector whose elements are zeros except for the k^*th element which is c, where $k^* = \min\{k : k \in \arg\max_{i \in \mathcal{K}} x_i\}$. For example, when $K = 2$, the scheduler H in (5.2) becomes

$$H(\mathbf{x}) = \begin{cases} (c, 0), & \text{if } x^1 \geq x^2, x^1 \geq c \\ (0, c), & \text{if } x^1 < x^2, x^2 \geq c, \\ \mathrm{Proj}_{\mathcal{R}}(\mathbf{x}), & \text{if } x^1 < c, x^2 < c. \end{cases} \tag{5.3}$$

For $t \in \mathbb{N}$, the dynamics of the workloads of queue $k \in \mathcal{K}$ is

$$Q^k_{t-1} = [Q^k_t - R^k_t]^+ + A^k_t, \tag{5.4}$$

where for $x \in \mathbb{R}$, $[x]^+ := \max\{0, x\}$. We assume that the arrival vector A_t happens any time in $(-t, -t+1)$ but cannot be served in that timeslot t.

In this chapter, we are interested in the asymptotic probabilities of the *finite-horizon* and *infinite-horizon* workloads. The finite-horizon workload, denoted by $Q_{0,T}$, is the workload at time 0, assuming the initial condition at time $-T$ is $Q_T \in \mathcal{R}$. The index T in $Q_{0,T}$ reminds us of this initial condition.[1] The infinite-horizon workload, Q, is defined as $Q = Q(A) := \lim_{T\to\infty} Q_{0,T}(A_{(0,T]})$. We assume that the limit exists but may be infinite. It can be shown that Q is the stationary workload when the system is stable. We will use the function G_T to mean $G_T(A_{(0,T]}) = Q_{0,T}(A_{(0,T]})$ and the function G to mean $G(A) = Q(A)$. To aid in describing our results we further define $G^{\mathbf{a}}_T$ and $G^{\mathbf{a}}$ in the following way:

Definition 5.1. For a function $F : \mathcal{X} \mapsto \mathcal{Y}$ and $x \in \mathcal{X}$, we define

$$F^x := \{y \in \mathcal{Y} : (\exists x_n \to x) \text{ such that } F(x_n) \to y\}. \tag{5.5}$$

[1] The initial condition is normally taken to be the zero vector but the result remains valid even when the initial condition is within $\mathcal{R}$. With $Q_T \in \mathcal{R}$, we always have the workload at time $-T+1$ be $Q_{T-1} = [Q_T - H(Q_T)]^+ + A_T = A_T$ from the non-idling condition that we imposed on the server allocation mechanism.

Note that $F^{(\cdot)}$ is a set-valued mapping. It is single-valued at x where F is continuous (i.e., $F^x = \{F(x)\}$).

We consider a sequence of queueing systems indexed by $L \in \mathbb{N}$ and will be interested in the behavior of the queueing system L as L becomes large. For each user $k \in \mathcal{K}$ and system indexed by L, we assume a stationary arrival process of work brought into the system given by a sequence $A^{k,L} := (A_t^{k,L}, t \in \mathbb{N})$ where $A_t^{k,L} \in \mathbb{R}_+$ is the work (in bits) brought in at time $-t$ into the queue of user k. The arrivals to different queues/users are mutually independent. We follow the many-sources scaling regime on the system with index L. The arrival process to each queue k is assumed to be an average of L i.i.d. processes, i.e., $A^{k,L} := \frac{1}{L}\sum_{i=1}^{L} A^{k,(i)}$, where each $A^{k,(i)}$ is an independent identically distributed copy of a stationary process A. We denote the mean arrival rate by $\mu := EA_1^{k,L} = EA_1$. Also let $A^L := (A^{k,L}, k \in \mathcal{K})$ be the sequence of arrival vectors.

5.1.1 Main Results

Assuming that the sequence of the arrival processes $\{A^L\}$ satisfies a many-sources sample path LDP with a continuous rate function (Assumptions 5.3 and 5.5, respectively, given in Section 5.2), the main results of the chapter are the following LDP's for the finite and infinite-horizon workloads. We also provide a simplification of the rate functions when the arrival processes have i.i.d. increments.

Theorem 5.1. *For $t \in \mathbb{N}$, the sequence of the finite-horizon workloads*

$\{Q_{0,t}(A^L_{(0,t]}) := G_t(A^L_{(0,t]})\}$ *satisfies an LDP on* $\mathbb{R}^K_+$ *with the rate function* I_t*, where for* $\mathbf{b} \in \mathbb{R}^K_+$

$$I_t(\mathbf{b}) = \inf_{\mathbf{x} \in \mathbb{R}^{K \times t}_+ : G^{\mathbf{x}}_t \ni \mathbf{b}} I^{\sharp}{}_t(\mathbf{x}) \tag{5.6}$$

Theorem 5.2. *If* $K\mu < c$*, the sequence of infinite-horizon workloads* $\{Q(A^L) := G(A^L)\}$ *satisfies an LDP on* $\mathbb{R}^K_+$ *with rate function* J*, where for* $\mathbf{b} \in \mathbb{R}^K_+$

$$J(\mathbf{b}) = \inf_{a \in \mathcal{D}^K_\mu : G^a \ni \mathbf{b}} I^{\sharp}(a). \tag{5.7}$$

In the above results, a denotes a sequence taking values in $\mathbb{R}^K_+$ and $\mathcal{D}^K_\mu$ is a special subset of sequences taking values in $\mathbb{R}^K_+$ which will be clarified in Section 5.2.1 .

5.2 Background and Assumptions

5.2.1 Topology for Sample Paths

Since a large deviations principle is defined with topological entities and since we will deal with continuity and convergence of the workload mappings, we need to precisely specify the topology for the space of the arrival sample paths. We use the scaled uniform topology as in [82] for our analysis. Let $\mathcal{D}$ denote the space of sample paths (non-negative discrete-time functions), i.e., $\mathcal{D} := \{x : \mathbb{N} \mapsto \mathbb{R}_+\}$, and let $\mathcal{D}^K$ be the K cartesian product of $\mathcal{D}$. Let $||\cdot||_u$ be the scaled uniform norm on $\mathcal{D}$, i.e., $||x||_u := \sup_{t \in \mathbb{N}} \left| \frac{x(0,t]}{t} \right|$ for all $x \in \mathcal{D}$ while for all $a = (a^k, k \in \mathcal{K}) \in \mathcal{D}^K$, where $a^k \in \mathcal{D}$, the scaled uniform norm of a is $||a||_u := \max_{k \in \mathcal{K}} ||a^k||_u$. Define a subspace $\mathcal{D}_\mu$ of $\mathcal{D}$ which contains all the arrival paths whose average arrival rate is equal to the expected rate μ, i.e., $\mathcal{D}_\mu := \left\{x \in \mathcal{D} : \lim_{t\to\infty} \frac{x(0,t]}{t} = \mu\right\}$ and $\mathcal{D}_\mu^K$ the K products of $\mathcal{D}_\mu$. Again, we equip $\mathcal{D}_\mu$ and $\mathcal{D}_\mu^K$ with the scaled uniform topology. For metric spaces like $\mathbb{R}_+^n$, $n \in \mathbb{N}$, we use the square uniform topology with the square metric ρ [59], where $\rho(\mathbf{x}, \mathbf{y}) := \max_{i\in\{1,\ldots,n\}} |x_i - y_i|$.

5.2.2 Large Deviations Principle

The following definition of a large deviations principle is taken from [82]. For an excellent full introduction to the theory, definitions, and tools, see [18] and for queueing applications, see [29].

Definition 5.2 (Large deviations principle)**.** A sequence of random variables X^L in a Hausdorff space $\mathcal{X}$ with σ-algebra $\mathcal{B}$ is said to satisfy a large deviations principle (LDP) with good rate function I if, for any set $B \in \mathcal{B}$,

$$-\inf_{x\in B^o} I(x) \le \liminf_{L\to\infty} \frac{1}{L} \log \Pr(X^L \in B) \le \limsup_{L\to\infty} \frac{1}{L} \log \Pr(X^L \in B) \le -\inf_{x\in \bar{B}} I(x), \tag{5.8}$$

where B^o and $\bar{B}$ are the interior and the closure of B, respectively, and the rate function $I : \mathcal{X} \mapsto \mathbb{R}_+ \cup \{\infty\}$ has compact level sets, where the level sets are defined as $\{x : I(x) \le \alpha\}$, for $\alpha \in \mathbb{R}$. If X^L is a mapping from $\mathbb{N}$ to $\mathbb{R}$ describing sample path of a random sequence, the LDP is referred to as a *sample path* LDP.

We are interested in finding an LDP for the sequence of the workloads $Q(A^L)$ and $Q_{0,T}(A^L_{(0,T]})$, assuming the following sample path LDP of the arrival processes A^L.

5.2.3 Sample Path LDP of Arrival Processes

The following sample path LDP for the sequence of arrival processes A^L is the starting point of our analysis.

Assumption 5.3 (Many-sources sample path LDP)**.** The sequence $\{A^L\}$ satisfies a sample path LDP in $\mathcal{D}_\mu^K$ equipped with the scaled uniform topology with rate function $I^\sharp$, where the rate function $I^\sharp$ is given as

$$I^\sharp(a) := \sup_{t\in\mathbb{N}} {I^\sharp}_t(a_{(0,t]}) = \lim_{t\to\infty} {I^\sharp}_t(a_{(0,t]}) \tag{5.9}$$

for $a \in \mathcal{D}_\mu^K$, where for $\mathbf{x} = (\mathbf{x}^k \in \mathbb{R}_+^t, k \in \mathcal{K}) \in \mathbb{R}_+^{Kt}$,

$${I^\sharp}_t(\mathbf{x}) := \sum_{k=1}^K \Lambda_t^*(\mathbf{x}^k), \tag{5.10}$$

and Λ_t^* is the convex conjugate or Fenchel-Legendre transform of Λ_t:

$$\Lambda_t^*(\mathbf{y}) := \sup_{\theta\in\mathbb{R}^t} \theta \cdot \mathbf{y} - \Lambda_t(\theta), \qquad \text{for } \mathbf{y} \in \mathbb{R}^t, \tag{5.11}$$

$$\Lambda_t(\theta) := \log E \exp\left(\theta \cdot A_{(0,t]}\right), \qquad \text{for } \theta \in \mathbb{R}^t. \tag{5.12}$$

Remark 5.4. Assumption 5.3 implies that the sequence $\{A^L\}$ also satisfies an LDP on $\mathcal{D}^K$ equipped with the scaled uniform topology, with rate function $I^\sharp$ where $I^\sharp(a) = \infty$ for $a \in \mathcal{D}^K/\mathcal{D}_\mu^K$ [29]. It is shown in [29, Lemma 7.8] that under Assumption 5.3, $\Lambda_t^*(\cdot)$ is non-negative, Λ_t^* is convex, and $\Lambda_t^*(\mu\mathbf{1}_t) = 0$, where $\mathbf{1}_n$ is the vector of all ones in $\mathbb{R}^n$. Hence, ${I^\sharp}_t(\mu\mathbf{1}_{Kt}) = 0$ and ${I^\sharp}_t$ is convex.

In this chapter, we also assume the following continuity condition on the rate function $I^\sharp$ in (5.9):

Assumption 5.5. $I^\sharp$ is continuous on its effective domain defined as $\{x \in \mathcal{D}^K : I^\sharp(x) < \infty\}$.

Remark 5.6. As shown in [82] and [29], the above many-sources sample path LDP (Assumption 5.3) holds when the underlying arrival process A satisfies mild regularity conditions. This implies that several standard stationary processes used for traffic modeling, such as i.i.d. increment processes, Markov-modulated, a general class of Gaussian, and fractional Brownian processes (for long-range dependent or heavy-tailed traffic), satisfy Assumptions 5.3 and 5.5.

5.2.4 Garcia's Extended Contraction Principle

The contraction principle (see [18, p. 126]) says that if we have an LDP for a sequence of random variables, we can effortlessly obtain LDP's for a whole other class of random sequences that are obtained via continuous transformations. However, due to the inherent discontinuity in the max-weight scheduling function, the usual contraction principle fails to provide sufficient structure. Instead, we will utilize the following powerful extension of the contraction principle for quasi-continuous transformations on metric spaces, given by Garcia [32]. First, let us provide the definition and condition of the quasi-continuity:

Definition 5.7 (Quasi-continuity)**.** A (single-valued) function $F : \mathcal{X} \mapsto \mathcal{Y}$ is *quasi-continuous* at $x \in \mathcal{X}$ if for every $y \in F^x$ and every pair (U, V) of neighborhoods of x and y respectively, there is a nonempty open subset U_0 of U such that $F(U_0) \subseteq V$. We say that F is quasi-continuous if it is quasi-continuous at every point of its domain. We also say that F is *strictly quasi-continuous* at $x \in \mathcal{X}$ if F is quasi-continuous but not continuous at x.

Fact 5.8. *[32, Theorem 3.2] If $\mathcal{X}, \mathcal{Y}$ are complete metric spaces, a function $F : \mathcal{X} \mapsto \mathcal{Y}$ is quasi-continuous if and only if for each $x \in \mathcal{X}$, there is a sequence $\{x_n\}$ such that $x_n \to x, F(x_n) \to F(x)$, and such that for all n, F is continuous at x_n.*

Remark 5.9. The definition of quasi-continuity is similar to that of continuity, where F is continuous at $x \in \mathcal{X}$ if for every neighborhood V of $F(x)$, there is a neighborhood U_0 of x such that $F(U_0) \subseteq V$ [59]. Obviously, every continuous function is quasi-continuous. A step function $F : \mathbb{R} \mapsto \mathbb{R}$, where $F(x) = 0$ for $x < 0$, $F(x) = 1$ for $x \geq 0$, is quasi-continuous. But if $F(0) = 1/2$, then F is not quasi-continuous. From this example, we can infer that our scheduling function H is quasi-continuous.

Remark 5.10. An interesting property is that if F is a continuous function and G is a quasi-continuous function, then $F \circ G$ is quasi-continuous but $G \circ F$ is not necessarily quasi-continuous [32].

Fact 5.11 (Garcia's Extended Contraction Principle)**.** *Assume $\Omega \overset{X^L}{\to} \mathcal{X} \overset{F}{\to} \mathcal{Y}$, $\mathcal{X}, \mathcal{Y}$ are metric spaces, and $\{X^L\}$ satisfies a large deviation principle with good rate function $I^\sharp$. If at every x with $I^\sharp(x) < \infty$, F is quasi-continuous and $I^\sharp$ is continuous, then $\{F(X^L)\}$ satisfies the LDP with rate function given by*

$$I(y) = \inf \left\{ I^\sharp(x) : y \in F^x \right\}. \tag{5.13}$$

Hence, given Assumption 5.5, the LDP's for the sequences of finite- and infinite-horizon workloads would follow as a direct consequence of the quasi-continuity of the mappings G_t and G. The quasi-continuity of the workload mappings is inherited from the quasi-continuity of the scheduler H.

5.3 Analysis: LDP's for Workloads

In this section, we present the main result of the chapter: LDP's for the sequences of the finite- and infinite-horizon workloads. We first establish an LDP for the sequence of the finite-horizon workloads.

5.3.1 LDP for Finite-Horizon Workloads

In this section, for $t \in \mathbb{N}$, we establish an LDP for finite-horizon workloads $\{Q^L_{0,t} := G_t(A^L_{(0,t]})\}$. The approach is to first show that the mapping $G_t : \mathbb{R}_+^{K\times t} \mapsto \mathbb{R}_+^K$ is quasi-continuous, then use Garcia's extended contraction principle to obtain an LDP for the finite-horizon workloads from the LDP assumption for $\{A^L_{(0,t]}\}$.

Lemma 5.12. *For $t \in \mathbb{N}$, G_t is quasi-continuous on $\mathbb{R}_+^{K\times t}$ with respect to the uniform topology.*

Proof. See Appendix. The idea of the proof relies on the quasi-continuity of the scheduler H and the linear dependence of the workload Q_s at time $-s$ on A_{s+1} for all $s \in (0, t-1]$. □

Now, as already discussed, the proof of Theorem 5.1 is complete. We refer to the corresponding rate function, I_t, as the finite-horizon rate function. Next, we discuss the LDP for the infinite-horizon workloads.

5.3.2 LDP for Infinite-Horizon Workloads

In this section, we establish an LDP of the sequence of the infinite-horizon workloads $\{Q^L = G(A^L)\}$ where $A^L \in \mathcal{D}^K$. Similar to the last section, we first show that the mapping G is quasi-continuous on $\mathcal{D}^K_\mu$ when $K\mu < c$, and then use Garcia's extended contraction principle to establish the desired LDP.

Lemma 5.13. *If $K\mu < c$, the mapping G is quasi-continuous on $\mathcal{D}^K_\mu$ with respect to the scaled uniform topology.*

Proof. See Appendix. The main idea is to use the fact that the sum (over all queues) workload process behaves like that of a single queue. □

Again, the above lemma and Garcia's extended contraction principle to the sequence of $\{A^L\}$ immediately give the LDP for the sequence of the infinite-horizon workload in Theorem 5.2. Recall that the set $\mathcal{D}^K_\mu$ contains all arrival sample paths a such that $I^\sharp(a) < \infty$ and $E[a^k_t] = \mu$ for all $k \in \mathcal{K}$ and $t \in \mathbb{N}$.

Let us now consider the problem of calculating the rate function. Eqn. (5.7) suggests that the rate function J, where $J(\mathbf{b}) = \inf_{a \in \mathcal{D}^K_\mu : G^a \ni \mathbf{b}} I^\sharp(a)$, could be interpreted as the minimum-cost solution among all paths $a \in \mathcal{D}^K_\mu$ such that $\mathbf{b} \in G^a$, where the cost of the path a is $I^\sharp(a)$ and convex. Hence, the problem of finding the rate functions is a deterministic optimal control problem like those in [67,75].

The expressions for the rate functions I_t and J in (5.6) and (5.7) are of little use in their current forms, as their computation is far from straight forward. In the next section, we simplify the rate functions when the arrival processes are limited to having i.i.d. increments.

5.4 I.I.D. Increments: Simplified Rate Functions

In this section, we give a calculation of the finite-horizon and infinite-horizon rate functions in the case when the arrivals have i.i.d. increments. In this case, the cost of a sample path $a \in \mathcal{D}^K$, which is $I^\sharp(a)$, is additive and the total cost of any arrival sample path is the sum of the cost over all timeslots and queues. This property helps us to simplify the calculation of the rate functions.

Consider the underlying arrival process A to be a process with i.i.d. increments, e.g., a compound Poisson arrival process with exponential packet length (see [42]). For these i.i.d. increment arrival processes, it is easy to show that for $\mathbf{x} \in \mathbb{R}_+^t$, $\Lambda_t^*(\mathbf{x}) = \sum_{i=1}^t \Lambda^*(x_i)$, where Λ^* is the Fenchel-Legendre transform of Λ and $\Lambda(\theta) = \log E \exp(\theta A_1)$ [29]. Hence, for a finite vector $\mathbf{a} = (a_i^k, k \in \mathcal{K}, i \in (0,t]) \in \mathbb{R}_+^{K \times t}$, the cost ${I^\sharp}_t(\mathbf{a})$ in (5.10) can be written as

$$
{I^\sharp}_t(\mathbf{a}) = \sum_{i=1}^{t} \mathcal{X}^A(\mathbf{a}_i), \tag{5.14}
$$

where we define $\mathcal{X}^A(\mathbf{x}) := \sum_{k=1}^K \Lambda^*(x^k)$, for $\mathbf{x} \in \mathbb{R}_+^K$, as the per-timeslot cost of a K-dimensional sample path. Next, we simplify the rate functions for the infinite-horizon and finite-horizon workloads, respectively.

5.4.1 Infinite-Horizon Rate Function

The following lemma expresses the infinite-horizon rate function J as the infimum of the finite-horizon rate functions I_t over all time t.

Lemma 5.14. *For i.i.d. increment arrivals and $K\mu < c$, the infinite-horizon rate function J is simplified as*

$$J(\mathbf{b}) = \inf_{t \geq 1} I_t(\mathbf{b}). \tag{5.15}$$

Proof. The cost of a sample path over time is the sum of the cost of arrivals in all timeslots. As in the proof of Lemma 5.13, for $a \in \mathcal{D}_\mu^K$ where $K\mu < c$, we can find $t := s^*(a)$ such that $Q_t(a) \in \mathcal{R}$. Hence, for a such that $\mathbf{b} \in G^a$, one can reduce the cost of the path by setting $a_v = \mu$ for all $v > t$ while keeping $G^a \ni \mathbf{b}$. This is because $\mathcal{X}^A(\mu \mathbf{1}_K) = 0$ and implies that $I^\sharp(a) = I^\sharp{}_t(a_{(0,t]})$. On the other hand, since $Q_t(a) \in \mathcal{R}$, we can write $\mathbf{b} \in G_t^{a_{(0,t]}}$. All of these imply that

$$J(\mathbf{b}) = \inf_{a \in \mathcal{D}_\mu^K : G^a \ni \mathbf{b}} I^\sharp(a) = \inf_{t \geq 1} \inf_{\mathbf{x} \in \mathbb{R}_+^{Kt} : G_t^{\mathbf{x}} \ni \mathbf{b}} I^\sharp{}_t(\mathbf{x}) = \inf_{t \geq 1} I_t(\mathbf{b}),$$

by the definition of $I_t(\mathbf{b})$ in (5.6). □

With this simplification available, we now look at the finite-horizon rate function I_t in more details.

5.4.2 Finite-Horizon Rate Function

In this subsection, we provide a further simplified expression of the finite-horizon rate function I_t.

Lemma 5.15. *For $t \in \mathbb{N}$, the finite-horizon rate function I_t is simplified as*

$$I_t(\mathbf{b}) = \min_{u \in (0,t]} \inf_{\mathbf{x} \in \mathbb{A}(u,\mathbf{b})} I^{\sharp}{}_u(\mathbf{x}) \tag{5.16}$$

for $\mathbf{b} \in \mathbb{R}_+^K$, where

$$\mathbb{A}(u,\mathbf{b}) := \left\{ a \in \mathbb{R}_+^{K \times u} : \mathbf{b} \in G_u^a, G_{u-v}(a_{(v,u]}) \notin \mathcal{R}, \forall v \in [1, u-1] \right\}. \tag{5.17}$$

Proof. This follows the idea from the proof of Lemma 5.14. Let $t \in \mathbb{N}$. For $a \in \mathbb{R}_+^{K \times t}$ such that $\mathbf{b} \in G_t^a$, we let

$$u = \min \left\{ t, \min\{ s \in [1, t-1] : Q_s = G_{t-s}(a_{(s,t]}) \in \mathcal{R} \} \right\}.$$

In other words, $-u$ is the last time the workload vector is inside the capacity region $\mathcal{R}$ before time 0. By definition of I_t, we already know that the workload vector starts initially inside $\mathcal{R}$ at time $-t$. With this definition of u, we have $Q_v \notin \mathcal{R}$ for all $v \in [1, u-1]$. We can find another path $\tilde{a} \in \mathbb{R}_+^{Kt}$ with a reduced cost while keeping the workloads at time $-u+1$ to 0 (i.e., Q_{u-1} to Q_0) intact by setting $\tilde{a}_v = \mu \mathbf{1}_K, \forall v \in (u,t]$ and $\tilde{a}_v = a_v$ otherwise. Since $\mathcal{X}^A(\mu \mathbf{1}_K) = 0$, we have $I^{\sharp}{}_t(a) \geq I^{\sharp}{}_t(\tilde{a}) = I^{\sharp}{}_u(a_{(0,u]})$ and yet $\mathbf{b} \in G_u^{\tilde{a}_{(0,u]}} = G_t^{\tilde{a}}$. Since by definition $Q_v = G_{u-v}(a_{(v,u]})$ for $v \in [1, u-1]$, we have

$$\begin{aligned} I_t(\mathbf{b}) &= \inf_{\mathbf{x} \in \mathbb{R}_+^{Kt} : G_t^{\mathbf{x}} \ni \mathbf{b}} I^{\sharp}{}_t(\mathbf{x}) \\ &= \min_{u \in (0,t]} \inf_{\mathbf{x} \in \mathbb{R}_+^{Ku} : \mathbf{b} \in G^{\mathbf{x}}, G_{u-v}(\mathbf{x}_{(v,u]}) \notin \mathcal{R}} I^{\sharp}{}_u(\mathbf{x}) = \min_{u \in (0,t]} \inf_{\mathbf{x} \in \mathbb{A}(u,\mathbf{b})} I^{\sharp}{}_u(\mathbf{x}), \end{aligned}$$

where $\mathbb{A}(u, \mathbf{b})$ is defined as in (5.17). □

Remark 5.16. The above lemma reduces the set of feasible sample paths to the set $\mathbb{A}(u, \mathbf{b})$ for $u \in (0, t]$. It is interesting to note the property of the sample paths in this set. For any $\mathbf{x} \in \mathbb{A}(u, \mathbf{b})$, we have $\hat{Q}_0(\mathbf{x}) = \hat{\mathbf{x}}(0, u] - c(u-1) = \hat{\mathbf{b}}$, recalling that the $\hat{\cdot}$ notation is the sum over queues. There is no wastage of service capacity over the $u-1$ timeslots because $\forall v \in [1, u-1]$, $Q_v = G_{u-v}(\mathbf{x}_{(v,u]}) \notin \mathcal{R}$ and hence $\hat{Q}_v > c$. That is, any sample path $\mathbf{x} \in \mathbb{A}(u, \mathbf{b})$ has its sum of the arrivals over time $(0, u]$ and queues equal to $\hat{\mathbf{x}}(0, u] = \hat{\mathbf{b}} + c(u-1)$.

In addition, an immediate implication of Lemma 5.15 is that we can rewrite J in (5.7) as

$$J(\mathbf{b}) = \inf_{t \geq 1} I_t(\mathbf{b}) = \inf_{t \geq 1} \min_{u \in (0,t]} \inf_{\mathbf{x} \in \mathbb{A}(u,\mathbf{b})} I^{\sharp}{}_u(\mathbf{x}) = \inf_{t \geq 1} \inf_{\mathbf{x} \in \mathbb{A}(t,\mathbf{b})} I^{\sharp}{}_t(\mathbf{x}). \tag{5.18}$$

If we denote t^* as the optimizer of the last equation, then t^* is called the *critical timescale* (see [82]). It can be interpreted that t^* is the length of time which the buffers are most likely to take to fill from “empty” level (more precisely, anywhere within $\mathcal{R}$) to a given level $\mathbf{b}$.

Note that for fixed $u \in \mathbb{N}$, $\inf_{\mathbf{x} \in \mathbb{A}(u,\mathbf{b})} I^{\sharp}{}_u(\mathbf{x})$ is a optimization problem, with a convex cost function $I^{\sharp}{}_u(\cdot)$ and a set $\mathbb{A}(u, \mathbf{b})$ of feasible solutions .real $K(u-1)$-dimensional This problem is difficult to solve analytically. Since the cost function $I^{\sharp}{}_u(\mathbf{x}) = \sum_{i=1}^{u} \mathcal{X}^A(\mathbf{x}_i)$ is additive, a possible numerical method is the numerical backwards induction of dynamic programming. However, the method suffers from the curse of dimensionality and hence is not practical for large u and

b. Hence, we turn our attention to finding some simplified bounds of the rate functions. This can be done by employing the additivity and convexity of the rate function ${I^\sharp}_t$. Next we derive some bounds when $K = 2$ queues.

5.4.3 Properties of the Minimum-Cost Sample Paths

Here, we see that the convexity of the cost function Λ^* induces two properties for the optimal paths.

Property 5.17. *Constant-speed linear path is the cheapest.* Among all arrival sample paths $\mathbf{x} \in \mathbb{R}^t_+$ to a queue in an interval of t timeslots, with the only constraint is to reach a common end point $\mathbf{x}(0,t] = d$ at the end of time t, the cheapest or minimum-cost path is the constant-speed linear path, where the arrival in each timeslot is equal to d/t.

Proof. This is because the path cost function is additive, i.e., $\Lambda^*_t(\mathbf{x}) = \sum_{i=1}^t \Lambda^*(x_i)$, and the per-timeslot cost function Λ^* is convex. Applying Jensen's inequality [64] gives

$$\Lambda^*_t(\mathbf{x}) = \sum_{i=1}^{t} \Lambda^*(x_i) \geq t\Lambda^* \left(\frac{1}{t} \sum_{i=1}^{t} x_i \right) = t\Lambda^*(d/t),$$

with equality when $x_i = d/t$ for all i. See an illustration in Figure 5.1(a). □

Property 5.18. *Constant-speed linear path closest to the equal line is the cheapest.* For $K = 2$, among constant-speed linear paths $\mathbf{a} \in \mathbb{R}^{Kt}_+$ with the sum $\hat{\mathbf{a}}(0,t] = d$, lying on the line perpendicular to the equal line, the path with destination closer to the equal line has a cheaper cost.

Proof. Since the arrival paths are constant-speed linear path, without loss of generality we can consider arrival paths in a single timeslot. Consider path $\mathbf{x} = (x, d-x) \in \mathbb{R}_+^2$ and $\mathbf{y} = (y, d-y) \in \mathbb{R}_+^2$, where $y > x > d/2$. That is, $\mathbf{x}$ is closer to the equal line (closer to the point $(d/2, d/2)$) than $\mathbf{y}$ does. The costs of paths $\mathbf{x}$ and $\mathbf{y}$ are $\Lambda^*(x) + \Lambda^*(d-x)$ and $\Lambda^*(y) + \Lambda^*(d-y)$, respectively. By convexity of Λ^*, we have $\Lambda^*(x) + \Lambda^*(d-x) \leq \Lambda^*(y) + \Lambda^*(d-y)$, and hence, $\mathbf{x}$ is cheaper than $\mathbf{y}$. □

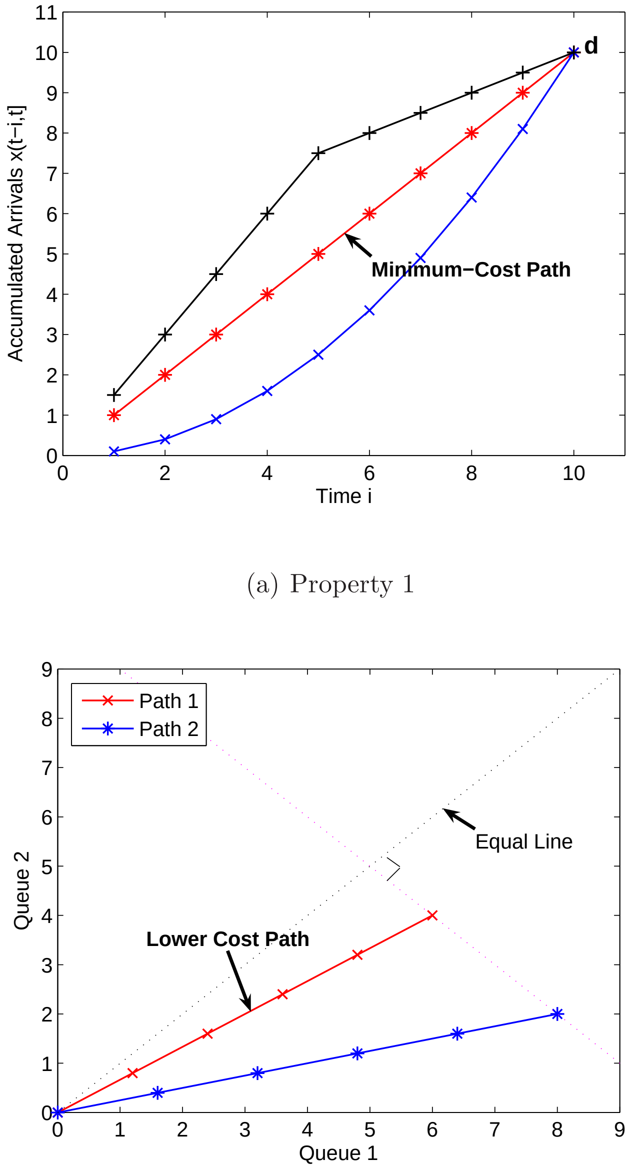
These properties are also used in [9, 67, 88] for large-deviations analysis of scheduling disciplines. Next, we use these properties to calculate I_2 and bounds on I_t for $t \in \mathbb{N}$.

5.4.4 Example: Calculation of I_2

Here we look at an example for calculation of the finite-horizon rate function I_t to illustrate that the calculation is rather involved. For simplicity, we consider the case when $t = 2$ and $K = 2$ queues. From (5.16), $I_2(\mathbf{b})$ for $\mathbf{b} \in \mathbb{R}_+^2$ can be written as

$$I_2(\mathbf{b}) = \min\left\{\mathcal{X}^A(\mathbf{b}), \inf_{(\mathbf{x}_1,\mathbf{x}_2)\in\mathbb{A}(2,\mathbf{b})} \mathcal{X}^A(\mathbf{x}_1) + \mathcal{X}^A(\mathbf{x}_2)\right\}, \tag{5.19}$$

where $\mathbb{A}(2,\mathbf{b}) = \left\{(\mathbf{a}_1,\mathbf{a}_2) \in \mathbb{R}_+^4 : \mathbf{a}_2 \notin \mathcal{R}, \mathbf{b} \in G_2^{(\mathbf{a}_1,\mathbf{a}_2)}\right\}$. The workload at time zero is $\mathbf{q}_0 = G_2(\mathbf{a}_1,\mathbf{a}_2) = \mathbf{a}_1 + [\mathbf{a}_2 - H(\mathbf{a}_2)]^+$, which is equal to $\mathbf{a}_1 + \mathbf{a}_2 - H(\mathbf{a}_2)$ since $\mathbf{a}_2 \notin \mathcal{R}$. On the other hand, we require $\mathbf{q}_0 = \mathbf{b}$. Hence, using the scheduler H given in (5.3), we can express $\mathbb{A}(2,\mathbf{b})$ as $\mathbb{A}(2,\mathbf{b}) = \mathbb{A}_{(1)} \cup \mathbb{A}_{(2)} \cup \mathbb{A}_{(3)}$, where



(a) Property 1

(b) Property 2

Figure 5.1 Two properties of the minimum-cost sample paths: (a) Property 1: the minimum-cost path is the constant-speed linear path. (b) Property 2: Path 1 which is closer to the Equal Line has a lower cost than Path 2.

$\mathbb{A}_{(j)} \subseteq \mathbb{R}^4_+, j = 1, 2, 3$, are defined as

$$\mathbb{A}_{(1)} := \{(\mathbf{a}_1, \mathbf{a}_2) \in \mathbb{R}^4_+ : a^1_2 \geq a^2_2, a^1_2 \geq c, \mathbf{a}_1 + \mathbf{a}_2 = \mathbf{b} + (c, 0)\}$$

$$\mathbb{A}_{(2)} := \{(\mathbf{a}_1, \mathbf{a}_2) \in \mathbb{R}^4_+ : a^1_2 \leq a^2_2, a^2_2 \geq c, \mathbf{a}_1 + \mathbf{a}_2 = \mathbf{b} + (0, c)\}$$

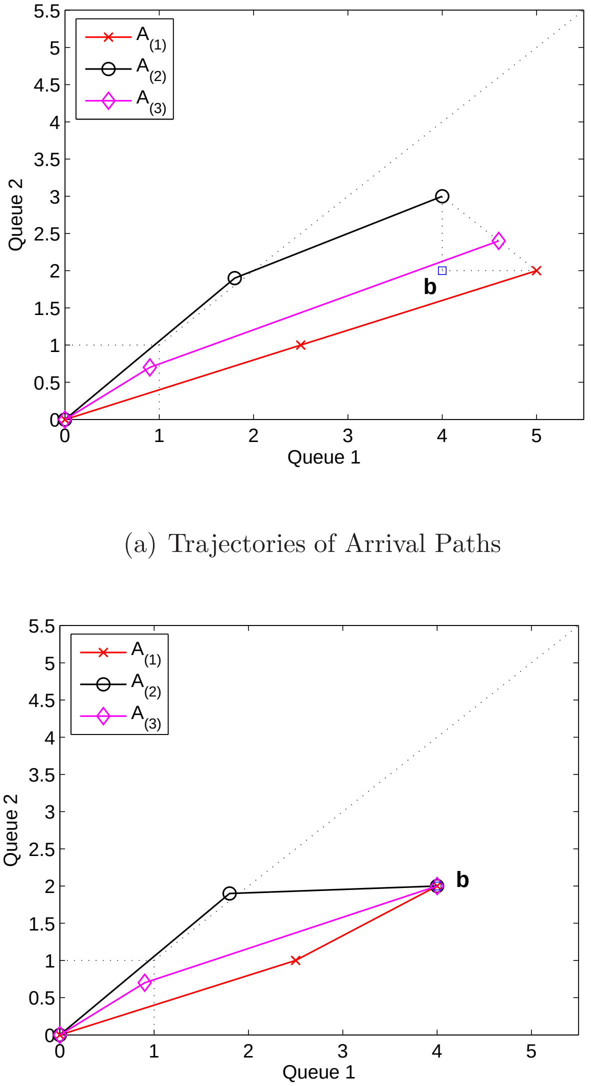
$$\mathbb{A}_{(3)} := \{(\mathbf{a}_1, \mathbf{a}_2) \in \mathbb{R}^4_+ : a^1_2 \leq c, a^2_2 \leq c, a^1_2 + a^2_2 \geq c, \mathbf{a}_1 + \mathbf{a}_2 = \mathbf{b} + \mathrm{Proj}_{\mathcal{R}}(\mathbf{a}_2)\}.$$

Hence, the second term in the RHS of (5.19) can be rewritten as

$$\inf_{\mathbf{a} \in \mathbb{A}(2,\mathbf{b})} \mathcal{X}^A(\mathbf{a}_1) + \mathcal{X}^A(\mathbf{a}_2) = \min_{j \in [1,3]} \inf_{\mathbf{a} \in \mathbb{A}_{(j)}} \mathcal{X}^A(\mathbf{a}_1) + \mathcal{X}^A(\mathbf{a}_2).$$

Trajectories of some examples of the (accumulated) arrival sample paths are illustrated in Figure 5.2(a) and their corresponding workload trajectories in Figure 5.2(b). Figure 5.2(a) shows example trajectories of the accumulated arrival sample paths $A_{(j)} \in \mathbb{A}_{(j)}, j = 1, 2, 3$, in the calculation of $I_2(\mathbf{b})$, where $\mathbf{b} = (4, 2)$, $c = 1$. For example, when $j = 1$, $A_{(1)} = (\mathbf{a}_{(1),1}, \mathbf{a}_{(1),2}) \in \mathbb{A}_{(1)}$ shows that $\mathbf{a}_{(1),2} = (2.5, 1)$ and $\mathbf{a}_{(1),1} + \mathbf{a}_{(1),2} = (5, 2)$. Figure 5.2(b) shows the workload paths $\mathbf{q}^{(j)}$ corresponding to the arrival path $A_{(j)}, j = 1, 2, 3$. For example, the figure shows that $\mathbf{q}^{(1)}_1 = \mathbf{a}_{(1),2} = (2.5, 1)$ and $\mathbf{q}^{(1)}_0 = \mathbf{a}_{(1),1} + \mathbf{a}_{(1),2} - (c, 0) = (4, 2)$.

This example underlines the difficulty in finding the rate function even for small timescales. We expect that the number of constrained sets like $A_{(j)}$ will grow exponentially with time duration t. However, the example gives us some insight on how to find some simple upper and lower bounds of I_t for any $t \in \mathbb{N}$.



(a) Trajectories of Arrival Paths

(b) Trajectories of Workload Paths

Figure 5.2 Example of accumulated arrival and workload paths for calculation of $I_2(\mathbf{b})$.

5.4.5 Bounds on I_t

In this subsection, we find simple expressions that give lower or upper bounds of $\inf_{\mathbf{x}\in\mathbb{A}(u,\mathbf{b})} I^{\sharp}{}_u(\mathbf{x})$, which in turn give the bounds on I_t and J. We focus on $K=2$ but similar result can be obtained for general K.

Lemma 5.19. *For $K=2$, $\mathbf{b}\in\mathbb{R}_+^2$, $I_t(\mathbf{b})$ can be bounded as*

$$I_t(\mathbf{b}) \geq \min_{u\in(0,t]} u\mathcal{X}\left(\frac{1}{u}Proj_{\mathbb{X}(u,\mathbf{b})}(\mathbf{0})\right) \tag{5.20}$$

and when $\mathbf{b}\notin[0,c)^2$,

$$I_t(\mathbf{b}) \leq \min_{u\in(0,t]} u\mathcal{X}\left(\frac{1}{u}(\mathbf{b}+(u-1)H(\mathbf{b}))\right), \tag{5.21}$$

where the convex set $\mathbb{X}(u,\mathbf{b})\subseteq\mathbb{R}_+^2$ is defined as

$$\mathbb{X}(u,\mathbf{b}) := \{\mathbf{b}+(v_1,v_2) : v_1+v_2=(u-1)c, v_1,v_2\geq 0\}. \tag{5.22}$$

Proof. Let $\mathbf{b}\in\mathbb{R}_+^2$, time $u\in(0,t]$, arrival path $\mathbf{a}\in\mathbb{A}(u,\mathbf{b})$, and $\mathbf{q}_i\in\mathbb{R}_+^2$ be the workload vector at time $-i$ for $i\in(0,u]$. We first show the lowerbound (5.20). As we have noted earlier that for $\mathbf{a}\in\mathbb{A}(u,\mathbf{b})$, the $[\cdot]^+$ function can be removed from the queue dynamics. Hence, we have $\mathbf{a}(0,u]=\mathbf{b}+\sum_{i=1}^{u-1}H(\mathbf{q}_i)$, where $\mathbf{q}_u\in\mathcal{R}$ and $\mathbf{q}_i\notin\mathcal{R}$ for all $i\in(0,u-1]$. Using this and the fact that $H(\mathbf{q}_i)\in\{(v_1,v_2) : v_1+v_2=c, v_1,v_2\geq 0\}$, for all $i\in(0,u-1]$, we have $\mathbf{a}(0,u]\in\mathbb{X}(u,\mathbf{b})$ where $\mathbb{X}(u,\mathbf{b})$ is defined above. Now, given any point $\mathbf{d}\in\mathbb{X}(u,\mathbf{b})$, the constant-speed linear path with increments of $\mathbf{d}/u$ is the minimum-cost path among all the paths with the same destination (using Property 1). In addition, among all the paths to destinations in $\mathbb{X}(u,\mathbf{b})$, the closest constant-speed linear paths $\mathbf{a}^*$ to the equal line is the minimum-cost path (using Property 2). Since the closest point in $\mathbb{X}(u,\mathbf{b})$ to the equal line is $\text{Proj}_{\mathbb{X}(u,\mathbf{b})}(\mathbf{0})$, we have $\mathbf{a}^* = (\mathbf{a}_i^* = \frac{1}{u}\text{Proj}_{\mathbb{X}(u,\mathbf{b})}(\mathbf{0}), i \in (0,u])$. Since the set of paths with destination in $\mathbb{X}(u,\mathbf{b})$ includes all paths in $\mathbb{A}(u,\mathbf{b})$, from (5.6) we have the lowerbound (5.20):

$$\begin{aligned}
I_t(\mathbf{b}) &= \min_{u\in(0,t]} \inf_{\mathbf{x}\in\mathbb{A}(u,\mathbf{b})} \sum_{i=1}^{K} \mathcal{X}^A(\mathbf{x}_i) \\
&\geq \min_{u\in(0,t]} \inf_{\mathbf{x}\in\mathbb{R}_+^{Kt}:\mathbf{x}(0,u]\in\mathbb{X}(u,\mathbf{b})} \sum_{i=1}^{K} \mathcal{X}^A(\mathbf{x}_i) \\
&= \min_{u\in(0,t]} u\mathcal{X}\left(\frac{1}{u}\text{Proj}_{\mathbb{X}(u,\mathbf{b})}(\mathbf{0})\right).
\end{aligned}$$

To show the upperbound (5.21), we only need to show that the constant-speed linear path $\mathbf{a} = (\mathbf{a}_i = \frac{1}{u}(\mathbf{b} + (u-1)H(\mathbf{b})), i \in (0,u])$, is in $\mathbb{A}(u,\mathbf{b})$, when $\mathbf{b} \notin [0,c)^2$. Without loss of generality, we consider only when $b^1 \geq b^2$ and $b^1 \geq c$. In this case, $H(\mathbf{b}) = (c,0)$ and the queue dynamics gives

$$\mathbf{q}_i = \frac{(u-i)}{u}(\mathbf{b} + (u-1)(c,0)) - (u-1-i)(c,0),$$

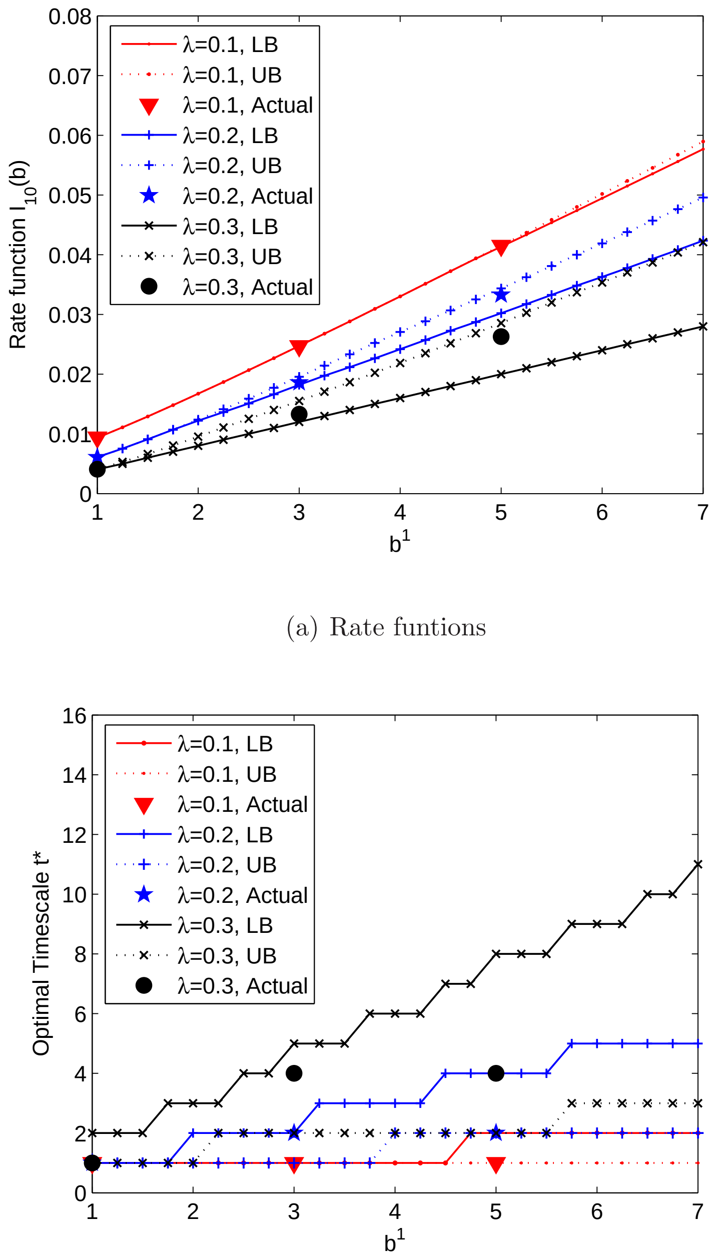
for all $i \in (0,u-1]$. Since $b^1 \geq c$, we have $q_i^1 \geq c$ and $q_i^1 \geq q_i^2$, and hence $H(\mathbf{q}_i) = (c,0)$ for all $i \in (0,u-1]$. Hence, $\mathbf{a} \in \mathbb{A}(u,\mathbf{b})$). □

Next we look at the tightness of the above bounds for an example of compound Poisson source process with exponential packet size. We expect the tightness to depend on the traffic loads.

5.4.6 Comparison of the Bounds: Numerical Examples

Here we illustrate the tightness of the bounds given in Lemma 5.19, via an example of compound Poisson source process with exponential packet size (the CPE process with $g(N) = N$, described in Section 3.1.2). Let the average packet arrival rate denoted by λ and the average packet size denoted by $1/\mu$. The function Λ^* for this process is simple and given as in (3.18). Figure 5.3(a) shows the upper and lower bounds and the actual values of I_t, for $t = 10$, at $\mu = 0.01$, $c = 1$, and various values of $\lambda = 0.1, 0.2, 0.3$ and when $\mathbf{b} = (b^1, b^2 = 1)$ for various values of b^1. Figure 5.3(b) shows the corresponding minimizing t^* for the bounds and the actual expression of I_t. We note that for all $\mathbf{b}$ in this example, $J(\mathbf{b})$ is actually equal to $I_t(\mathbf{b})$ for $t = 10$ since all optimizing t^* is less than 10 (see (5.18)). This example shows that, in the range of $\mathbf{b}$ in consideration, both bounds are tight and almost coincide when the traffic load is small, i.e., $\lambda = 0.1$. However, when the traffic load is higher, the lowerbound becomes loose while the upperbound is still considerably tight.

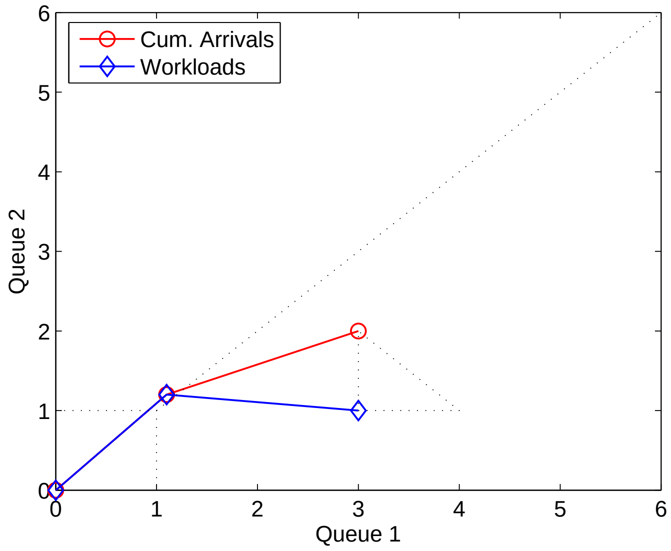
It is interesting to note the optimal timescale t^* which the queues most likely to take to reach the level $\mathbf{b}$. Figure 5.3(b) shows that, for example, it is most likely to take only two timeslots for CPE process with $\lambda = 0.2$ to reach the buffer level $\mathbf{b} = (3, 1)$, while the most likely timescale is four timeslots when the traffic load is higher ($\lambda = 0.3$). Figure 5.3(c) and Figure 5.3(d) show the optimal trajectories of the accumulated arrival process and the workload process for $\lambda = 0.2$ and 0.3, respectively.



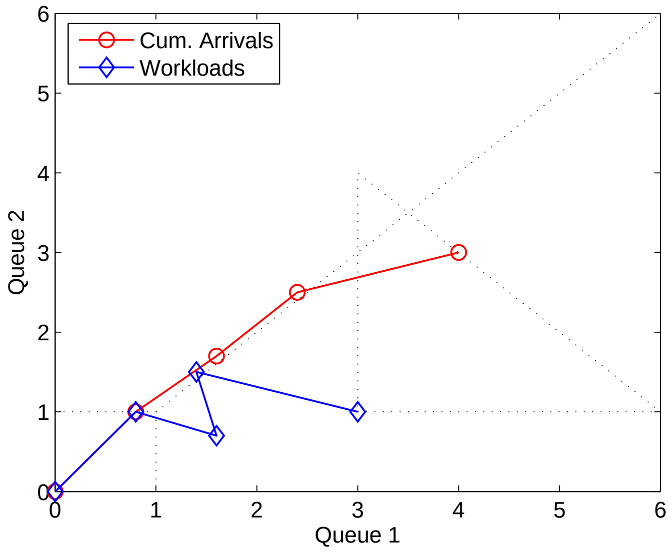
(a) Rate funtions

(b) Optimal Timescales

Figure 5.3 Example of the rate function $I_{10}(\mathbf{b})$ and its upper and lower bounds, and their corresponding optimizing t^* and optimal trajectories, when $\mathbf{b} = (b^1, b^2 = 1)$.



(c) Optimal Trajectory for $\mathbf{b} = (3, 1)$ and $\lambda = 0.2$



(d) Optimal Trajectory for $\mathbf{b} = (3, 1)$ and $\lambda = 0.3$

Figure 5.3 Example of the rate functions, continued.

5.5 Summary

In this chapter, we have established a many-sources LDP for the stationary (infinite-horizon) workload for multi-queue single-server system with simplex capacity, operated under the maximum-weight scheduling with the arrival processes assumed to satisfy a many-sources sample path LDP. To extend the LDP of the arrival processes to the LDP of the workloads, we employed Garcia's extended contraction principle, which applies to quasi-continuous mappings. Along the way, we also establish an LDP for the finite-horizon workload. We gave the associated rate functions and the expression of the infinite-horizon rate function in term of the finite-horizon ones, when the arrivals processes have i.i.d. increments.

Acknowledgment

This chapter, in part, appears in the following publication. The dissertation author was the primary investigator and author of this paper.

- S. Kittipiyakul, T. Javidi, and V. G. Subramanian, "Many-sources large deviations for max-weight scheduling," to appear in the *46th Annual Allerton Conference on Communication, Control, and Computing (Allerton'08)*.

I would like to thank my coauthors Dr. Vijay G. Subramanian and Prof. Tara Javidi.

Chapter 6

Conclusions

In this work, we have examined several resource allocation problems for wireless data communications of delay-sensitive and bursty data. We considered the scenario where the channel state information at the transmitter (CSIT) is available and the scenario when it is not. In the first scenario, the knowledge of the channel states, together with the queue state information, can be utilized to improve the packet delay performance of the system. In particular, in this work we studied the problem of delay-optimal subcarrier allocation problem in OFDMA downlink systems. On the other hand, in the second scenario when the CSIT is not available, the question of interest is how to select the optimal operating PHY parameters, when the network layer performance (delay violation probability), as well as the physical layer performance (channel decoding error probability), are jointly considered. Given the knowledge of the source and channel statistics, we studied how to set up the systems so that they operate at their best performance in term of the asymptotic high-SNR total error probability. We considered both single-user setting and multi-user setting with dynamic queue-aware rate schedulers. As there are always extensions to any piece of work, below we provide some suggestions for future extensions.

6.1 Future Work

6.1.1 Future Work for Chapter 2

In Chapter 2, we used a dynamic programming approach to show that the MTLB server allocation policy is delay-optimal when the connectivities follow a binary on-off model. We established this result for $N = 2$ users. The difficulty in showing the optimality for general N lies in the constraint that the server allocation must be integral, i.e., a server can be assigned to only one queue. When this constraint is relaxed to a fluid allocation constraint, i.e., any server can be assigned to serve multiple queues as long as it serves no more than one packet in total, we showed that the MTLB-F policy, which is the fluid version of the MTLB policy, is optimal for the general N case. Some interesting and important extensions are as follow.

Extension 6.1. Extension to general N with the integral server allocation constraint. Form the above results, an interesting extension is a verification of the following conjecture:

Conjecture 6.2. *In the case of $N > 2$ statistically symmetric users with the on-off* *channel connectivity and the integral server allocation constraint, the MTLB policy is optimal.*

The conjecture seems intuitive due to the assumed symmetry of the users and the convexity of the cost function. The conjecture has indeed been established in several special cases. For the case of single server and Bernoulli arrivals, Tassiulas and Ephremides [76] proved the optimality of the LCQ policy, which coincides with the MTLB policy for $K = 1$ server. For the case of multiple servers and Bernoulli arrivals but with vector connectivity (i.e., each user is either connected to all servers or none) and the integral server allocation constraint, the LCQ policy (a more generalized multi-queue version of the LCQ policy in [76]) is optimal [31]. This LCQ policy serves the longest connected queues – this is equivalent to the MTLB policy with the integral server allocation constraint. It is interesting to note the complimentary roles of the stochastic coupling and majorization techniques used in [76] and [31] and the dynamic programming technique we employed in this chapter. These roles, in nature, are closely related to the results discussed in [50].

Extension 6.3. Generalization to channel connectivity models more general than the binary model, as well as to heterogeneous users. This extension is of practical concern, but it is challenging to find an optimal policy and even more to establish its optimality. Although we have proposed several heuristic algorithms that perform well at various traffic loads in the case of homogeneous users, the question of optimal scheduling policy remains open.

6.1.2 Future Work for Chapter 3

In Chapter 3, our analysis assumed no feedback, no retransmission, and a static allocation of the multiplexing rate r and coding block length T. Some interesting extensions, especially when these assumptions are relaxed, are as follow:

Extension 6.4. Improvement of the system performance by adjusting r and T according to the current queue length. For example, when the queue length is short, it seems intuitive that reducing T improves the channel error performance, possibly at the cost of longer delays of the bits that arrive later. However, since in high-SNR analysis the probability of error is asymptotically dominated by the worst case probability, it is not clear whether such adaptive mechanism will improve the asymptotic channel error performance.

Extension 6.5. Allowing retransmission mechanism. With retransmission, the diversity of the channel can be improved considerably [22] but at the cost of random transmission delays.

Extension 6.6. Fine-tuning the high-SNR asymptotic analysis for the regime of finite SNR. Our result focuses on the asymptotic high-SNR approximation and the notion of SNR error exponent as a measure of performance. This view of communication systems provides a tractable and intuitive characterization of various suggested schemes in the high-SNR regime. It would be interesting to fine-tune the analysis and verify the results via simulations when SNR is finite.

6.1.3 Future Work for Chapter 4

In Chapter 4, we bounded the delay violation probability by the performances of a static rate scheduler and a largest-delay-first rate scheduler with simplex rate region. Some suggestions that can extend the utilization of the system resources in other aspects are as follow:

Extension 6.7. Time-diversity: It is interesting to extend our study to include the time-diversity, as a result of either coding over multiple *fixed* coherence times as in Chapter 3 or coding over *random* coherence times as in hybrid ARQ (see [22, 37, 38]). This extension requires the asymptotic high-SNR delay violation probability that must be valid for finite and small delay bound D. Although the asymptotic buffer overflow probability result we developed in Chapter 5 could be used in the multi-user study of time-diversity when the delay violation probability is replaced with the buffer overflow probability, a direct analysis in the delay performance like that in [72] would be more beneficial in highlighting the effect of different coding block durations.[1]

Extension 6.8. Cooperative multiple-access channel: As mentioned in Section 3.4.2, cooperation among users can substantially improve the reliability of communication by providing a form of virtual MIMO communications [65], but at the cost of additional delays. It is interesting to extend the single-user study in Section 3.4.2

[1] We note that an analysis on the delay performance in the multi-user setting with dynamic scheduler is challenging since in this case the delay that a bit will experience in the system is not simply related to the queue length the bit sees upon its arrival. This is unlike the single-user case in Chapter 3, where the delay can be immediately calculated from the queue length because the server capacity to the queue is always fixed.

to the multi-user setting of cooperative multiple-access channel (see [3]). However, such extension requires a DMT tradeoff result for this channel.

6.1.4 Future Work for Chapter 5

Finally, in Chapter 5, we analyzed the LDP of the stationary workload processes of the maximum-weight scheduling policy when the rate region $\mathcal{R}$ is simplex. To establish the LDP of the stationary workload process, we showed the quasi-continuity of the stationary workload process mapping G. Some suggestions for future works are as follow:

Extension 6.9. Extension to the case where $\mathcal{R}$ is any compact and convex region. This extension will complement the recent result by Subramanian in [74] which studies the LDP of the *finite-horizon* workload process in the large-buffer framework. Assuming the following conjecture, the main difficulty in establishing the LDP for the *infinite-horizon* workload for general $\mathcal{R}$ seems to lie in showing the quasi-continuity of the mapping G as in Lemma 5.13, in particular, in establishing a similar result as in Claim D.4.

Conjecture 6.10. *Given any convex and compact region $\mathcal{R}$, there exists a quasi-continuous, max-weight scheduling function H. With H being quasi-continuous, the finite-horizon workload mapping G_t, for $t \in \mathbb{N}$, is quasi-continuous.*

We believe that the first part on H in the above conjecture is not too difficult to verify because the maximum-weight type schedulers are invariant with respect to scalings of the queues, i.e., a scaling of all queue lengths by the same factor at any given time does not change the scheduling choice.[2] For the second part about G_t, we believe that this is an immediate result from the quasi-continuity of H (see the proof of Lemma 5.12).

Extension 6.11. Consideration of non-i.i.d. arrivals. In this chapter we characterized the rate functions in the case of i.i.d. bit-arrival processes. An interesting extension is to consider non-i.i.d. (correlated in time) arrivals, which are of practical interest since realistic traffic streams, such as video, usually have bursts of arrivals which are correlated in time. For these processes, we expect that the characterization of the rate functions in the multi-user setting could benefit from the results and proof techniques already established in the single-user setting (e.g., see [82]).

[2] See more discussion of this property in [73] which considers a large-deviations analysis of the Exponential (EXP) scheduler, which is not invariant to the scalings of the queues.

Appendix A

Appendix for Chapter 2

A.1 Existence Proof of MTLB: Proof of Theorem 2.1

In this section, we prove Theorem 2.1 and Proposition 2.9 using the notions of alternating, balancing, and throughput-increasing paths, which are the concepts taken from graph literature [35]. We note that some of the results here are useful in the next appendices. Note that the discussions in this section (both the existence proof as well as the construction of the MTLB allocation) are valid for general N.

A.1.1 Alternating, Balancing, and Throughput-Increasing Paths

For convenience and simplicity of the proofs, we adopt the language of graph literature. Let $G = (V, U, C)$ be a bipartite graph with U the set of queues, V the set of servers, and edge set $C \subseteq V \times U$ representing the set of connectivities between queues and servers. Each allocation matrix $W = [w_{i,j}]$ can be thought of an edge set where an edge $(v, u) \in W$ if $w_{v,u} = 1$. Hence, an edge set W is a (feasible) allocation (i.e., $W \in \mathcal{W}(C)$) if each vertex $v \in V$ is incident with exactly one edge in C. Furthermore, W is non-idling (i.e., $W \in \mathcal{W}(\mathbf{b}, C) \subseteq \mathcal{W}(C)$) if W is feasible and $\sum_{i=1}^{K} w_{iu} \leq b_u$ for each $u \in U$.

Definition A.1. A vertex in V that is incident to any edge in the allocation is called *matched*, and *unmatched* otherwise. A queue or vertex u in U with $b_u - \sum_{i=1}^{K} w_{iu} > 0$ is called *non-empty*, and *empty* otherwise.

Definition A.2. For a given allocation $W \subseteq \mathcal{W}(\mathbf{b}, C)$ in G, an *alternating path* $S(W, u_0, u_k)$ with respect to W is a sequence of edges with distinct vertices,

$$S(W, u_0, u_k) := \{(v_1, u_0), (v_1, u_1), (v_2, u_1), \ldots, (v_k, u_k)\}, \tag{A.1}$$

from a queue $u_0 \in U$ to a queue $u_k \in U$, through matched servers, with $v_i \in V$, $u_i \in U$, $(v_i, u_{i-1}) \in C \backslash W$, and $(v_i, u_i) \in W$ for each $i = 1, \ldots, k$.

Definition A.3. An alternating path $S(W, u_0, u_k)$ is called a *balancing path* if it satisfies $b_{u_0} - \sum_{i=1}^{K} w_{i,u_0} \geq b_{u_k} - \sum_{i=1}^{K} w_{i,u_k} + 2$.

For convenience, we treat paths as a sequence of vertices. For example, we write

$$S(W, u_0, u_k) = (u_0, v_1, u_1, v_2, \ldots, v_k, u_k)$$

to show $S(W, u_0, u_k)$ as a sequence of vertices alternatively taken from U and V starting from u_0 and ending at u_k.

Definition A.4. A *throughput-increasing path* relative to W, from an unmatched server $v_0 \in V$ to a non-empty queue $u_k \in U$, is a sequence of distinct vertices (or equivalently, a sequence of edges)

$$I(W, v_0, u_k) := (v_0, u_1, v_1, u_2, \ldots, v_{k-1}, u_k)$$

with $v_i \in V$, $u_i \in U$, $(v_{i-1}, u_i) \in C \backslash W$, and $(v_i, u_i) \in W$ for each i.

Definition A.5. For the alternating path $S = S(W, u_0, u_k)$ given in (A.1), $W^a(S)$ is the *alternating allocation* of the allocation W along an alternating path S if server v_l is reassigned to serve queue u_{l-1}, $\forall l = 1, \ldots, k$. If, in addition S is a balancing path, then $W^a(S)$ is specifically called the *balancing* allocation and denoted by $W^b(S)$. In a similar fashion, $W^t(I)$ is called the *throughput-increasing* allocation if I is a throughput-increasing path and $W^t(I)$ assigns server v_l to serve queue u_{l+1}, $\forall l = 0, \ldots, k-1$. Equivalently, we can write $W^a(S) = W \oplus S$, $W^b(S) = W \oplus S$, and $W^t(I) = W \oplus I$, where $A \oplus B := (A \backslash B) \cup (B \backslash A)$ for any sets A, B.

An example of some alternating path and alternating allocation is shown in Figure A.1. It is easy to see that $W^a(S(W, u_0, u_k)) \in \mathcal{W}(\mathbf{b}, C)$ if u_0 is non-empty under W. Obviously, $W^b(S) \in \mathcal{W}(\mathbf{b}, C)$. Note that the alternating allocations (when u_0 is non-empty) and the balancing allocations leave the cardinality of the allocation (i.e., throughput) unchanged, while throughput-increasing allocations increase the current throughput by one. In addition, when u_0 is non-empty under W, the allocations W and $W^a(S = S(W, u_0, u_k))$ result in the leftover queues that are identical except for queues u_0 and u_k. In other words, if we denote the leftover queues under W and $W^a(S)$ as $\mathbf{l} = \mathbf{b} - \mathbf{1}W$ and $\mathbf{l}^a = \mathbf{b} - \mathbf{1}W^a(S)$, respectively, then $l^a_{u_0} = l_{u_0} - 1$, $l^a_{u_k} = l_{u_k} + 1$, and $l^a_u = l_u$, for all $u \in U \backslash \{u_0, u_k\}$.

Notice that the above notion of throughput-increasing path is conceptually related to the notion of the alternating path in the graph matching literature [57]. Likewise, our notion of balancing path is related to the notion of the cost-reducing path in [35] where cost is the "unbalancedness" of the queues.

A.1.2 Proof of Existence of MTLB Policy

The following Proposition is used to find the necessary and sufficient condition for policies to satisfy (**C1**) and show the existence of the MTLB policy (Theorem 2.1).

Proposition A.6. *An allocation achieves the maximum throughput (**C1**) if and only if it has no throughput-increasing paths.*

Proof. Obviously, if there is a throughput-increasing path for a given allocation W, then W does not achieve the maximum throughput. To show that not having any throughput-increasing paths is a sufficient condition for achieving the maximum

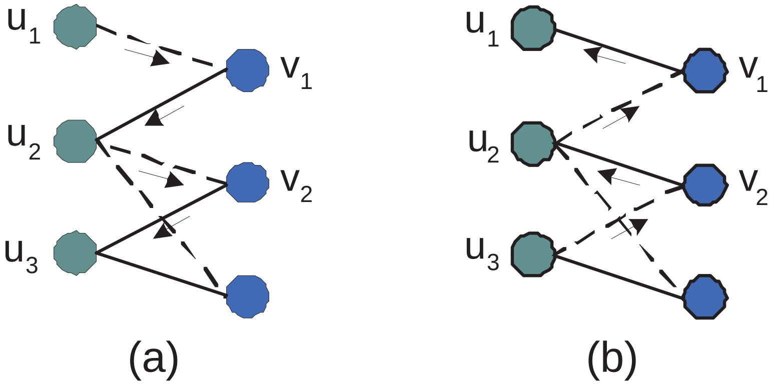


Figure A.1 Example of an alternating path and the alternating allocation from queue u_1 to queue u_3 (a) Alternating path $S = (u_1, v_1, u_2, v_2, u_3)$. The dotted and solid lines show the connectivities while the solid lines show the allocation. (b) The solid lines show the alternating allocation $W^a(S)$.

throughput, the proof follows the standard graph technique used in [57] which turns the problem into a maximum network flow problem. We refer interested readers to [57] for a more detailed proof. □

Theorem 2.1. *For any given* $(\mathbf{b}, C)$*, an MTLB allocation always exists.*

Proof. Without loss of generality, consider $(\mathbf{b}, C)$ such that $\mathcal{W}(\mathbf{b}, C) \neq \emptyset$. Let $\mathcal{W}^{\mathrm{MT}} = \mathcal{W}^{\mathrm{MT}}(\mathbf{b}, C) \subseteq \mathcal{W}(\mathbf{b}, C)$ contain all (maximum-throughput) allocations satisfying (**C1**) and $\mathcal{W}^{\mathrm{LB}} = \mathcal{W}^{\mathrm{LB}}(\mathbf{b}, C) \subseteq \mathcal{W}(\mathbf{b}, C)$ contain all (load-balancing) allocations satisfying (**C2**). Since there is a finite number of servers, clearly the maximum throughput is finite and there exists a maximum-throughput allocation. In other words, $\mathcal{W}^{\mathrm{MT}} \neq \emptyset$.

Now, we show that $\mathcal{W}^{\mathrm{LB}} \neq \emptyset$. That is, there must exist an allocation $W^{\mathrm{LB}} \in \mathcal{W}(\mathbf{b}, C)$ such that $l(W^{\mathrm{LB}}) \leq_{\mathrm{LQO}} l(W)$ for all $W \in \mathcal{W}(\mathbf{b}, C)$, where we let $l(W) := \mathbf{b} - \mathbf{1}W$ be the leftover queue vector under any allocation W. Since the $\leq_{\mathrm{LQO}}$ ordering is basically a lexicographic ordering, we can always rank any two elements in $\mathcal{W}(\mathbf{b}, C)$ using the $\leq_{\mathrm{LQO}}$ relation[1] and find the least element (not necessarily unique), say W^{LB}, such that $l(W^{\mathrm{LB}}) \leq_{\mathrm{LQO}} l(W)$ for all $W \in \mathcal{W}(\mathbf{b}, C)$.

Next, we show the existence of an MTLB allocation, satisfying (**C1**) and (**C2**), by showing that $\mathcal{W}^{\mathrm{LB}} \cap \mathcal{W}^{\mathrm{MT}} \neq \emptyset$. Using Proposition A.6, we observe that, for any allocation satisfying (**C2**) but not (**C1**), there would be idle servers that could have been assigned via some throughput-increasing allocations to serve more packets. With such allocations of the idle servers, the queues will be no less balanced than before. Thus, we have shown that there exists an MTLB allocation.

□

A.1.3 Necessary and Sufficient Condition for MTLB Policy

The following Proposition gives a necessary and sufficient condition for the MTLB policy. This result will be useful in the proof of the optimality of the MTLB policy in the next Appendix.

Proposition A.7. *Any allocation satisfying the maximum-throughput condition (**C1**) also satisfies the load-balancing condition (**C2**) if and only if it has no balancing path.*

[1] For any $W, W' \in \mathcal{W}(\mathbf{b}, C)$, either $l(W) \leq_{\mathrm{LQO}} l(W')$ or $l(W') \leq_{\mathrm{LQO}} l(W)$, with the exception that $l(W) \leq_{\mathrm{LQO}} l(W')$ and $l(W') \leq_{\mathrm{LQO}} l(W)$, when $l(W') = \pi(l(W))$ for some permutation π.

Proof. Without loss of generality, consider $(\mathbf{b}, C)$ such that $\mathcal{W}(\mathbf{b}, C) \neq \emptyset$. The *only if* part is obvious, i.e., if there is a balancing path S relative to an allocation $W \in \mathcal{W}(\mathbf{b}, C)$, then $\mathbf{b} - \mathbf{1}W^b(S) \leq_{\text{LQO}} \mathbf{b} - \mathbf{1}W$ but $\mathbf{b} - \mathbf{1}W \not\leq_{\text{LQO}} \mathbf{b} - \mathbf{1}W^b(S)$, i.e., W does not satisfy (**C2**).

What remains is the *if* part: if a maximum-throughput allocation does not satisfy (**C2**), then it has at least one balancing path. Let $W \in \mathcal{W}(\mathbf{b}, C)$ be a maximum-throughput allocation (satisfying (**C1**)) but is not the most balanced (not satisfying (**C2**)). We show that a balancing path relative to W must exist. Since at least one MTLB allocation exists (by Theorem 2.9), we let $W^* \in \mathcal{W}(\mathbf{b}, C)$ be an MTLB allocation. If more than one MTLB allocations exist, we pick W^* such that the number of edges in the symmetric difference $W^* \oplus W = (W^* \backslash W) \cup (W \backslash W^*)$ is minimized among all MTLB allocations. That is, W^* is the "closest" MTLB allocation to W, i.e., among all MTLB allocations, W^* requires the minimum number of servers to be reassigned to get to W. Now, let G_d be the subgraph of the bipartite graph $G = (V, U, C)$ induced by the edges of $W^* \oplus W$. Color the edges of $W^* \backslash W$ green and the edges of $W \backslash W^*$ red. Direct the green edges from V to U and the red edges from U to V. Let the leftover queue vectors under W and W^* be $\mathbf{l} = \mathbf{b} - \mathbf{1}W$ and $\mathbf{l}^* = \mathbf{b} - \mathbf{1}W^*$, respectively.

We claim that for every directed path P in G_d from $u_1 \in U$ to $u_2 \in U$, we have

$$l^*_{u_1} \leq l^*_{u_2}. \tag{A.2}$$

To see this, let $P = (u_1, \ldots, u_2)$ be a directed path in G_d. By the choice of the directions for the edges, P must be alternating between red and green edges. If $l^*_{u_2} < l^*_{u_1} - 1$ then P is a balancing path for W^*, and $\mathbf{b} - \mathbf{1}W^b(P) \leq_{\text{LQO}} \mathbf{b} - \mathbf{1}W^*$ but $\mathbf{b} - \mathbf{1}W^* \not\leq_{\text{LQO}} \mathbf{b} - \mathbf{1}W^b(P)$, contradicting to the assumption that W^* satisfy (**C2**). Similarly, if $l^*_{u_2} = l^*_{u_1} - 1$ then by alternating the assignment of the V-vertices along P, we can get another MTLB allocation W^{**} such that the number of edges in $W^{**} \oplus W$ is strictly less than that in $W^* \oplus W$, in contradiction to the choice of W^*. Hence, we must have that $l^*_{u_2} \geq l^*_{u_1}$. Using a similar argument, we can also show that G_d is acyclic.

Since both W^* and W achieve the maximum throughput, we have that $\sum_{i=1}^N l_i = \sum_{i=1}^N l^*_i$. But since W does not satisfy the LB condition (**C2**), there must exist $u_1 \in U$ such that

$$l_{u_1} < l^*_{u_1}. \tag{A.3}$$

Obviously, there is a red edge directed out of u_1. Starting from u_1 we build an alternating red-green path P' in G_d as follows: (1) From an arbitrary vertex $u \in U$ (including u_1), if there is a red edge directed out of u and $l_u \leq l_{u_1} + 1$, we build P' by arbitrarily select one of the red edges directed out of u. (2) From an arbitrary $v \in V$, we build P' by following the single green edge directed out of v. Such a green edge always exist.[2] (3) Otherwise, stop.

Using the fact that G_d is acyclic, P' is well-defined and finite. Let $u_2 \in U$ be the final vertex on the path. These are two possible cases:

Case 1: $l_{u_2} > l_{u_1} + 1$. In this case, we reverse the order of nodes in P' to

[2] Otherwise, combining the red edge coming into v with W^* would have yielded a non-idling feasible allocation with additional one packet throughput, a contradiction to (**C1**).

arrive at a balancing path relative to W.

Case 2: There is no red edge directed out of u_2. Thus, P' arrived at u_2 via a green edge. This means that u_2 is served at least one more packet under W^*, relative to W. Hence, $l_{u_2} > l^*_{u_2}$. This together with (A.2) and (A.3) give $l_{u_2} > l^*_{u_2} \geq l^*_{u_1} > l_{u_1}$, which means $l_{u_2} \geq l_{u_1} + 2$. Reversing the order of nodes in P' gives a balancing path relative to W.

Since there exists a balancing path in both cases, we have the assertion of the proposition. □

The above Proposition states that all MTLB allocations are such that they have no balancing path. The results in the proof that G_d is acyclic and has finite edges immediately implies the following result:

Corollary A.8. *The minimum number of balancing allocations required to turn any maximum throughput allocation satisfying (**C1**) to an MTLB allocation is finite.*

A.1.4 Proof of Proposition 2.9

The equivalence of the MWM matching on the equivalent bipartite graph and the MTLB allocation (Proposition 2.9 in Section 2.3.3) can be proved as follows:

Proof of Proposition 2.9. Since all weights are strictly positive, the MWM matching on the equivalent bipartite graph necessarily matches all possible servers and hence the equivalent allocation (defined in Definition 2.8) achieves the maximum-throughput condition (**C1**). We prove the load-balancing condition (**C2**) by contradiction. Suppose the maximum weight matching M results in the allocation W that achieves the maximum throughput but does not produce the most balanced queues. From Proposition A.7, we know that there must exist a balancing path $S(W, j, i)$ from some queue j to queue i such that $b_j - w_j \geq b_i - w_i + 2$, where $w_s = \sum_{m=1}^{K} w_{m,s}$ for $s = i, j$. Let us denote the balancing allocation of W along $S(W, j, i)$ as W^b. Let M^b be the equivalent matching to W^b. According to M, node a_{iw_i} is matched and $a_{j(w_j+1)}$ is not, while the reverse is true for M^b. In other words, $\psi(M^b) - \psi(M) = b_j - w_j - (b_i - w_i + 1) \geq 1$. But this is a contradiction to the assumption that M is the maximum weight matching on the equivalent bipartite graph. □

A.2 Supporting Lemmas for Theorem 2.2

In this appendix we establish the proofs for lemmas 2.15 to 2.24 stated in Section 2.4.3. The first lemma establishes the strict monotonicity of v_n for all $n = 0, \ldots, T$. Hence, v_n satisfies (**B.1**) for all n as well.

Lemma 2.15. *$v_n(\mathbf{b})$ is strictly increasing on $\mathbf{b}$ for all $n = 0, \ldots, T$, i.e., $\mathbf{b}' > \mathbf{b} \Rightarrow v_n(\mathbf{b}') > v_n(\mathbf{b})$.*

Proof. Since v_n is linearly related to V_n^* by (2.13), it suffices to show the strict monotonicity of $V_n^*(\mathbf{b}, C)$ for any C. We show by induction.

Induction Basis: $V_0^*(\mathbf{b}, C) = \phi(\mathbf{b}) = \sum_{j=1}^N \xi(b_j)$ is strictly increasing by the assumption of ξ.

Induction Step: Assume $V_{n-1}^*(\mathbf{b}', C) > V_{n-1}^*(\mathbf{b}, C)$ for any $\mathbf{b}' > \mathbf{b}$, then

$$\begin{aligned}
V_n^*(\mathbf{b}', C) &= \phi(\mathbf{b}') + \min_{W' \in \mathcal{W}(\mathbf{b}', C)} E_{\mathbf{a}, \tilde{\mathbf{C}}}[V_{n-1}^*(\mathbf{b}' - \mathbf{1}W' + \mathbf{a}, \tilde{C})] \\
&\geq \phi(\mathbf{b}') + \min_{W' \in \mathcal{W}(\mathbf{b}', C)} E_{\mathbf{a}, \tilde{\mathbf{C}}}[V_{n-1}^*(\mathbf{b} - \mathbf{1}W(W') + \mathbf{a}, \tilde{C})] \\
&\geq \phi(\mathbf{b}') + \min_{W \in \mathcal{W}(\mathbf{b}, C)} E_{\mathbf{a}, \tilde{\mathbf{C}}}[V_{n-1}^*(\mathbf{b} - \mathbf{1}W + \mathbf{a}, \tilde{C})] \\
&> \phi(\mathbf{b}) + \min_{W \in \mathcal{W}(\mathbf{b}, C)} E_{\mathbf{a}, \tilde{\mathbf{C}}}[V_{n-1}^*(\mathbf{b} - \mathbf{1}W + \mathbf{a}, \tilde{C})] \\
&= V_n^*(\mathbf{b}, C),
\end{aligned}$$

where, for each allocation $W' \in \mathcal{W}(\mathbf{b}', C)$, we define $W(W') \in \mathcal{W}(\mathbf{b}, C)$ as the allocation that assigns to each queue j the same number of servers (the same servers) as W' does unless the queue is empty, in which case it assigns only b_j. In other words, $\mathbf{1}W(W') = \mathbf{b} - [\mathbf{b} - \mathbf{1}W']^+$. In light of this, the first inequality holds by the induction hypothesis and noticing that $\mathbf{b} - \mathbf{1}W(W') = [\mathbf{b} - \mathbf{1}W']^+ \leq [\mathbf{b}' - \mathbf{1}W']^+ \leq \mathbf{b}' - \mathbf{1}W'$, where the $[\cdot]^+$ is removed because we know that $\mathbf{b}' \geq \mathbf{1}W'$ from $W' \in \mathcal{W}(\mathbf{b}', C)$. The second inequality holds because $W(W') \in \mathcal{W}(\mathbf{b}, C)$. The third inequality is a result of the strict monotonicity of ϕ. □

The following Lemma 2.16 shows that the MTLB policy is optimal at horizon $n+1$ if we know that $v_n \in \mathcal{F}$.

Lemma 2.16. *If $v_n \in \mathcal{F}$, then the MTLB policy is optimal at horizon $n+1$.*

Proof. We need to show that $v_n(\mathbf{b} - \mathbf{1}W^*) = \min_{W \in \mathcal{W}(\mathbf{b}, C)} v_n(\mathbf{b} - \mathbf{1}W)$ when

$W^* = [w^*_{ij}] \in \mathcal{W}(\mathbf{b}, C)$ is an MTLB allocation.

We first show that W^* must satisfy the maximum-throughput condition (**C1**). Assume W^* does not satisfy **(C1)**, i.e., $\sum_{i,j} w^*_{ij} < L$, where L is the maximum achievable throughput. By Proposition A.6, there exists at least one throughput-increasing path I. The throughput-increasing allocation $W^t(I)$ results in one more throughput and smaller leftover queues, i.e., $\mathbf{b} - \mathbf{1}W^t(I) < \mathbf{b} - \mathbf{1}W^*$. Hence, by Lemma 2.15, $v_n(\mathbf{b} - \mathbf{1}W^t(I)) < v_n(\mathbf{b} - \mathbf{1}W^*)$, a contradiction with the optimality of W^*.

Next, we show that W^* must also satisfy (**C2**). Assume W^* satisfies **(C1)** but not **(C2)**. Hence, by Proposition A.7, there must exist a balancing path $S = S(W^*, i, k)$ for some queues i, k. Let W' be the corresponding balancing allocation. Let $\mathbf{l}^* = \mathbf{b} - \mathbf{1}W^*$ and $\mathbf{l}' = \mathbf{b} - \mathbf{1}W'$. Since S is a balancing path and W' is the balancing allocation, we know that $l^*_i \geq l^*_k + 2$ and $\mathbf{l}' = R_{ik}(\mathbf{l}^*)$. Using this fact and the assumption that $v_n \in \mathcal{F}$ (hence, v_n satisfies (**B.6**)), we have $v_n(\mathbf{l}') = v_n(R_{ik}(\mathbf{l}^*)) \leq v_n(\mathbf{l}^*)$. Hence, W' is also optimal. Since any maximum-throughput allocations can be made to some MTLB allocation via some finite sequence of balancing allocations (Corollary A.8), we have that any MTLB allocation is also optimal. □

The rest of the appendix provides Lemmas 2.22 to 2.24, necessary to establish that $v_{n+1} \in \mathcal{F}$ if $v_n \in \mathcal{F}$, as discussed in the proof of Theorem 2.2. The next lemma shows that v_n satisfies (**B.2**) for all $n = 0, \ldots, T$.

Lemma 2.22. *$v_n(\mathbf{b})$, $n = 0, \ldots, T$, is permutation invariant on $\mathbf{b}$, i.e., $v_n(\pi(\mathbf{b})) =$*

$v_n(\mathbf{b})$ *for any permutation function* π.

Proof. From (2.13) and Assumption **(A1)**, it suffices to show the permutation invariance property of $V_n^*(\mathbf{b}, C)$ for any $(\mathbf{b}, C)$.

Induction Basis: $V_0^*(\mathbf{b}, C) = \phi(\mathbf{b}) = \sum_{j=1}^N \xi(b_j)$ is clearly permutation invariant.

Induction Step: Assume $V_{n-1}^*(\pi(\mathbf{b}), \Pi_\pi(C)) = V_{n-1}^*(\mathbf{b}, C)$, then

$$\begin{aligned} V_n^*(\pi(\mathbf{b}), \Pi_\pi(C)) &= \min_{W \in \mathcal{W}(\pi(\mathbf{b}), \Pi_\pi(C))} E_{\mathbf{a}, \tilde{\mathbf{C}}}[V_{n-1}^*(\pi(\mathbf{b}) - \mathbf{1}W + \mathbf{a}, \tilde{C})] + \phi(\pi(\mathbf{b})) \\ &= \min_{W \in \mathcal{W}(\pi(\mathbf{b}), \Pi_\pi(C))} E_{\mathbf{a}, \tilde{\mathbf{C}}}[V_{n-1}^*(\pi(\mathbf{b}) - \mathbf{1}W + \pi(\mathbf{a}), \Pi_\pi(\tilde{C}))] + \phi(\mathbf{b}), \end{aligned}$$

where the last equality is a direct result of Assumptions **(A1)** and **(A2)**. Now, using the fact that

$$W \in \mathcal{W}(\mathbf{b}, C) \Leftrightarrow \Pi_\pi(W) \in \mathcal{W}(\pi(\mathbf{b}), \Pi_\pi(C))$$

and the induction hypotheses, we have

$$\begin{aligned} V_n^*(\pi(\mathbf{b}), \Pi_\pi(C)) &= \min_{W \in \mathcal{W}(\mathbf{b}, C)} E_{\mathbf{a}, \tilde{\mathbf{C}}}[V_{n-1}^*(\pi(\mathbf{b}) - \mathbf{1}\Pi_\pi(W) + \pi(\mathbf{a}), \Pi_\pi(\tilde{C}))] + \phi(\mathbf{b}) \\ &= \min_{W \in \mathcal{W}(\mathbf{b}, C)} E_{\mathbf{a}, \tilde{\mathbf{C}}}[V_{n-1}^*(\mathbf{b} - \mathbf{1}W + \mathbf{a}, \tilde{C})] + \phi(\mathbf{b}) \\ &= V_n^*(\mathbf{b}, C). \end{aligned}$$

□

Next, we establish Lemmas 2.23 and 2.24. Specifically, given that $\hat{v}_n \in \hat{\mathcal{F}}$, we show that $E_{\mathbf{a}, \mathbf{C}}\left[\min_{W \in \mathcal{W}(C)} \hat{v}_n(\mathbf{b} - \mathbf{1}W + \mathbf{a})\right]$ satisfies (**B.3**) to (**B.5**) in Lemma 2.23 and satisfies (**B.6**) in Lemma 2.24. Without loss of generality, we consider $i=1$ and $j=2$ in (**B.3**) to (**B.6**). Note that from now on we are working with the relaxed problem where overallocation is allowed, i.e., we are considering allocations in $\mathcal{W}(C)$, instead of $\mathcal{W}(\mathbf{b},C)$. Before we proceed, for notational convenience, we write

$$\mathcal{T}_n^{\mathbf{a},C}(\mathbf{b}) := \min_{W\in\mathcal{W}(C)} \hat{v}_n(\mathbf{b}-\mathbf{1}W+\mathbf{a}), \tag{A.4}$$

and define the set of optimal allocations as follows:

Definition A.9. Define $\mathcal{X}^*(\mathbf{b},C)$ to be the set of all optimal (not necessarily non-idling) allocations when the state of the system is $(\mathbf{b},C)$. In other words, at horizon $n+1$,

$$\mathcal{X}^*(\mathbf{b},C) := \{W^*\in\mathcal{W}(C) : v_n([\mathbf{b}-\mathbf{1}W^*]^+) = \min_{W\in\mathcal{W}(C)} v_n([\mathbf{b}-\mathbf{1}W]^+)\} \tag{A.5}$$

We are now ready to show the following important lemma:

Lemma 2.23. *Assuming $N=2$ and $\hat{v}_n\in\hat{\mathcal{F}}$. For any state $\mathbf{b}$, $E_{\mathbf{a},\mathbf{C}}\left[\mathcal{T}_n^{\mathbf{a},C}(\mathbf{b})\right]$ satisfies (**B.3**), (**B.4**), and (**B.5**).*

Proof. By using Fact 2.19, it suffices to show that $\mathcal{T}_n^{\mathbf{a},C}(\mathbf{b})$ satisfies conditions (**B.3**) to (**B.5**) for any realization $(\mathbf{a},C)$ of the arrival and connectivity processes. From (**B.3**) to (**B.5**) and the definition of $\mathcal{T}_n^{\mathbf{a},C}(\mathbf{b})$ in (A.4), it is equivalent to show the non-negativity of the following quantities, respectively:

$$\begin{aligned}[i]\ &\min_{W\in\mathcal{W}(C)} \hat{v}_n(\mathbf{b}'+\mathbf{e}_1+\mathbf{e}_2-\mathbf{1}W) + \min_{W\in\mathcal{W}(C)} \hat{v}_n(\mathbf{b}'-\mathbf{1}W)\\ &- \min_{W\in\mathcal{W}(C)} \hat{v}_n(\mathbf{b}'+\mathbf{e}_1-\mathbf{1}W) - \min_{W\in\mathcal{W}(C)} \hat{v}_n(\mathbf{b}'+\mathbf{e}_2-\mathbf{1}W),\end{aligned} \tag{A.6}$$

$$[ii] \min_{W\in\mathcal{W}(C)} \hat{v}_n(\mathbf{b}'+\mathbf{e}_1-\mathbf{1}W) - 2\min_{W\in\mathcal{W}(C)} \hat{v}_n(\mathbf{b}'-\mathbf{1}W) + \min_{W\in\mathcal{W}(C)} \hat{v}_n(\mathbf{b}'-\mathbf{e}_1-\mathbf{1}W), \tag{A.7}$$

and

$$\begin{aligned}[iii] \min_{W\in\mathcal{W}(C)} \hat{v}_n(R_{12}(\mathbf{b}')-\mathbf{1}W) - 2\min_{W\in\mathcal{W}(C)} \hat{v}_n(\mathbf{b}'-\mathbf{1}W)&\\ + \min_{W\in\mathcal{W}(C)} \hat{v}_n(R_{21}(\mathbf{b}')-\mathbf{1}W),&\end{aligned} \tag{A.8}$$

where we let $\mathbf{b}' := \mathbf{a}+\mathbf{b}$ for convenience. The non-negativity of (A.6) to (A.8) is shown by using the assumption that $\hat{v}_n \in \hat{\mathcal{F}}$ (hence, $\hat{v}_n$ satisfies conditions (**B.1**) to (**B.6**)) and the fact that the MTLB policy is optimal at horizon $n+1$ (Lemma 2.16).

Since the MTLB policy is optimal at horizon $n+1$, we first make the following important observation:[3]

Observation A.10. For $N = 2$ users, there exists an MTLB allocation $W^* \in \mathcal{X}^*(\mathbf{b}', C)$ at horizon $n+1$ such that $W^* \in \mathcal{X}^*(\mathbf{b}', C) \cap \mathcal{X}^*(\mathbf{b}'+\mathbf{e}_1, C) \cap \mathcal{X}^*(\mathbf{b}'+\mathbf{e}_1+\mathbf{e}_2, C)$.

This is because 1) adding one packet to each queue does not create any balancing paths, i.e., $W^* \in \mathcal{X}^*(\mathbf{b}'+\mathbf{e}_1+\mathbf{e}_2, C)$; and 2) W^* can always be chosen such that it gives priority to serving queue 1, hence, adding one packet to queue 1 does not create any balancing paths, i.e., $W^* \in \mathcal{X}^*(\mathbf{b}'+\mathbf{e}_1, C)$.

Now, with this choice of $W^* \in \mathcal{X}^*(\mathbf{b}', C) \cap \mathcal{X}^*(\mathbf{b}'+\mathbf{e}_1, C) \cap \mathcal{X}^*(\mathbf{b}'+\mathbf{e}_1+$

[3] As we will discuss later, this observation which is essential in proving the lemma does not hold in the case of general $N(\geq 3)$.

$\mathbf{e}_2, C)$ and $\mathbf{d} := \mathbf{b}' - \mathbf{1}W^*$, we rewrite some terms in (A.6) to (A.8):

$$\min_{W \in \mathcal{W}(C)} \hat{v}_n(\mathbf{b}' - \mathbf{1}W + \mathbf{e}_1 + \mathbf{e}_2) = \hat{v}_n(\mathbf{d} + \mathbf{e}_1 + \mathbf{e}_2), \tag{A.9}$$

$$\min_{W \in \mathcal{W}(C)} \hat{v}_n(\mathbf{b}' - \mathbf{1}W) = \hat{v}_n(\mathbf{d}), \tag{A.10}$$

$$\min_{W \in \mathcal{W}(C)} \hat{v}_n(\mathbf{b}' + \mathbf{e}_1 - \mathbf{1}W) = \hat{v}_n(\mathbf{d} + \mathbf{e}_1). \tag{A.11}$$

Now, we are ready to show that $\mathcal{T}_n^{\mathbf{a},C}(\mathbf{b})$ satisfies (**B.3**) to (**B.5**), respectively.

(i) $\mathcal{T}_n^{\mathbf{a},C}(\mathbf{b})$ satisfies (**B.3**).

$$\begin{aligned}
&\min_{W \in \mathcal{W}(C)} \hat{v}_n(\mathbf{b}'{+}\mathbf{e}_1{+}\mathbf{e}_2{-}\mathbf{1}W) + \min_{W \in \mathcal{W}(C)} \hat{v}_n(\mathbf{b}'{-}\mathbf{1}W) \\
&\quad - \min_{W \in \mathcal{W}(C)} \hat{v}_n(\mathbf{b}'{+}\mathbf{e}_1{-}\mathbf{1}W) - \min_{W \in \mathcal{W}(C)} \hat{v}_n(\mathbf{b}'{+}\mathbf{e}_2{-}\mathbf{1}W) \\
&= \hat{v}_n(\mathbf{d} + \mathbf{e}_1 + \mathbf{e}_2) + \hat{v}_n(\mathbf{d}) - \hat{v}_n(\mathbf{d} + \mathbf{e}_1) - \min_{W \in \mathcal{W}(C)} \hat{v}_n(\mathbf{b}'{+}\mathbf{e}_2{-}\mathbf{1}W) \\
&\geq \hat{v}_n(\mathbf{d} + \mathbf{e}_1 + \mathbf{e}_2) + \hat{v}_n(\mathbf{d}) - \hat{v}_n(\mathbf{d} + \mathbf{e}_1) - \hat{v}_n(\mathbf{d} + \mathbf{e}_2) \\
&\geq 0,
\end{aligned}$$

where the equality is due to (A.9)-(A.11), the first inequality is due to the observation that $W^* \in \mathcal{X}^*(\mathbf{b}', C) \subseteq \mathcal{W}(C)$ (but not necessarily in $\mathcal{X}^*(\mathbf{b}' + \mathbf{e}_2, C)$), and the last inequality holds because $\hat{v}_n \in \hat{\mathcal{F}}$ and hence satisfying condition (**B.3**).

(ii) $\mathcal{T}_n^{\mathbf{a},C}(\mathbf{b})$ satisfies (**B.4**).

Using (A.10) and (A.11), this is equivalent to showing

$$\hat{v}_n(\mathbf{d} + \mathbf{e}_1) - 2\hat{v}_n(\mathbf{d}) + \min_{W \in \mathcal{W}(C)} \hat{v}_n(\mathbf{b}' - \mathbf{e}_1 - \mathbf{1}W) \geq 0$$

To show this, we consider the following two cases, depending on whether $W^* \in \mathcal{X}^*(\mathbf{b}' - \mathbf{e}_1, C)$ or not.

Case 1: If $W^* \in \mathcal{X}^*(\mathbf{b}' - \mathbf{e}_1, C)$, then

$$\hat{v}_n(\mathbf{d} + \mathbf{e}_1) - 2\hat{v}_n(\mathbf{d}) + \min_{W \in \mathcal{W}(C)} \hat{v}_n(\mathbf{b}' - \mathbf{e}_1 - \mathbf{1}W)$$
$$= \hat{v}_n(\mathbf{d} + \mathbf{e}_1) - 2\hat{v}_n(\mathbf{d}) + \hat{v}_n(\mathbf{d} - \mathbf{e}_1) \geq 0,$$

since $\hat{v}_n$ satisfies (**B.4**).

Case 2: $W^* \notin \mathcal{X}^*(\mathbf{b}' - \mathbf{e}_1, C)$. Thus, there exists a balancing path $S(W^*, 2, 1)$ from queue 2 to queue 1 and $d_2 = d_1 + 1$. Note that since there are only two queues, this balancing path is simply a balancing path $(2, v, 1)$ for some server $v \in V$, which is connected to both queues but is being assigned to queue 1 under W^*. Hence, the balancing allocation $(W^*)^b(S)$ is in $\mathcal{X}^*(\mathbf{b}' - \mathbf{e}_1, C)$ and we have $\mathbf{1}(W^*)^b(S) = R_{12}(\mathbf{1}W^*) = \mathbf{1}W^* - \mathbf{e}_1 + \mathbf{e}_2$. In other words,

$$\begin{aligned}
&\hat{v}_n(\mathbf{d} + \mathbf{e}_1) - 2\hat{v}_n(\mathbf{d}) + \min_{W \in \mathcal{W}(C)} \hat{v}_n(\mathbf{b}' - \mathbf{e}_1 - \mathbf{1}W) \\
&= \hat{v}_n(\mathbf{d} + \mathbf{e}_1) - 2\hat{v}_n(\mathbf{d}) + \hat{v}_n(\mathbf{b}' - \mathbf{e}_1 - \mathbf{1}(W^*)^b(S)) \\
&= \hat{v}_n(\mathbf{d} + \mathbf{e}_1) - 2\hat{v}_n(\mathbf{d}) + \hat{v}_n(\mathbf{d} - \mathbf{e}_2) \\
&= \hat{v}_n(\mathbf{d} + \mathbf{e}_1) - \hat{v}_n(\mathbf{d}) - \hat{v}_n(\pi_{12}(\mathbf{d})) + \hat{v}_n(\mathbf{d} - \mathbf{e}_2) \\
&= \hat{v}_n(\mathbf{d} + \mathbf{e}_1) - \hat{v}_n(\mathbf{d}) - \hat{v}_n(\mathbf{d} + \mathbf{e}_1 - \mathbf{e}_2) + \hat{v}_n(\mathbf{d} - \mathbf{e}_2) \\
&\geq 0,
\end{aligned}$$

where the second equality holds because $\mathbf{b}' - \mathbf{e}_1 - \mathbf{1}(W^*)^b(S) = \mathbf{d} - \mathbf{e}_2$, the third equality holds because $\hat{v}_n$ satisfies (**B.2**), the fourth equality follows from $d_2 = d_1 + 1$, and the last inequality because $\hat{v}_n$ satisfies (**B.3**).

(iii) $\mathcal{T}_n^{\mathbf{a},C}(\mathbf{b})$ satisfies (**B.5**).

This is equivalent to showing

$$\min_{W\in\mathcal{W}(C)} \hat{v}_n(R_{12}(\mathbf{b}') - \mathbf{1}W) - 2\hat{v}_n(\mathbf{d}) + \min_{W\in\mathcal{W}(C)} \hat{v}_n(R_{21}(\mathbf{b}') - \mathbf{1}W) \geq 0 \qquad \text{(A.12)}$$

To show this, we consider the following three cases.

Case 1: If $W^* \in \mathcal{X}^*(R_{21}(\mathbf{b}'), C) \cap \mathcal{X}^*(R_{12}(\mathbf{b}'), C)$, then we are done since $\hat{v}_n$ satisfies (**B.5**).

Case 2: If $W^* \notin \mathcal{X}^*(R_{21}(\mathbf{b}'), C)$, then there exists a balancing path $S(W^*, 1, 2)$. Since $W^* \in \mathcal{X}^*(\mathbf{b}' + \mathbf{e}_1, C)$ (i.e., W^* was chosen to give priority to serving queue 1), we must have $d_1 = d_2$. Hence, $(W^*)^b(S) \in \mathcal{X}^*(R_{21}(\mathbf{b}'), C)$. Furthermore, $\mathbf{1}(W^*)^b(S) = R_{21}(\mathbf{1}W^*)$. Thus, for this case, the LHS of (A.12) becomes:

$$\begin{aligned}
&\min_{W\in\mathcal{W}(C)} \hat{v}_n(R_{12}(\mathbf{b}') - \mathbf{1}W) - 2\hat{v}_n(\mathbf{d}) + \min_{W\in\mathcal{W}(C)} \hat{v}_n(R_{21}(\mathbf{b}') - \mathbf{1}W) \\
&= \min_{W\in\mathcal{W}(C)} \hat{v}_n(R_{12}(\mathbf{b}') - \mathbf{1}W) - 2\hat{v}_n(\mathbf{d}) + \hat{v}_n(\mathbf{d}) \\
&= \min_{W\in\mathcal{W}(C)} \hat{v}_n(R_{12}(\mathbf{b}') - \mathbf{1}W) - \hat{v}_n(\mathbf{d}) \qquad \text{(A.13)}
\end{aligned}$$

Next we consider the two following subcases:

Case 2.1: If $W^* \in \mathcal{X}^*(R_{12}(\mathbf{b}'), C)$, then (A.13) is equal to $\hat{v}_n(R_{12}(\mathbf{d})) - \hat{v}_n(\mathbf{d}) \geq 0$, because $d_1 = d_2$ and $\hat{v}_n \in \hat{\mathcal{F}}$.

Case 2.2: If $W^* \notin \mathcal{X}^*(R_{12}(\mathbf{b}'), C)$, then there exists a balancing path $S(W^*, 2, 1)$. Hence, $(W^*)^b(S) \in \mathcal{X}^*(R_{12}(\mathbf{b}'), C)$, ensuring that $\mathbf{1}(W^*)^b(S) = R_{12}(\mathbf{1}W^*)$. Thus, (A.13) is equal to zero.

Case 3: If $W^* \in \mathcal{X}^*(R_{21}(\mathbf{b}'), C)$ but $W^* \notin \mathcal{X}^*(R_{12}(\mathbf{b}'), C)$, then, by the permutation invariance property of $\hat{v}_n$, this reduces to Case 2.1 above. □

Lemma 2.24. *Assuming $N = 2$ and $\hat{v}_n \in \hat{\mathcal{F}}$. For any state $\mathbf{b}$ such that $b_1 \geq b_2+1$, $E_{\mathbf{a},\mathbf{C}}[\mathcal{T}_n^{\mathbf{a},C}(\mathbf{b})]$ satisfies condition (**B.6**).*

Proof. To show that $E_{\mathbf{a},\mathbf{C}}[\mathcal{T}_n^{\mathbf{a},C}(\mathbf{b})]$ meets condition (**B.6**) is equivalent to show the non-negativity of $E_{\mathbf{a},\mathbf{C}}\left[\mathcal{T}_n^{\mathbf{a},C}(\mathbf{b}) - \mathcal{T}_n^{\mathbf{a},C}(R_{12}(\mathbf{b}))\right]$. For convenience, let us define

$$Z^{\mathbf{a},C}(\mathbf{b}) := \mathcal{T}_n^{\mathbf{a},C}(\mathbf{b}) - \mathcal{T}_n^{\mathbf{a},C}(R_{12}(\mathbf{b})). \tag{A.14}$$

Using the permutation invariance property of the arrival and connectivity processes (Assumptions **A1** and **A2**), we can rewrite $E_{\mathbf{a},C}[Z^{\mathbf{a},C}(\mathbf{b})]$ as

$$E_{\mathbf{a},\mathbf{C}}\left[Z^{\mathbf{a},C}(\mathbf{b})\right] = \frac{1}{2}E_{\mathbf{a},\mathbf{C}}\left[Z^{\mathbf{a},C}(\mathbf{b}) + Z^{\pi_{12}(\mathbf{a}),\Pi_{\pi_{12}}(C)}(\mathbf{b})\right].$$

Thus, it suffices to show that, for any $(\mathbf{a}, C)$ and $b_1 \geq b_2 + 1$,

$$Z^{\mathbf{a},C}(\mathbf{b}) + Z^{\pi_{12}(\mathbf{a}),\Pi_{\pi_{12}}(C)}(\mathbf{b}) \geq 0.$$

We show this by noticing that

$$\begin{aligned} Z^{\mathbf{a},C}(\mathbf{b}) + Z^{\pi_{12}(\mathbf{a}),\Pi_{\pi_{12}}(C)}(\mathbf{b}) &= Z^{\mathbf{a},C}(\mathbf{b}) + Z^{\mathbf{a},C}(\pi_{12}(\mathbf{b})) \\ &= Z^{\mathbf{a},C}(\mathbf{b}) + \mathcal{T}_n^{\mathbf{a},C}(\pi_{12}(\mathbf{b})) - \mathcal{T}_n^{\mathbf{a},C}(\pi_{12}(R_{12}(\mathbf{b}))) \\ &= \mathcal{T}_n^{\mathbf{a},C}(\mathbf{b}^0) - \mathcal{T}_n^{\mathbf{a},C}(\mathbf{b}^1) + \mathcal{T}_n^{\mathbf{a},C}(\mathbf{b}^M) - \mathcal{T}_n^{\mathbf{a},C}(\mathbf{b}^{M-1}) \end{aligned} \tag{A.15}$$

where $M := b_1 - b_2$ (≥ 1) and $\mathbf{b}^m := \mathbf{b} - m\mathbf{e}_1 + m\mathbf{e}_2$, for $m = 0, \ldots, M$. The first equality follow from the permutation invariance property, while the second and third equalities follow from (A.14). Note that $\pi_{12}(\mathbf{b}) = b_2\mathbf{e}_1 + b_1\mathbf{e}_2 = \mathbf{b} - (b_1 - b_2)\mathbf{e}_1 + (b_1 - b_2)\mathbf{e}_2 = \mathbf{b}^M$, $\pi_{12}(R_{12}(\mathbf{b})) = \mathbf{b}^{M-1}$, $\mathbf{b}^{m+1} = R_{12}(\mathbf{b}^m)$, and $\mathbf{b}^{m-1} = R_{21}(\mathbf{b}^m)$.

Now notice that if $M = 1$, then the RHS of (A.15) is zero. If $M \geq 2$, we have

$$\begin{aligned}
&\mathcal{T}_n^{\mathbf{a},C}(\mathbf{b}^0) - \mathcal{T}_n^{\mathbf{a},C}(\mathbf{b}^1) - \mathcal{T}_n^{\mathbf{a},C}(\mathbf{b}^{M-1}) + \mathcal{T}_n^{\mathbf{a},C}(\mathbf{b}^M) \\
&= \sum_{m=1}^{M-1} \left\{ \mathcal{T}_n^{\mathbf{a},C}(\mathbf{b}^{m-1}) - 2\mathcal{T}_n^{\mathbf{a},C}(\mathbf{b}^m) + \mathcal{T}_n^{\mathbf{a},C}(\mathbf{b}^{m+1}) \right\} \\
&= \sum_{m=1}^{M-1} \{ \mathcal{T}_n^{\mathbf{a},C}(R_{21}(\mathbf{b}^m)) - 2\mathcal{T}_n^{\mathbf{a},C}(\mathbf{b}^m) + \mathcal{T}_n^{\mathbf{a},C}(R_{12}(\mathbf{b}^m)) \} \\
&\geq 0,
\end{aligned}$$

where the inequality holds because $\hat{v}_n \in \hat{\mathcal{F}}$ and, from Lemma 2.23, $\mathcal{T}_n^{\mathbf{a},C}(\cdot)$ satisfies condition (**B.5**). □

Remark A.11. The proofs for Lemmas 2.23 and 2.24 are valid only for $N = 2$. The main difficulty in the extension to the general case of $N > 2$ is with Lemma 2.23, Observation A.10, where we cannot claim that there exists an MTLB allocation $W^* \in \mathcal{X}^*(\mathbf{b}', C)$ such that $W^* \in \mathcal{X}^*(\mathbf{b}', C) \cap \mathcal{X}^*(\mathbf{b}' + \mathbf{e}_1, C) \cap \mathcal{X}^*(\mathbf{b}' + \mathbf{e}_1 + \mathbf{e}_2, C)$. This reflects a major obstacle that adding one packet to queue 1 and/or queue 2 may generate a balancing path in the original optimal allocation. For general N, we will need to explore more cases and require extra convexity properties of functions in $\mathcal{F}$. Currently, we do not know which extra conditions are needed and how to show that these conditions of v_n carried over to v_{n+1}. Therefore, the extension to the general case of $N > 2$ remains open.

A.3 Supporting Lemmas and Proof of Theorem 2.3

In this appendix, we prove the optimality of the MTLB-F policy for the fluid server allocation relaxation. We assume that the cost function $\phi(\mathbf{b})$ is monotonically increasing, permutation invariant, and convex on $\mathbf{b} \in \mathbb{R}_+^N$. It is easy to see that the strict monotonicity and permutation invariance of v_n in the fluid relaxation (Lemmas 2.15 and 2.22) still hold for all n. Next we show the convexity of v_n for all n.

Lemma A.12. *$v_n(\mathbf{b})$ is convex on $\mathbf{b}$ for all $n = 0, 1, \ldots, T$.*

Proof. It suffices to show the convexity of $V_n^*(\mathbf{b}, C)$ for every C. We show this by induction: $V_0^*(\mathbf{b}, C) = \phi(\mathbf{b})$ is convex on $\mathbf{b}$. Assume $V_{n-1}^*(\mathbf{b}, C)$ convex on $\mathbf{b}$. For $i = 1, 2$, let $W^i \in \mathcal{W}^f(\mathbf{b}^i, C)$ be an optimal allocation at time n for $(\mathbf{b}^i, C)$, i.e.,

$$V_n^*(\mathbf{b}^i, C) = \phi(\mathbf{b}^i) + E_{\mathbf{a},\tilde{C}}\left[V_{n-1}^*(\mathbf{b}^i + \mathbf{a} - \mathbf{1}W^i, \tilde{C})\right]. \tag{A.16}$$

Then, for any $\beta \in [0, 1]$, let $W^\beta = \beta W^1 + (1 - \beta)W^2$ and $\mathbf{b}^\beta = \beta\mathbf{b}^1 + (1 - \beta)\mathbf{b}^2$. We can easily see that $W^\beta \in \mathcal{W}^f(\mathbf{b}^\beta, C)$ since it satisfies the conditions (a') to (d) in Definition 2.28. Then, we have

$$\begin{aligned}
V_n^*(\mathbf{b}^\beta, C) &= \phi(\mathbf{b}^\beta) + \min_{W \in \mathcal{W}^f(\mathbf{b}^\beta, C)} E_{\mathbf{a},\tilde{C}} \left[V_{n-1}^*(\mathbf{b}^\beta + \mathbf{a} - \mathbf{1}W, \tilde{C}) \right] \\
&\le \phi(\mathbf{b}^\beta) + E_{\mathbf{a},\tilde{C}} \left[V_{n-1}^*(\mathbf{b}^\beta + \mathbf{a} - \mathbf{1}W^\beta, \tilde{C}) \right] \\
&\le \phi(\mathbf{b}^\beta) + E_{\mathbf{a},\tilde{C}} \left[\beta V_{n-1}^*(\mathbf{b}^1 + \mathbf{a} - \mathbf{1}W^1, \tilde{C}) \right] \\
&\quad + E_{\mathbf{a},\tilde{C}} \left[(1-\beta) V_{n-1}^*(\mathbf{b}^2 + \mathbf{a} - \mathbf{1}W^2, \tilde{C}) \right] \\
&\le \beta \left(\phi(\mathbf{b}^1) + E_{\mathbf{a},\tilde{C}} \left[V_{n-1}^*(\mathbf{b}^1 + \mathbf{a} - \mathbf{1}W^1, \tilde{C}) \right] \right) \\
&\quad + (1-\beta) \left(\phi(\mathbf{b}^2) + E_{\mathbf{a},\tilde{C}} \left[V_{n-1}^*(\mathbf{b}^2 + \mathbf{a} - \mathbf{1}W^2, \tilde{C}) \right] \right) \\
&= \beta V_n^*(\mathbf{b}^1, C) + (1-\beta) V_n^*(\mathbf{b}^2, C),
\end{aligned}$$

where the second inequality follows from the induction hypothesis, the last inequality from the convexity of the cost function ϕ, and the last equality from (A.16). □

The notions of alternating, balancing, and throughput-increasing paths and allocations can be generalized as follows:

Definition A.13. For $\epsilon > 0$, an *ϵ-alternating path* from queue $u_0 \in U$ to queue $u_k \in U$ with respect to $W \in \mathcal{W}^f(\mathbf{b}, C)$ is a sequence of distinct vertices

$$S(W, u_0, u_k, \epsilon) := (u_0, v_1, u_1, v_2, \ldots, v_k, u_k),$$

with $v_i \in V$, $u_i \in U$, $w_{v_i,u_i} \ge \epsilon$ for each $i = 1, \ldots, k$. An ϵ-alternating path $S(W, u_0, u_k, \epsilon)$ is called an *ϵ-balancing path* if

$$\left(b_{u_0} - \sum_{i=1}^{K} w_{i,u_0} \right) - \left(b_{u_k} - \sum_{i=1}^{K} w_{i,u_k} \right) \ge 2\epsilon.$$

Definition A.14. For $\epsilon > 0$, an *ϵ-throughput-increasing path* relative to $W \in \mathcal{W}^f(\mathbf{b}, C)$, is a sequence of distinct vertices

$$I(W, v_0, u_k, \epsilon) := (v_0, u_1, v_1, u_2, \ldots, v_{k-1}, u_k)$$

with (a) $v_i \in V$, $u_i \in U$, and $w_{v_i,u_i} \geq \epsilon$ for each i, (b) v_0 is not fully matched, i.e., $1 - \sum_{j=1}^{N} w_{v_0,j} \geq \epsilon$, and (c) u_k is non-empty, i.e., $b_{u_k} - \sum_{i=1}^{K} w_{i,u_k} \geq \epsilon$.

Definition A.15. Given an ϵ-balancing path $S = S(W, u_0, u_k, \epsilon)$ for $\epsilon > 0$, an ϵ-balancing allocation $W^{b,\epsilon}(S)$ balances the queue u_0 and u_k by ϵ packets by reassigning server v_l to serve ϵ packets more from queue u_{l-1} and ϵ packets less from queue u_l, $\forall l = 1, \ldots, k$.

Definition A.16. Given an ϵ-throughput-increasing path $I = I(W, v_0, u_k, \epsilon)$ for $\epsilon > 0$, an ϵ-throughput-increasing allocation $W^{t,\epsilon}(I) \in \mathcal{W}^f(\mathbf{b}, C)$ achieves additional throughput of ϵ packets by assigning v_0 to serve ϵ more packets from u_1 and reassigning server v_l to serve ϵ packets more from queue u_{l+1} and ϵ packets less from queue u_l, $\forall l = 1, \ldots, k-1$.

Note that the throughput-increasing and balancing paths previously considered under the integral server allocation (Definitions A.3 and A.4) are equivalent to the fluid versions (Definitions A.13 and A.14), when we take $\epsilon = 1$.

We see that similar results as in Appendix A.1 hold for the ϵ-throughput-increasing and ϵ-balancing paths as well. The existence of the MTLB-F policy could be similarly established as in Theorem 2.1 using the following result:

Proposition A.6'. *An allocation achieves the maximum throughput (**C1**) if and only if it has no ϵ-throughput-increasing paths for any $\epsilon > 0$.*

Proof. Similar to the proof in Proposition A.6. □

Proposition A.7'. *Any allocation satisfying the maximum-throughput condition (**C1**) also satisfies the load-balancing condition (**C2**) if and only if it has no ϵ-balancing path for any $\epsilon > 0$.*

Proof. The proof is very similar to that of Proposition A.7 but here we allow $\epsilon \in (0, 1]$ packets to be reallocated. It suffices to show that if a maximum-throughput allocation $W = [w_{i,j}] \in \mathcal{W}^f(\mathbf{b}, C)$ does not satisfy (**C2**), then W has at least one ϵ-balancing path for some $\epsilon > 0$. Let $W^* = [w^*_{i,j}] \in \mathcal{W}^f(\mathbf{b}, C)$ be an MTLB-F allocation, chosen such that $||W^* - W||$ is minimized, where $||X|| := \sum_{i,j} |x_{ij}|$ for any matrix $X = [x_{ij}]$. Now let G_d be the weighted subgraph of the bipartite graph $G = (V, U, C)$ induced by the allocation difference matrix $W^* - W$. Specifically, G_d contains an edge (v, u) for $v \in V$ and $u \in U$ if and only if $|w^*_{v,u} - w_{v,u}| > 0$. We assign the weight of the edge (v, u) as $|w^*_{v,u} - w_{v,u}|$. Color the edges (v, u) of G_d green if $w^*_{v,u} - w_{v,u} > 0$, and red if $w^*_{v,u} - w_{v,u} < 0$. Direct the green edges from V to U and the red edges from U to V. Let the leftover queue vectors under W and W^* be $\mathbf{l} = \mathbf{b} - \mathbf{1}W$ and $\mathbf{l}^* = \mathbf{b} - \mathbf{1}W^*$, respectively.

We claim that for every directed path P in G_d from $u_1 \in U$ to $u_2 \in U$, we have

$$l^*_{u_1} \le l^*_{u_2}. \tag{A.17}$$

To see this, let $P = (u_1, \ldots, u_2)$ be a directed path in G_d. By the choice of the directions for the edges, P must be alternating between red and green edges. Let ϵ' be the minimum of the weights of the edges along this path ($\epsilon' > 0$ by construction of G_d). If $l^*_{u_1} > l^*_{u_2}$ then P is an ϵ-balancing path for W^* with $\epsilon = \min\{\epsilon', (l^*_{u_1} - l^*_{u_2})/2\} > 0$, contradicting to the assumption that W^* satisfy (**C2**).

Next, we claim that G_d is acyclic. Assume G_d is cyclic, i.e., there is a red-green alternating and directed path $P = (u_1, \ldots, u_1)$ in G_d, for some $u_1 \in U$, with the minimum weight $\epsilon > 0$ along this path. Then, we get another MTLB-F allocation $W^{**} = [w^{**}_{i,j}]$ by letting $w^{**}_{v,u} = w^*_{v,u} - \epsilon$ for all green edges $(v, u) \in P$, $w^{**}_{v,u} = w^*_{v,u} + \epsilon$ for all red edges $(v, u) \in P$, and $w^{**}_{v',u'} = w^*_{v',u'}$ for all edges (v', u') not in P. We see that W^{**} is closer to W than W^* is to W because $|w^{**}_{v,u} - w_{v,u}| = |w^*_{v,u} - w_{v,u}| - \epsilon \geq 0$ for each edge (v, u) in P. In other words, $||W^{**} - W|| < ||W^* - W||$, in contradiction to the choice of W^*. Hence, we must have that G_d is acyclic.

Now, since both W^* and W achieve the maximum throughput, we have that $\sum_{i=1}^N l_i = \sum_{i=1}^N l^*_i$. But since W does not satisfy the LB condition (**C2**), there must exist $u_1 \in U$ such that

$$l_{u_1} < l^*_{u_1}. \tag{A.18}$$

Obviously, there is a red edge directed out of u_1. Starting from u_1 we build an alternating red-green and directed path P' in G_d as follows: (1) From an arbitrary vertex $u \in U$ (including u_1), if there is a red edge directed out of u and $l_u \leq l_{u_1}$, we build P' by arbitrarily select one of the red edges directed out of u. (2) From an arbitrary $v \in V$, we build P' by arbitrarily follow one of the green edges directed out of v. Such a green edge always exist.[4] (3) Otherwise, stop.

Using the fact that G_d is acyclic, P' is well-defined and finite. Let $u_2 \in U$ be the final vertex on the path and $\epsilon' > 0$ be the minimum weight along P'. There are two possible cases:

Case 1: $l_{u_2} > l_{u_1}$. In this case, we reverse the order of nodes in P' to arrive at an ϵ-balancing path relative to W, where $\epsilon = \min\{\epsilon', (l_{u_2} - l_{u_1})/2\} > 0$.

Case 2: There is no red edge directed out of u_2. Thus, P' arrived at u_2 via a green edge (v, u_2) for some $v \in V$. This means that u_2 is served at least $w^*_{v,u_2} - w_{v,u_2} > 0$ packets more under W^*, relative to W. Hence, $l_{u_2} > l^*_{u_2}$. This together with (A.17) and (A.18) give $l_{u_2} > l^*_{u_2} \geq l^*_{u_1} > l_{u_1}$, which means $l_{u_2} > l_{u_1}$. Reversing the order of nodes in P' gives a ϵ-balancing path relative to W, $\epsilon = \min\{\epsilon', (l_{u_2} - l_{u_1})/2\} > 0$.

Since there exists an ϵ-balancing path in both cases for some $\epsilon > 0$, we have the assertion of the proposition. □

Using the above results, we can show the following:

Theorem 2.3. *For the problem (**P**) with the fluid server allocation relaxation, the MTLB-F policy is optimal.*

[4] Otherwise, combining one of the red edges coming into v with W^* would have yielded a non-idling feasible allocation with additional positive packet throughput, a contradiction to (**C1**).

Proof. We need to show that

$$v_n(\mathbf{b} - \mathbf{1}W^*) = \min_{W \in \mathcal{W}^f(\mathbf{b},C)} v_n(\mathbf{b} - \mathbf{1}W), \tag{A.19}$$

where $W^* \in \mathcal{W}^f(\mathbf{b}, C)$ is MTLB-F. Since $\mathcal{W}^f(\mathbf{b}, C)$ is convex and compact, there exists an optimal allocation W^*. Similarly as in the proof of Lemma 2.16, we can show that W^* must satisfy (**C1**) using Proposition A.6' and the strict monotonicity of v_n (Lemma 2.15).

Now assume W^* satisfies (**C1**) but not (**C2**). By Proposition A.7', there must exist an ϵ-balancing path $S = S(W^*, i, k, \epsilon)$ relative to W^*, for some $\epsilon > 0$ and some queues $i, k \in U$. Let $W' = W^{b,\epsilon}(S)$ be the corresponding ϵ-balancing allocation. Let $\mathbf{l}^* = \mathbf{b} - \mathbf{1}W^*$ and $\mathbf{l}' = \mathbf{b} - \mathbf{1}W'$. We then have $l'_i = l^*_i - \epsilon$, $l'_k = l^*_k + \epsilon$, and $l'_u = l^*_u$ for all $u \neq i, k$. By the convexity and permutation invariance properties of v_n, we can show that $v_n(\mathbf{l}') \leq v_n(\mathbf{l}^*)$ [31] as follows: Since $\mathbf{l}^*$ and $\mathbf{l}'$ differ only in the ith and kth components, it suffices to consider a function v_n of two variables. We notice that $(l^*_i - \epsilon, l^*_k + \epsilon)$ lies on the interval joining (l^*_i, l^*_k) and (l^*_k, l^*_i). Hence, for some $\gamma \in [0, 1]$, we have $(l^*_i - \epsilon, l^*_k + \epsilon) = \gamma(l^*_i, l^*_k) + (1 - \gamma)(l^*_k, l^*_i)$. Using the convexity and the permutation invariance of v_n, we obtain $v_n(l^*_i - \epsilon, l^*_k + \epsilon) \leq \gamma v_n(l^*_i, l^*_k) + (1 - \gamma)v_n(l^*_k, l^*_i) = v_n(l^*_i, l^*_k)$. Hence, $v_n(\mathbf{l}') \leq v_n(\mathbf{l}^*)$ and W' is also optimal but more balanced than W^*. Thus, we can conclude that any MTLB-F allocation is optimal. □

Appendix B

Appendix for Chapter 3

B.1 Proof of Proposition 3.6

Proposition 3.6. *Consider a g-smoothly-scaling process $A^{(N)}$ with the limiting g-scaled log moment generation function Λ. Let $S_t^{(N)} = \sum_{i=1}^{t} A_i^{(N)}$, for $t \in \mathbb{N}$. Then, for $a > \lambda t$, we have*

$$\lim_{N\to\infty} \frac{1}{g(N)} \log Pr\left(\frac{S_t^{(N)}}{N} > a\right) = -t\Lambda^*(a/t), \tag{B.1}$$

where Λ^* is the convex conjugate of Λ.

Proof. Let $n = g(N)$ and $Y_t^{(n)} = \frac{g(N)}{N} S_t^{(N)}$. From (3.7) and the property of Λ for the g-smoothly-scaling process, we have

$$\Lambda_{Y_t}(\theta) := \lim_{n\to\infty} \frac{1}{n} \log E[e^{\theta Y_t^{(n)}}] = \Lambda_{S_t}(\theta) = t\Lambda(\theta),$$

which exists for each $\theta \in \mathbb{R}$ as an extended real number and is finite in a neighborhood of $\theta = 0$, essentially smooth, and lower-semicontinuous. Then, the Gärtner-

Ellis theorem (Theorem 2.11 in [29]) shows that $Y_t^{(n)}/n$ (which, in this case, is equivalent to $S_t^{(N)}/N$) satisfies the *large deviations principle* (LDP) in $\mathbb{R}$ with good convex rate function

$$\Lambda_{Y_t}^*(x) := \sup_{\theta \in \mathbb{R}} \ \theta x - \Lambda_{Y_t}(\theta) = \sup_{\theta \in \mathbb{R}} \ \theta x - t\Lambda(\theta) = t\Lambda^*(x/t).$$

For $a > E[\frac{S_t^{(N)}}{N}] = \lambda t$, the LDP result gives the assertion of the proposition (see Lemma 2.6 and Theorem 2.8 in [29]). □

B.2 Proof of Lemma 3.10

Lemma 3.10. *Given $g \in \mathcal{G}$, $T \in \mathbb{T}$, $r > \lambda$, a batch service of rNT every T timeslots, and a g-smoothly-scaling bit-arrival process characterized by the limiting g-scaled log moment generation function Λ, the decay rate of $P_{delay}(r,T)$ is given by the function I, i.e.,*

$$\lim_{N \to \infty} \frac{1}{g(N)} \log P_{delay}(r,T) = -I(r,T) \tag{B.2}$$

where

$$I(r,T) = \min_{\substack{t \in \mathbb{Z}^+: \\ tT+T-1-k>0}} (tT + T - 1 - k)\Lambda^* \left(r + \frac{(D+1-2T)r}{tT+T-1-k} \right), \tag{B.3}$$

and $k = D(\bmod T)$. In addition, $I(r,T)$ is lower-semicontinuous and increasing on r.

Proof. Let $g \in \mathcal{G}$, $T \in \mathbb{T} = \{1, 2, \ldots, \lfloor \frac{D}{2} \rfloor\}$, $r > \lambda$, and $k = D(\bmod T)$. Without loss of generality, we assume that $I(r,T) < \infty$.

For any given SNR ρ and $N = \log \rho$, there are $A_t^{(N)}$ bits arriving at time t. The queue is being served exactly at times mT, for $m \in \mathbb{Z}$, with an instantaneous removal of the oldest $RT = rNT$ bits. The corresponding queue dynamics for the queue size $Q_t^{(N)}$, at time t, are as follows.

$$Q_t^{(N)} = \begin{cases} \left[Q_{t-1}^{(N)} + A_t^{(N)} - TR\right]^+, & \text{if } t = mT,\ m \in \mathbb{Z}, \\ Q_{t-1}^{(N)} + A_t^{(N)}, & \text{otherwise,} \end{cases} \tag{B.4}$$

where $Q_{-\infty}^{(N)} \equiv 0$. Since the arrival process is stationary and the system started empty at time $-\infty$, then $Q_i^{(N)}$ has the same steady-state distribution as that of $Q_{mT+i}^{(N)}$, $m \in \mathbb{Z}$, for each $i = 0, \ldots, T-1$. The delay at time i also has the same steady-state distribution as the delay at time $mT + i$. Since $P_{\text{delay}}(r, T)$, as a function of r, T, is defined as the probability of the steady-state delay being greater than D, we have

$$\begin{aligned} P_{\text{delay}}(r, T) &:= \Pr(\text{steady-state delay of a bit} > D) \\ &= \frac{1}{T} \sum_{i=0}^{T-1} \Pr(\text{s-s delay of a bit arriving at time } i > D), \end{aligned} \tag{B.5}$$

where the equality holds since the arrivals are independent across time. From Lemma B.2 in Appendix B.5, we have that the delay violation probability of *any* bit arriving at time i is *asymptotically* equal to the delay violation probability of the *last* bit arriving at time i, (B.5) becomes

$$P_{\text{delay}}(r, T) \overset{g}{=} \frac{1}{T} \sum_{i=0}^{T-1} \Pr(\mathcal{Q}_i^{(N)}) \overset{g}{=} \sum_{i=0}^{T-1} \Pr(\mathcal{Q}_i^{(N)}), \tag{B.6}$$

where $\mathcal{Q}_i^{(N)}$ denotes the event that the last bit arriving at timeslot i violates the delay bound D. This holds because T is a constant independent of ρ. Hence, (B.6) says that P_{delay} is asymptotically equal to the sum of $\Pr(\mathcal{Q}_i^{(N)})$.

Next, we relate the event $\mathcal{Q}_i^{(N)}$ to a condition on the queue length $Q_i^{(N)}$, for $i = 0, \ldots, T-1$. To do this, we need to describe the condition that the delay of the last bit arriving at timeslot i violates the delay bound D. Upon arrival, the last bit sees $Q_i^{(N)}$ bits (including itself) waiting in the queue. Since the batch service happens exactly in multiples of T, the bit must wait $T-i$ timeslots for the next service to start and another $\left\lceil \frac{Q_i^{(N)}}{RT} \right\rceil T$ timeslots for all $Q_i^{(N)}$ bits (including the last bit) to get served and be decoded. Hence, the last bit arriving at time i violates the delay bound D if, and only if,

$$T - i + \left\lceil \frac{Q_i^{(N)}}{RT} \right\rceil T > D.$$

Let $\Omega^{(N)}$ contains all measurable random events. The condition above implies that the delay violation event for the last bit is given as

$$\mathcal{Q}_i^{(N)} := \{\omega \in \Omega^{(N)} : T - i + \left\lceil \frac{Q_i^{(N)}(\omega)}{RT} \right\rceil T > D\}. \tag{B.7}$$

Using (B.4) and (B.7), we show in Lemma B.1 of Appendix B.4 that

$$P_{\text{delay}}(r, T) \overset{g}{=} \Pr(\mathcal{Q}_{T-1-k}^{(N)}) \overset{g}{=} \Pr(Q_{T-1-k}^{(N)} > (D - T - k)R). \tag{B.8}$$

Intuitively, this means that $P_{\text{delay}}(r, T)$ is asymptotically equal to $\Pr(\mathcal{Q}_{T-1-k}^{(N)})$, equivalently $P_{\text{delay}}(r, T)$ is asymptotically equal to the probability that the last bit arriving at time $T - 1 - k$ sees a queue length greater than $(D - T - k)R$ bits.

Finally, using (B.8), what remains is to establish that

$$\begin{aligned}
&\lim_{N\to\infty} \frac{\log \Pr(Q^{(N)}_{T-k-1} > (D-T-k)rN)}{g(N)} \\
&= -I(r,T) \\
&= - \min_{\substack{t\in\mathbb{Z}^+:\\ tT+T-1-k>0}} (tT+T-k-1)\Lambda^*\left(r + \frac{(D+1-2T)r}{tT+T-k-1}\right). \qquad \text{(B.9)}
\end{aligned}$$

For notational simplicity, let $i := T-1-k$ and $q := (D-T-k)r$. Note that $q > ri \geq 0$ since $T \in \{1, 2, \dots, \lfloor \frac{D}{2} \rfloor\}$ and $k = D(\mod T)$. Now, since

$$\frac{q + rTt}{Tt+i} = r + \frac{(D+1-2T)r}{tT+T-k-1},$$

it is sufficient to show that

$$\lim_{N\to\infty} \frac{\log \Pr(Q^{(N)}_i > Nq)}{g(N)} = - \min_{\substack{t\in\mathbb{Z}^+:\\ tT+i>0}} (Tt+i)\Lambda^*(\frac{q+rTt}{Tt+i}). \qquad \text{(B.10)}$$

We separately show (matching) upper and lower bounds.

First, we show the lower bound. By using the queue dynamics in (B.4) recursively and the assumption of $Q^{(N)}_{-\infty} = 0$, the queue length $Q^{(N)}_i$ is related to the arrivals $A^{(N)}_j$, $j \leq i$, in the following manner:

$$Q^{(N)}_i = \sup_{t\in\mathbb{Z}^+} \left(\sum_{j=-tT+1}^{i} A^{(N)}_j - rtTN \right), \qquad \text{(B.11)}$$

where we use the convention that $\sum_{j=1}^{0} A^{(N)}_j \equiv 0$. Using this relation and the fact that $q > 0$, we have

$$\begin{aligned}
\Pr(Q^{(N)}_i > Nq) &= \Pr\left(\sup_{t\in\mathbb{Z}^+} \sum_{j=-tT+1}^{i} A^{(N)}_j - rtTN > Nq \right) \\
&= \Pr\left(\sup_{\substack{t\in\mathbb{Z}^+:\\ tT+i>0}} \sum_{j=-tT+1}^{i} A^{(N)}_j - rtTN > Nq \right).
\end{aligned}$$

Now, for any fixed $t \in \mathbb{Z}^+$ so that $tT + i > 0$, we have

$$\begin{aligned}\Pr(Q_i^{(N)} > Nq) &\geq \Pr\left(\sum_{j=-tT+1}^{i} A_j^{(N)} - rtTN > Nq\right) \\ &= \Pr\left(\sum_{j=1}^{tT+i} A_j^{(N)} > N(q + rTt)\right) \\ &= \Pr\left(\frac{S_{Tt+i}^{(N)}}{N} > q + rTt\right).\end{aligned}$$

Taking the limit of both sides and using Proposition 3.6, we have

$$\liminf_{N\to\infty} \frac{\log \Pr(Q_i^{(N)} > Nq)}{g(N)} \geq -(Tt+i)\Lambda^*(\frac{q + rTt}{Tt+i}). \tag{B.12}$$

Since t is arbitrary, maximizing the RHS over t gives the appropriate lower bound:

$$\liminf_{N\to\infty} \frac{\log \Pr(Q_i^{(N)} > Nq)}{g(N)} \geq -\inf_{\substack{t\in\mathbb{Z}^+:\\ tT+i>0}} (Tt+i)\Lambda^*(\frac{q + rTt}{Tt+i}). \tag{B.13}$$

For the upper bound, we use the following result from Lemma B.3 in Appendix B.6:

$$\limsup_{N\to\infty} \frac{\log \Pr(Q_i^{(N)} > Nq)}{g(N)} \leq -\inf_{\substack{t\in\mathbb{Z}^+:\\ tT+i>0}} (Tt+i)\Lambda^*\left(\frac{q + rTt}{Tt+i}\right),$$

noting that the RHS is strictly greater than $-\infty$, by assumption. Hence, the lower and upper bounds coincide and (B.10) holds.

To complete the proof, we show the properties of $I(r, T)$ for $T \in \mathbb{T}$. First, I is increasing on $r \geq \lambda$ because $\Lambda^*(x)$ is increasing on $x \geq \lambda$ (Lemma 2.7 in [29]). Second, $I(r, T)$ is lower-semicontinuous on r because I is the minimum of a number of function Λ^* which are lower-semicontinuous (Lemma 2.7 in [29]). □

Approximation 3.11. *Relaxing the integer constraint in (3.14) gives the lower* *bound of I as*

$$I(r,T) \geq \delta_r r(D+1-2T) =: I_{ir}(r,T), \tag{B.14}$$

where

$$\delta_r = \sup\{\theta > 0 : \Lambda(\theta) < \theta r\}. \tag{B.15}$$

Proof. By the definition of I, we have

$$\begin{aligned} I(r,T) &= \min_{\substack{t\in\mathbb{Z}^+: \\ tT+T-1-k>0}} (tT+T-1-k)\Lambda^*\big(r + r\frac{D-2T+1}{tT+T-1-k}\big) \\ &\geq \min_{\tau\in\mathbb{R}^+} \tau\Lambda^*\big(r + r\frac{D-2T+1}{\tau}\big) \\ &= \delta_r r(D-2T+1), \end{aligned}$$

where the last equality is a result of Lemma 3.4 of [29] with δ_r defined as in (B.15). □

B.3 Proof of Theorem 3.1

Proof of Theorem 3.1. Recall that:

$$P_{\text{tot}}(r,T) := P_{\text{ch}}(r,T) + (1 - P_{\text{ch}}(r,T))P_{\text{delay}}(r,T), \tag{B.16}$$

where, from (3.2),

$$P_{\text{ch}}(r,T) \doteq \rho^{-d_{\text{ch}}(r,T)} \tag{B.17}$$

and, from Lemma 3.10,

$$P_{\text{delay}}(r,T) \overset{g}{=} e^{-I(r,T)g(\log\rho)}. \tag{B.18}$$

Case 1: when $\lim_{N\to\infty} \frac{g(N)}{N} = \gamma \in (0, \infty)$. We have

$$P_{\text{delay}}(r,T) \doteq \rho^{-\gamma I(r,T)} \tag{B.19}$$

and

$$P_{\text{tot}}(r,T) \doteq \rho^{-\min\{\gamma I(r,T),\ d_{\text{ch}}(r,T)\}}. \tag{B.20}$$

The optimal negative SNR exponent of P_{tot} is

$$\begin{aligned} d^* &:= \sup_{\substack{r\in(\lambda,r_{\max})\\ T\in\mathbb{T}}} \left\{ \lim_{\rho\to\infty} -\frac{\log P_{\text{tot}}(r,T)}{\log\rho} \right\} \\ &= \sup_{\substack{r\in(\lambda,r_{\max})\\ T\in\mathbb{T}}} \left\{ \min\left\{\gamma I(r,T),\ d_{\text{ch}}(r,T)\right\}\right\} \\ &= \max_{T\in\mathbb{T}} \left\{ \sup_{r\in(\lambda,r_{\max})} \left\{ \min\left\{\gamma I(r,T),\ d_{\text{ch}}(r,T)\right\}\right\} \right\}. \end{aligned} \tag{B.21}$$

We first solve the optimization sub-problem within the bracket for any given integer $T \in \mathbb{T}$. Because $I(r,T)$ is increasing on $r \geq \lambda$ while $d_{\text{ch}}(r,T)$ is strictly decreasing on $r \in [0, r_{\max}]$, the sub-problem is solved by the optimal choice of multiplexing gain when the coding duration is fixed at T as

$$r^*(T) := \inf\{r \in (\lambda, r_{\max}) : \gamma I(r,T) = d_{\text{ch}}(r,T)\}. \tag{B.22}$$

Hence, (B.21) is solved with the optimal coding duration T^*, given as

$$T^* = \arg\max_{T\in\mathbb{T}} \ \gamma I(r^*(T), T),$$

and the optimal multiplexing gain r^*, given as

$$r^* = r^*(T^*).$$

Note that, since $I(r,T) > 0$ when $r > \lambda$ and $d_{\text{ch}}(r,T) > 0$ when $r < r_{\max}$, it is guaranteed that $r^*(T) \in (\lambda, r_{\max})$.

Case 2: when $\lim_{N\to\infty} \frac{g(N)}{N} = 0$ and $\lim_{N\to\infty} \frac{g(N)}{\log N} = \infty$. In this case, for all $r \in (\lambda, r_{\max})$ and all $T \in \mathbb{T}$, we have $P_{\text{delay}}(r,T)$ asymptotically dominates $P_{\text{ch}}(r,T)$ and hence $P_{\text{tot}}(r,T)$ is asymptotically equal to $P_{\text{delay}}(r,T)$. Since, for any $T \in \mathbb{T}$, $I(r,T)$ is increasing on $r > \lambda$, we have

$$\begin{aligned}\sup_{\substack{r\in(\lambda,r_{\max}),\\ T\in\mathbb{T}}} \lim_{\rho\to\infty} \frac{-\log P_{\text{tot}}(r,T)}{g(\log\rho)} &\leq \max_{T\in\mathbb{T}} \left\{ \sup_{r\in(\lambda,r_{\max})} I(r,T) \right\} \\ &= \max_{T\in\mathbb{T}} \ I(r_{\max},T).\end{aligned}$$

Case 3: when $\lim_{N\to\infty} \frac{g(N)}{N} = \infty$. This case is an opposite of Case 2. Here, $P_{\text{tot}}(r,T)$ is asymptotically equal to $P_{\text{ch}}(r,T)$ for all $r \in (\lambda, r_{\max})$ and all $T \in \mathbb{T}$. Since $d_{\text{ch}}(r,T)$ is decreasing on r and increasing on T, we have

$$\begin{aligned}\sup_{\substack{r\in(\lambda,r_{\max}),\\ T\in\mathbb{T}}} \lim_{\rho\to\infty} \frac{-\log P_{\text{tot}}(r,T)}{\log\rho} &\leq \max_{T\in\mathbb{T}} \left\{ \sup_{r\in(\lambda,r_{\max})} d_{\text{ch}}(r,T) \right\} \\ &= \max_{T\in\mathbb{T}} \ d_{\text{ch}}(\lambda,T) \\ &= d_{\text{ch}}(\lambda, \left\lfloor \frac{D}{2} \right\rfloor).\end{aligned}$$

□

B.4 Proof of Lemma B.1

In this appendix, we prove the following lemma which is used in Appendix B.2.

Lemma B.1. *Consider $g \in \mathcal{G}$, $T \in \mathbb{T} = \{1, \ldots, \lfloor \frac{D}{2} \rfloor\}$, $r > \lambda$, a family of g-smoothly-scaling bit-arrival processes characterized by the limiting g-scaled log moment generation function Λ, and a periodic batch service of rNT bits at timeslots mT, $m \in \mathbb{Z}$. Let $Q_i^{(N)}$ be the queue length at time $i \in \{0, \ldots, T-1\}$. Then, the event $\mathcal{Q}_{T-k-1}^{(N)}$, defined as*

$$\mathcal{Q}_{T-k-1}^{(N)} = \left\{ \omega \in \Omega^{(N)} : k+1+\left\lceil \frac{Q_{T-k-1}^{(N)}(\omega)}{RT} \right\rceil T > D \right\},$$

with $k = D(\bmod T)$, asymptotically dominates $P_{delay}(r,T)$. In other words,

$$P_{delay}(r,T) \overset{g}{=} \Pr\big(Q_{T-k-1}^{(N)} > (D-T-k)r \log \rho\big). \tag{B.23}$$

Proof. Let $k = D(\bmod T)$ and $i \in \{0, \ldots, T-1\}$. Recall from (B.7) that

$$\mathcal{Q}_i^{(N)} = \{\omega \in \Omega^{(N)} : T - i + \left\lceil \frac{Q_i^{(N)}(\omega)}{RT} \right\rceil T > D\}.$$

Now using the observation that, for any $x, y \in \mathbb{R}$,

$$\lceil x \rceil > y \Leftrightarrow \lceil x \rceil > \lfloor y \rfloor \Leftrightarrow x > \lfloor y \rfloor,$$

we have

$$\begin{aligned} \mathcal{Q}_i^{(N)} &= \left\{ \omega : \frac{Q_i^{(N)}(\omega)}{RT} > \left\lfloor \frac{D+i-T}{T} \right\rfloor \right\} \\ &= \begin{cases} \big\{\omega : Q_i^{(N)}(\omega) > (D-T-k)R\big\}, & i \in [0, T-k-1] \\ \big\{\omega : Q_i^{(N)}(\omega) > (D-k)R\big\}, & i \in [T-k, T-1]. \end{cases} \end{aligned} \tag{B.24}$$

On the other hand, (B.6) implies that

$$\begin{aligned}
P_{\text{delay}}(r,T) &\overset{g}{=} \sum_{i=0}^{T-1} \Pr(\mathcal{Q}_i^{(N)}) \\
&= \sum_{i=0}^{T-k-1} \Pr(\mathcal{Q}_i^{(N)}) + \sum_{i=T-k}^{T-1} \Pr(\mathcal{Q}_i^{(N)}) \\
&\underset{(a)}{\overset{g}{=}} \Pr(\mathcal{Q}_{T-k-1}^{(N)}) + \Pr(\mathcal{Q}_{T-1}^{(N)}) \\
&\overset{g}{=} \max\{\Pr(\mathcal{Q}_{T-k-1}^{(N)}), \Pr(\mathcal{Q}_{T-1}^{(N)})\} \\
&\underset{(b)}{=} \max\{\Pr(Q_{T-k-1}^{(N)} > (D-T-k)R), \Pr(Q_{T-1}^{(N)} > (D-k)R)\} \\
&\underset{(c)}{=} \Pr(Q_{T-k-1}^{(N)} > (D-T-k)R), \qquad \text{(B.25)}
\end{aligned}$$

where the equality in (b) is from (B.24). Next, we establish the (asymptotic) equalities (a) and (c). For (a), we first need to show that

$$\sum_{j=0}^{T-k-1} \Pr(\mathcal{Q}_j^{(N)}) \overset{g}{=} \Pr(\mathcal{Q}_{T-k-1}^{(N)}). \qquad \text{(B.26)}$$

To establish this, we first observe that

$$Q_j^{(N)}(\omega) = Q_i^{(N)}(\omega) + \underbrace{A_{i+1}^{(N)}(\omega) + \ldots + A_j^{(N)}(\omega)}_{\geq 0} \geq Q_i^{(N)}(\omega), \qquad \text{(B.27)}$$

for all $\omega \in \Omega^{(N)}$ and $0 \leq i \leq j \leq T-1$. Hence, from (B.24), we have

$$\Pr(\mathcal{Q}_{T-k-1}^{(N)}) \geq \Pr(\mathcal{Q}_i^{(N)}), \quad i \in \{0, \ldots, T-k-1\},$$

which implies

$$\sum_{i=0}^{T-k-1} \Pr(\mathcal{Q}_i^{(N)}) \leq (T-k)\Pr(\mathcal{Q}_{T-k-1}^{(N)}). \qquad \text{(B.28)}$$

On the other hand, from the non-negativity of probability, we have

$$\sum_{i=0}^{T-k-1} \Pr(\mathcal{Q}_i^{(N)}) \geq \Pr(\mathcal{Q}_{T-k-1}^{(N)}). \qquad \text{(B.29)}$$

Combining (B.28) and (B.29), we have (B.26). Similarly, we can show that

$$\sum_{j=T-k}^{T-1} \Pr(\mathcal{Q}_j^{(N)}) \overset{g}{=} \Pr(\mathcal{Q}_{T-1}^{(N)}). \tag{B.30}$$

Combining (B.26) and (B.30), equality (a) in (B.25) is established.

To establish equality (c), it is sufficient to show that

$$\Pr(Q_0^{(N)} > D'R) \leq \Pr(Q_j^{(N)} > D'R) \leq \Pr(Q_0^{(N)} > (D'-T)R), \tag{B.31}$$

for any $D' > T$ and $j \in \{0, \ldots, T-1\}$. This is because for $j_1 = T-1$ and $D'_1 = D-k$, we get

$$P\big(Q_{T-1}^{(N)} > (D-k)R\big) \leq P\big(Q_0^{(N)} > (D-T-k)R\big),$$

while for $j_2 = T-k-1$ and $D'_2 = D-T-k$, we get

$$P\big(Q_0^{(N)} > (D-T-k)R\big) \leq P\big(Q_{T-k-1}^{(N)} > (D-T-k)R\big),$$

asserting (c).

We prove (B.31) in two steps. The lower bound directly follows from (B.27), i.e.,

$$Q_j^{(N)}(\omega) \geq Q_0^{(N)}(\omega), \ \forall \omega \in \Omega^{(N)}.$$

For the upper bound, we notice that, for $D' > T$ and

$$\omega \in \Big\{\omega \in \Omega^{(N)} : Q_j^{(N)}(\omega) > D'R\Big\} \subseteq \Big\{\omega \in \Omega^{(N)} : Q_j^{(N)}(\omega) > TR\Big\},$$

$Q_j^{(N)}(\omega)$ is related to $Q_T^{(N)}(\omega)$ as

$$\begin{aligned} Q_T^{(N)}(\omega) &= [Q_j^{(N)}(\omega) + A_{j+1}^{(N)}(\omega) + \cdots + A_T^{(N)}(\omega) - TR]^+ \\ &= Q_j^{(N)}(\omega) + A_{j+1}^{(N)}(\omega) + \cdots + A_T^{(N)}(\omega) - TR, \end{aligned}$$

where $[\cdot]^+$ is removed. As a result, we have

$$\begin{aligned}\Pr(Q_j^{(N)} > D'R) &= \Pr(Q_T^{(N)} - \left\{A_{j+1}^{(N)} + \ldots + A_T^{(N)}\right\} + TR > D'R) \\ &\le \Pr(Q_T^{(N)} > (D' - T)R) \\ &= \Pr(Q_0^{(N)} > (D' - T)R),\end{aligned}$$

where the last equality holds since $Q_T^{(N)}$ and $Q_0^{(N)}$ have the same stationary distribution. □

B.5 Proof of Lemma B.2

This appendix shows that the average probability of delay violation for bits that arrive at time i is asymptotically equal to the corresponding probability for the last bit arriving at that time. The proof is mainly based on the definition of the g-smoothly-scaling process.

Lemma B.2. *Consider $g \in \mathcal{G}$ and a family of g-smoothly-scaling bit-arrival processes $((A_t^{(N)}, t \in \mathbb{Z}), N \in \mathbb{N})$, characterized by the limiting g-scaled log moment generation function Λ. For any given N, let $W^{(N)}$ be a random variable having the same distribution as the steady-state distribution of the delay of a randomly chosen bit that arrives at time $i \in \{0, \ldots, T-1\}$ while $Z^{(N)}$ is a random variable having a distribution that is identical to the steady-state distribution of the delay for the last bit that arrives during time i. Then, for any $D > 0$,*

$$Pr(W^{(N)} > D) \overset{g}{=} Pr(Z^{(N)} > D). \tag{B.32}$$

Proof. We show (B.32) by showing the upper bound:

$$\Pr(W^{(N)} > D) \leq \Pr(Z^{(N)} > D) \tag{B.33}$$

and the lower bound:

$$\Pr(W^{(N)} > D) \overset{g}{\geq} \Pr(Z^{(N)} > D). \tag{B.34}$$

The upper bound is an immediate consequence of $W^{(N)}(\omega) \leq Z^{(N)}(\omega)$ for $\omega \in \Omega^{(N)}$. Below we prove the lower bound. We have

$$\Pr(W^{(N)} > D) = \sum_{a\in\mathbb{N}} \Pr(W^{(N)} > D|A_i^{(N)} = a)\Pr(A_i^{(N)} = a). \tag{B.35}$$

Now, given that $A_i^{(N)} = a$ bits arrive at time i, we index the a bits as bit 1 to a, where bit 1 arrives first and bit a arrives last. Given $A_i^{(N)} = a$, we let $W_j^{(N)}$ to be the steady-state delay of the j-th bit, $j \in \{1, \ldots, a\}$. Since the bit can have any index, from 1 to a, with equal probability of $1/a$, we have

$$\Pr(W^{(N)} > D|A_i^{(N)} = a) = \frac{1}{a}\sum_{j=1}^{a} \Pr(W_j^{(N)} > D|A_i^{(N)} = a).$$

Ignoring all but the last term in the sum, we have

$$\Pr(W^{(N)} > D|A_i^{(N)} = a) \geq \frac{1}{a}\Pr(W_a^{(N)} > D|A_i^{(N)} = a) = \frac{1}{a}\Pr(Z^{(N)} > D|A_i^{(N)} = a),$$

where the equality is a result of how $Z^{(N)}$ is defined. This means that

$$\begin{aligned}\Pr(W^{(N)} > D) &\geq \sum_{a\in\mathbb{N}} \frac{1}{a}\Pr(Z^{(N)} > D|A_i^{(N)} = a)\Pr(A_i^{(N)} = a)\\ &= \sum_{a\in\mathbb{N}} \frac{1}{a}\Pr(Z^{(N)} > D \text{ and } A_i^{(N)} = a).\end{aligned}$$

Now, for a given $\beta > 0$, define

$$B^{(N)} := \{b \in \mathbb{N} : b < e^{\beta g(N)}\}.$$

We can further lower bound $\Pr(W^{(N)} > D)$ as follows:

$$\begin{aligned}
\Pr(W^{(N)} > D) &\geq \sum_{a \in B^{(N)}} \frac{1}{a} \Pr(Z^{(N)} > D \text{ and } A_i^{(N)} = a) \\
&\geq e^{-\beta g(N)} \sum_{a \in B^{(N)}} \Pr(Z^{(N)} > D \text{ and } A_i^{(N)} = a) \\
&= e^{-\beta g(N)} \Pr(Z^{(N)} > D \text{ and } A_i^{(N)} \in B^{(N)}), \qquad \text{(B.36)}
\end{aligned}$$

where the second inequality holds because $1/a > e^{-\beta g(N)}$ for any $a \in B^{(N)}$.

Next, we show that $\Pr(A_i^{(N)} \in B^{(N)}) \to 1$ as $N \to \infty$. We do this by using the definition of the g-smoothly-scaling process: there exists $\theta > 0$ such that

$$\lim_{N \to \infty} \frac{\log E[e^{\theta A_i^{(N)} g(N)/N}]}{g(N)} = \Lambda(\theta) < \infty.$$

Hence, for any $\epsilon > 0$, there exists $N_0 = N_0(\epsilon)$ such that for all $N > N_0$, we have

$$g(N)(\Lambda(\theta) + \epsilon) > \log E[e^{\theta A_i^{(N)} g(N)/N}]. \qquad \text{(B.37)}$$

The RHS can be lower-bounded, for any $a_1 \in \mathbb{N}$:

$$\begin{aligned}
\log E[e^{\theta A_i^{(N)} g(N)/N}] &= \log \left(\sum_{a \in \mathbb{N}} \Pr(A_i^{(N)} = a) e^{\theta a g(N)/N} \right) \\
&\geq \log \left(\sum_{a \geq a_1} \Pr(A_i^{(N)} = a) e^{\theta a g(N)/N} \right) \\
&\geq \log \left(\Pr(A_i^{(N)} \geq a_1) e^{\theta a_1 g(N)/N} \right) \\
&= \theta a_1 \frac{g(N)}{N} + \log \Pr(A_i^{(N)} \geq a_1).
\end{aligned}$$

This together with (B.37) gives

$$\log \Pr(A_i^{(N)} \geq a_1) < g(N)[\Lambda(\theta) + \epsilon - \frac{\theta a_1}{N}],$$

for all $a_1 \in \mathbb{N}$. Now, we select $a_1 = e^{\beta g(N)}$ to get

$$\log \left(1 - \Pr(A_i^{(N)} \in B^{(N)})\right) = \log \Pr(A_i^{(N)} \geq e^{\beta g(N)}) < g(N)[\Lambda(\theta) + \epsilon - \frac{\theta e^{\beta g(N)}}{N}].$$

Since $\lim\limits_{N\to\infty} \frac{g(N)}{\log N} = \infty$, we, then, have

$$\Pr\left(A_i^{(N)} \in B^{(N)}\right) \to 1. \tag{B.38}$$

Finally, combining (B.38) and (B.36) implies that, for any $\beta > 0$,

$$\lim_{N\to\infty} \frac{\log \Pr(W^{(N)} > D)}{g(N)} \geq \lim_{N\to\infty} \frac{\log \Pr(Z^{(N)} > D)}{g(N)} - \beta.$$

Since β can be chosen arbitrarily small, we have the lower bound in (B.34), hence the assertion of the lemma. □

B.6 Proof of Lemma B.3

In this appendix, we prove the following lemma which is used in Appendix B.2.

Lemma B.3. *Consider $g \in \mathcal{G}$, $T \in \mathbb{T} = \{1, \dots, \lfloor \frac{D}{2} \rfloor\}$, $r > \lambda$, a family of g-smoothly-scaling bit-arrival processes characterized by the limiting g-scaled log moment generation function Λ, and a periodic batch service of rNT bits at timeslots mT, $m \in \mathbb{Z}$. Let $Q_i^{(N)}$ be the queue length at time $i \in \{0, \dots, T-1\}$. Then, for*

$q > ir$, we have

$$\limsup_{N\to\infty} \frac{\log Pr(Q_i^{(N)} > Nq)}{g(N)} \leq -\inf_{\substack{t\in\mathbb{Z}^+:\\ tT+i>0}} (Tt+i)\Lambda^*\left(\frac{q+rTt}{Tt+i}\right), \tag{B.39}$$

assuming that the RHS is strictly greater than $-\infty$.

Proof. The proof uses the same technique as in [29, Lemma 1.10 and 1.11]. Using (B.11), we have the following bound:

$$\begin{aligned}\Pr(Q_i^{(N)} > Nq) &= \Pr\left(\sup_{\substack{t\in\mathbb{Z}^+:\\ tT+i>0}} \sum_{j=-tT+1}^{i} A_j^{(N)} - rtTN > Nq\right)\\ &= \Pr\left(\sup_{\substack{t\in\mathbb{Z}^+:\\ tT+i>0}} S_{tT+i}^{(N)} - rtTN > Nq\right)\\ &\leq \sum_{t:t>-\frac{i}{T}} \Pr\left(S_{tT+i}^{(N)} > N(q+rTt)\right).\end{aligned}$$

Now, for any fixed $t_0 \in \mathbb{N}$, we have

$$\Pr(Q_i^{(N)} > Nq) \leq \sum_{-\frac{i}{T}<t\leq t_0} \Pr(S_{tT+i}^{(N)} > N(q+rTt)) + \sum_{t>t_0} \Pr(S_{tT+i}^{(N)} > N(q+rTt)). \tag{B.40}$$

Employing the principle of the largest term[1] gives

$$\begin{aligned}&\limsup_{N\to\infty} \frac{\log \Pr(Q_i^{(N)} > Nq)}{g(N)}\\ &\leq \max\Bigg(\max_{-\frac{i}{T}<t\leq t_0} \limsup_{N\to\infty} \frac{\log \Pr(S_{tT+i}^{(N)} > N(q+rTt))}{g(N)},\\ &\qquad \limsup_{N\to\infty} \frac{1}{g(N)} \log \sum_{t>t_0} \Pr(S_{tT+i}^{(N)} > N(q+rTt))\Bigg).\end{aligned} \tag{B.41}$$

[1] The principle of the largest term [29, Lemma 2.1]: Let a_n and b_n be sequences in $\mathbb{R}^+$. If $n^{-1}\log a_n \to a$ and $n^{-1}\log b_n \to b$, then $n^{-1}\log(a_n + b_n) \to \max(a, b)$. This extends easily to finite sums.

For the first term (the $t \leq t_0$ term) in the maximum, we use Proposition 3.6 to get

$$\max_{-\frac{i}{T}<t\leq t_0} \limsup_{N\to\infty} \frac{1}{g(N)} \log \Pr\left(\frac{S_{tT+i}^{(N)}}{N} > q + rTt\right) \leq \max_{-\frac{i}{T}<t\leq t_0} -(Tt+i)\Lambda^*\left(\frac{q+rTt}{Tt+i}\right)$$
$$\leq - \inf_{\substack{t\in\mathbb{Z}^+:\\ tT+i>0}} (Tt+i)\Lambda^*\left(\frac{q+rTt}{Tt+i}\right), \tag{B.42}$$

which is the RHS of (B.39) and finite by assumption.

Now, we show that we can select t_0 appropriately such that the second term (the $t > t_0$ term) in the RHS of (B.41) is also no greater than the RHS of (B.39). In other words, we show that there exists t_0 such that

$$\limsup_{N\to\infty} \frac{1}{g(N)} \log \sum_{t>t_0} \Pr\left(S_{tT+i}^{(N)} > N(q+rTt)\right) \leq - \inf_{\substack{t\in\mathbb{Z}^+:\\ tT+i>0}} (Tt+i)\Lambda^*\left(\frac{q+rTt}{Tt+i}\right). \tag{B.43}$$

This is shown by proving that there exist some $\theta > 0$ and $\epsilon > 0$ such that

$$\limsup_{N\to\infty} \frac{1}{g(N)} \log \sum_{t>t_0} \Pr\left(S_{tT+i}^{(N)} > N(q+rTt)\right) \leq -\epsilon\theta\big((t_0+1)T+i\big), \tag{B.44}$$

for all $t_0 \in \mathbb{N}$. Now, selecting

$$t_0 = \left\lceil \frac{1}{\epsilon\theta T} \inf_{\substack{t\in\mathbb{Z}^+:\\ tT+i>0}} (Tt+i)\Lambda^*\left(\frac{q+rTt}{Tt+i}\right) \right\rceil$$

provides (B.43).

To prove (B.44), we first use Chernoff bound as follows:

$$\begin{aligned}
&\sum_{t>t_0} \Pr(S_{tT+i}^{(N)} > N(q+rTt)) \\
&= \sum_{t>t_0} \Pr\left(e^{\frac{\theta g(N)}{N} S_{tT+i}^{(N)}} > e^{\frac{\theta g(N)}{N} N(q+rTt)}\right) \\
&\leq \sum_{t>t_0} e^{-\theta g(N)(q+rTt)} E[e^{\theta S_{tT+i}^{(N)} \frac{g(N)}{N}}] \\
&= \sum_{t>t_0} e^{-\theta g(N)(q+rTt)} (E[e^{\theta A_1^{(N)} \frac{g(N)}{N}}])^{tT+i} \\
&= \sum_{t>t_0} \exp\left(-g(N)(tT+i)\left[\theta\left(\frac{q+rtT}{tT+i}\right) - \frac{\log E[e^{\frac{\theta g(N)}{N} A_1^{(N)}}]}{g(N)}\right]\right), \qquad \text{(B.45)}
\end{aligned}$$

where θ is an arbitrary positive scalar and the second equality is a consequence of i.i.d. assumption on $A_t^{(N)}$.

Next, we use the convexity of Λ and the fact that $\Lambda'(0) = \lambda < r$ (Remark 3.5) to establish that there exist some $\theta > 0$ and $\epsilon > 0$ for which

$$\Lambda(\theta) < \theta(r - 2\epsilon). \qquad \text{(B.46)}$$

On the other hand, from (3.6), we know that $\frac{\log E[e^{\frac{\theta g(N)}{N} A_1^{(N)}}]}{g(N)} \to \Lambda(\theta)$. This means that there exists a $N_0 = N_0(\theta, \epsilon)$ such that, for all $N > N_0$,

$$\frac{\log E[e^{\frac{\theta g(N)}{N} A_1^{(N)}}]}{g(N)} < \Lambda(\theta) + \theta\epsilon.$$

Combining this with (B.46), we have

$$\frac{\log E[e^{\frac{\theta g(N)}{N} A_1^{(N)}}]}{g(N)} < \theta(r - 2\epsilon) + \theta\epsilon = \theta(r - \epsilon), \qquad \text{(B.47)}$$

for all $N > N_0$.

Hence, using (B.47), the term inside the square bracket in (B.45) can be bounded, uniformly over all $t > t_0$, as

$$\begin{aligned}
\theta\left(\frac{q+rtT}{tT+i}\right) - \frac{\log E[e^{\frac{\theta g(N)}{N}A_1^{(N)}}]}{g(N)} &= \theta\left(r + \frac{q-ir}{tT+i}\right) - \frac{\log E[e^{\frac{\theta g(N)}{N}A_1^{(N)}}]}{g(N)} \\
&> \theta r - \frac{\log E[e^{\frac{\theta g(N)}{N}A_1^{(N)}}]}{g(N)} \\
&> \theta r - \theta(r-\epsilon) \\
&= \theta\epsilon, \qquad \text{(B.48)}
\end{aligned}$$

where the first equality holds because $q > ir$, by assumption.

Inserting (B.48) into (B.45), we have (B.44):

$$\begin{aligned}
&\limsup_{N\to\infty} \frac{1}{g(N)} \log \sum_{t>t_0} \Pr\left(S_{tT+i}^{(N)} > N(q+rTt)\right) \\
&\leq \limsup_{N\to\infty} \frac{1}{g(N)} \log \sum_{t>t_0} \exp\left(-g(N)(tT+i)\theta\epsilon\right) \\
&= \limsup_{N\to\infty} \frac{1}{g(N)} \log\left(\frac{e^{-g(N)\theta\epsilon((t_0+1)T+i)}}{1-e^{-g(N)\theta\epsilon T}}\right) \\
&= -\epsilon\theta\big((t_0+1)T+i\big),
\end{aligned}$$

and, hence, the assertion of the lemma. □

Appendix C

Appendix for Chapter 4

C.1 Additional Assumption on Source Model

Here we give the additional assumption B on the source we consider in this paper. Assumption B is required in the proof of Lemma 4.12.

Assumption B. *(Sample path LDP for partial sum process (see [9, 89]))*

For an arrival sequence $\{S_1, S_2, \ldots\}$, for all $m \in \mathbb{N}$, for every $\epsilon_1, \epsilon_2 > 0$, and for every scalar $a_0, \ldots, a_{m-1}$, there exists $M > 0$ such that for all $n \geq M$ and all $k_0, \ldots, k_m$ with $1 = k_0 \leq k_1 \leq \cdots \leq k_m = n$,

$$\begin{aligned}
&\exp\left\{-n\epsilon_2 - \sum_{i=1}^{m-1}(k_{i+1} - k_i)\Lambda^*(a_i)\right\} \\
&\leq \Pr\left[\left|S_{k_{i+1}} - S_{k_i} - (k_{i+1} - k_i)a_i\right| \leq n\epsilon_1, i = 1, \ldots, m-1\right] \\
&\leq \exp\left\{n\epsilon_2 - \sum_{i=1}^{m-1}(k_{i+1} - k_i)\Lambda^*(a_i)\right\}
\end{aligned}$$

C.2 Proofs of Lemmas and Proposition

C.2.1 Proof of Lemma 4.9

Proof of Lemma 4.9. Since the symmetric static scheduler always assigns the service rate $C_{av}(d)$ to each queue, we have that

$$\Pr[D^1 > D] = \Pr[Q^1 > DC_{av}(d)]. \tag{C.1}$$

To be more specific, the statement holds because any bits delayed more than D timeslots see at least $DC_{av}(d)$ bits before them, and any bits delayed less than D timeslots must see less than $DC_{av}(d)$ bits before them. This is valid because of the first-come-first-serve discipline assumption. Hence, the two events $\{D^1 > D\}$ and $\{Q^1 > DC_{av}(d)\}$ are equivalent and have the same probability.

Now, since $\Pr[Q^1 > DC_{av}(d)]$ is equal to the tail probability for a buffer which is served at fixed capacity of $C_{av}(d)$ and whose arrival process is described by $\Lambda(\cdot)$ and satisfying LDP, one can calculate the tail probability for a single queue system with a fixed service rate c as (see [41], [29] and [80])

$$\lim_{B\to\infty} \frac{1}{B} \log \Pr[Q^1 > B] = -\theta^*$$

where θ^* is the largest positive root of equation $\frac{\Lambda(\theta)}{\theta} = c$. By replacing B with $DC_{av}(d)$ and c with $C_{av}(d)$ and using (4.10), we have

$$\lim_{\rho\to\infty} \frac{\log \Pr[Q^1 > DC_{av}(d)]}{\log \rho} = -\sigma_s(d) D Tr_{av}(d)$$

where $\sigma_s(d)$ is given as the solution to $\Lambda(\sigma_s(d)) = \sigma_s(d) C_{av}(d)$. □

C.2.2 Proof of Lemma 4.12

Before showing the proof of Lemma 4.12, we recall the following result on the asymptotic tail probability of the maximal weighted delay under the longest-weighted-delay-first (LWDF) scheduling discipline from [72] and simplify the result to our specific assumptions of symmetric users with LDF scheduling discipline.

Fact C.1. *(Theorem 2.2 in [72]) Consider a single server of fixed service rate 1 and K mutually independent source processes with stationary increments. The total number of information bits generated by source i ($i = 1, \dots, K$) is given by a sequence $\left\{\hat{S}_t^i, t = 1, 2, \dots\right\}$ where $\hat{S}_t^i$ is the cumulative total number of work arrived until time t from source i. We assume $\left\{\hat{S}_t^i, t = 1, 2, \dots\right\}$ satisfies LDP and sample path LDP (Assumptions A and B) with the convex decay function $\hat{\Lambda}_i^*$ and the convex log moment generating function $\hat{\Lambda}_i$. Assume $KE[\hat{S}_1^1] < 1$ for stability. Let α_i be the weight for user i (assuming $0 < \alpha_1 \leq \alpha_2 \leq \cdots \leq \alpha_K$). Consider the longest-weighted-delay-first (LWDF) scheduling discipline, which always assign the server to the longest waiting (i.e. head-of-the-line)) customer of the source i which has the maximal weighted delay. Then, the LWDF scheduling discipline maximizes the exponential decay rate of the stationary distribution of the maximal delay, among all causal and work-conserving scheduling disciplines. Furthermore, the probability is given as*

$$\limsup_{n\to\infty} \frac{1}{n} \log \Pr\left[\frac{1}{n} \max_{i \in 1,\dots,K} \hat{w}^i > 1\right] \leq -J_* \tag{C.2}$$

where and $\hat{w}^i$ is the stationary delay for user i, $i = 1\ldots, K$, and J_* is given as:

$$J_* = \min_{j;x_1,\ldots,x_j} \frac{1}{\gamma} \sum_{i=1}^{j} (1-\alpha_i\gamma)\hat{\Lambda}_i^*(x_i) \tag{C.3}$$

subject to

$$j \in \{1,\ldots,K\}, x_i > 0, \sum_{i=1}^{j} x_i > 1 \tag{C.4}$$

and

$$\frac{1}{\alpha_{j+1}} < \gamma = \frac{\sum_{i=1}^{j} x_i - 1}{\sum_{i=1}^{j} \alpha_i x_i} \le \frac{1}{\alpha_j} \tag{C.5}$$

with $\alpha_{K+1} \equiv \infty$.

Since in this paper we consider symmetric users and LDF scheduling which is the LWDF discipline with equal weights, the following corollary gives a specific expression of J_* which will be used to show Lemma 4.12.

Corollary C.2. *Under the assumptions of symmetric users with equal weights, i.e. $\hat{\Lambda}_i^* = \hat{\Lambda}^*$, $\hat{\Lambda}_i = \hat{\Lambda}$, and $\alpha_i = 1$ for all $i = 1,\ldots,K$, J_* in Fact C.1 is reduced to*

$$J_* = \sup\left\{\theta : \hat{\Lambda}(\theta) \le \theta/K\right\}. \tag{C.6}$$

Proof. Under the assumption of equal weights (i.e. $\alpha_i = 1$ for all $i = 1,\ldots,K$), there are feasible values of γ in (C.5) only when $j = K$. Hence, the minimization in (C.3) is reduced to

$$J_* = \min_{x_1,\ldots,x_K} \frac{1-\gamma}{\gamma} \sum_{i=1}^{K} \hat{\Lambda}^*(x_i) \tag{C.7}$$

subject to

$$\sum_{i=1}^{K} x_i > 1, x_i > 0, \ i = 1,\ldots,K \tag{C.8}$$

and

$$\frac{1}{\alpha_{K+1}} = 0 < \gamma = \frac{\sum_{i=1}^{K} x_i - 1}{\sum_{i=1}^{K} x_i} \leq 1. \tag{C.9}$$

However, we notice that condition (C.9) is satisfied with any choices of $\{x_i\}$ satisfying condition (C.8). Hence, plugging the expression of γ into (C.7), we get

$$J_* = \min_{x_1,\ldots,x_K} \frac{1}{\sum_{i=1}^{K} x_i - 1} \sum_{i=1}^{K} \hat{\Lambda}^*(x_i) \tag{C.10}$$

subject to (C.8).

We can simplify J_* further by using the convexity property of $\hat{\Lambda}^*$, i.e.

$$\frac{1}{K}\sum_{i=1}^{K} \hat{\Lambda}^*(x_i) \geq \hat{\Lambda}^*(\frac{\sum_{i=1}^{K} x_i}{K}) =: \hat{\Lambda}^*(\frac{Ka+1}{K}),$$

where we let a be such that $Ka = \sum_{i=1}^{K} x_i - 1 > 0$ by the condition in (C.8). The equality holds when $x_i = a + 1/K$ for all $i = 1, \ldots, K$. Hence, we can rewrite (C.10) and its conditions concisely as

$$J_* = \min_{a>0} \frac{\hat{\Lambda}^*\left(a + \frac{1}{K}\right)}{a}. \tag{C.11}$$

To finish the proof, we expand $\hat{\Lambda}^*$ using its definition, as follows:

$$\begin{aligned}
J_* &= \min_{a>0} \frac{1}{a}\hat{\Lambda}^*\left(a + 1/K\right) \\
&= \min_{a>0} \frac{1}{a} \sup_{\theta\in\mathbb{R}} \theta(a + 1/K) - \hat{\Lambda}(\theta) \\
&= \sup_{\theta\in\mathbb{R}} \min_{a>0} \theta + \frac{\theta/K - \hat{\Lambda}(\theta)}{a} \\
&= \sup_{\theta\in\mathbb{R}} \begin{cases} -\infty, & \text{if } \theta/K < \hat{\Lambda}(\theta), \\ \theta, & \text{if } \theta/K \geq \hat{\Lambda}(\theta) \end{cases} \\
&= \sup\left\{\theta : \hat{\Lambda}(\theta) \leq \theta/K\right\},
\end{aligned}$$

where the third equality holds because the function $\theta + \frac{\theta/K - \hat{\Lambda}(\theta)}{a}$ is convex on a and concave on θ (since $\hat{\Lambda}$ is convex). □

Proof of Lemma 4.12. Consider a scaled version of the system in Fact C.1 where the service rate is scaled to C (which is equal to KC_{av}) and the arrivals are also scaled up by C. We can think of this scaling as a change of measurement units. We denote D_s^i as the stationary delay of arrivals of user i, $i = 1, \ldots, K$, for this scaled single-server system with LDF scheduling. Since scaling of the service and arrivals do not change the distribution of the delays, we have from Fact C.1 that

$$\limsup_{n \to \infty} \frac{1}{n} \log \Pr\left[\frac{1}{n} \max_{i \in 1, \ldots, K} D_s^i > 1\right] \leq -J_*. \tag{C.12}$$

Noticing that the log moment generating function Λ of the scaled system is given as

$$\Lambda(\theta) = \lim_{t \to \infty} \frac{1}{t} \log E[e^{\theta C \hat{S}_t^1}] = \hat{\Lambda}(\theta C), \tag{C.13}$$

we have, by using Corollary C.2,

$$\begin{aligned} J_* &= \sup\left\{\theta : \hat{\Lambda}(\theta) \leq \theta/K\right\} \\ &= \sup\left\{\theta : \Lambda(\theta/C) \leq \theta/K\right\} \\ &= C \sup\left\{\tilde{\theta} : \Lambda(\tilde{\theta}) \leq \tilde{\theta}C/K = \tilde{\theta}C_{av}\right\} \\ &= C\sigma_s \end{aligned} \tag{C.14}$$

where $\sigma_s > 0$ is defined as the unique solution to $\Lambda(\sigma_s) = \sigma_s C_{av}$. The second equality in (C.14) follows by using (C.13); the third equality follows by letting $\tilde{\theta} = \theta/C$; and the last equality by using that fact that Λ is strictly convex and

$\Lambda'(0) = \lambda T \log \rho < C_{av}$ (the stability condition in (4.8) and the fact that $\hat{\Lambda}'(0)$ is the average arrival rate per source [41]) and hence the supremum is attained with $\tilde{\theta} = \sigma_s$.

Replacing n with D and $J_* = C\sigma_s = KC_{\text{av}}\sigma_s = K\sigma_s Tr_{\text{av}} \log \rho$ in (C.12), we have

$$\Pr\left[\max_{i\in 1,\ldots,K} D_s^i > D\right] \doteq \rho^{-K\sigma_s DTr_{\text{av}}},$$

for large value of D.

From symmetry, on the other hand, we have

$$\Pr[D_s^1 > D] \leq \Pr\left[\max_{i\in 1,\ldots,K} D_s^i > D\right] \leq K\Pr[D_s^1 > D].$$

This provides the assertion of the lemma:

$$P^l(d) := \Pr[D_s^1 > D] \doteq \rho^{-K\sigma_s DTr_{\text{av}}}.$$

□

C.2.3 Proof of Theorem 4.1

Proof of Theorem 4.1. We first show the existence of a solution d in (4.28) of Algorithm 1. The LHS term is decreasing on d and equal to 0 for $d \geq \bar{d}$, for some $\bar{d}$ such that the arrival rate $\lambda T \log \rho$ is equal to the service rate $C_{av}(\bar{d})$ (in which case, $\sigma_s(\bar{d}) = 0$). On another hand, the RHS term is increasing on d and is equal to 0 when $d = 0$. Hence, (4.28) must hold for some $d \in (0, \bar{d})$. Next, if d solving (4.28) is less than d_0, this d is the lower bound d_l^* (i.e. asymptotically maximizing the RHS term in (4.27)) and the upper bound d_u^* is obtained from maximizing the LHS in (4.27). The existence of d solving (4.29) can be shown similarly. □

C.2.4 Proof of Proposition 4.18

Proof of Proposition 4.18. The limiting log moment generating function $\Lambda(\cdot)$ for compound Poisson source with exponential packet length is derived in [44], which is

$$\begin{aligned}
\Lambda(\theta) &= \lim_{n\to\infty}\frac{1}{n}\log\mathbb{E}\left[\exp(\theta S_t^1)\right] \\
&= \lim_{n\to\infty}\frac{1}{n}\log\mathbb{E}\left[\exp(\theta\sum_{t=1}^{n}A_t^1)\right] \\
&= \log\mathbb{E}\left[\exp\theta A_1^1\right] \\
&= \begin{cases} \frac{\nu\theta}{\mu-\theta} & \text{if } \theta<\mu, \\ \\ \infty & \text{if } \theta\ge\mu. \end{cases} \qquad (C.15)
\end{aligned}$$

From Lemma 4.9, $\sigma_s(d)$ is the solution to $\Lambda(\sigma_s(d)) = \sigma_s(d)C_{av}(d)$. From (C.15), this reduces to finding $\sigma_s(d)$ such that

$$\begin{aligned}
\frac{\nu\sigma_s(d)}{\mu-\sigma_s(d)} &= \sigma_s(d)C_{av}(d) \\
\Leftrightarrow\ \frac{\mu\lambda T\log\rho}{\mu-\sigma_s(d)} &= r_{av}(d)T\log\rho
\end{aligned}$$

where we have replaced ν and $C_{\text{av}}(d)$ from (4.32) and (4.10). Hence, $\sigma_s(d) = \mu(1-\lambda/r_{av}(d))$. □

Appendix D

Appendix for Chapter 5

D.1 Proof of Lemma 5.12

Here we prove Lemma 5.12 which uses the following fact and lemmas, all three of which directly result from the definitions of quasi-continuity, continuity, and scheduler assignment $H(\cdot)$.

Fact D.1. *Assume $\mathcal{X} \overset{F,G}{\rightarrow} \mathcal{Y}$, $\mathcal{X}, \mathcal{Y}$ are metric spaces, and $x \in \mathcal{X}$. If F is quasi-continuous at x and G is continuous at x, then $F + G$ is quasi-continuous at x.*

Lemma 5.12. *For $t \in \mathbb{N}$, G_t is quasi-continuous on $\mathbb{R}_+^{K\times t}$ with respect to the uniform topology.*

Proof. Using our queueing equation we first observe the following recursive relation between G_t and G_{t-1} for any $t \in \{2, 3, \ldots\}$ and $\mathbf{x} = \mathbf{x}_{(0,t]} \in \mathbb{R}_+^{K\times t}$:

$$G_t(\mathbf{x}_{(0,t]}) = [G_{t-1}(\mathbf{x}_{(1,t]}) - H(G_{t-1}(\mathbf{x}_{(1,t]}))]^+ + \mathbf{x}_1, \tag{D.1}$$

where we used the fact that $Q_0(\mathbf{x}_{(0,t]}) = G_t(\mathbf{x}_{(0,t]})$, and $Q_1(\mathbf{x}_{(1,t]}) = G_{t-1}(\mathbf{x}_{(1,t]})$ when the initial backlog at time $-t$ is small, i.e., $Q_t \in \mathcal{R}$.

Equation (D.1) says that $G_t(\mathbf{x}_{(0,t]})$ depends linearly on $\mathbf{x}_1$. This implies the following simple but consequential observations:

Observation D.2. If G_t is strictly quasi-continuous at $\mathbf{x}_{(0,t]}$, then it is strictly quasi-continuous at $\tilde{\mathbf{x}}_{(0,t]} := (\tilde{\mathbf{x}}_1, \mathbf{x}_2, \ldots, \mathbf{x}_t)$ for any $\tilde{\mathbf{x}}_1 \in \mathbb{R}_+^K$. If G_t is continuous at $\mathbf{x}_{(0,t]}$, then it is also continuous at $\tilde{\mathbf{x}}_{(0,t]}$.

Observation D.3. If $G_t(\mathbf{x}^n_{(0,t]}) \to G_t(\mathbf{x}_{(0,t]})$ for a sequence $\{\mathbf{x}^n_{(0,t]}\}$ such that $\mathbf{x}^n_{(0,t]} \to \mathbf{x}_{(0,t]}$, then for any sequence $\{\tilde{\mathbf{x}}^n_{(0,t]} = (\tilde{\mathbf{x}}^n_1, \mathbf{x}^n_2, \ldots, \mathbf{x}^n_t)\}$ where $\tilde{\mathbf{x}}^n_1 \to \mathbf{x}_1$, we also have $G_t(\tilde{\mathbf{x}}^n_{(0,t]}) \to G_t(\mathbf{x}_{(0,t]})$.

Using the recursive relation in (D.1), we prove this lemma by induction on $t \in \mathbb{N}$. For $t = 1$, $G_1(\mathbf{a}_1) = \mathbf{a}_1$, hence G_1 is continuous on $\mathbb{R}_+^K$. Assuming that G_t is quasi-continuous on $\mathbb{R}_+^{Kt}$, we want to show that G_{t+1} is quasi-continuous on $\mathbb{R}_+^{K(t+1)}$. Using the fact that the $[\cdot]^+$ function is continuous, Remark 5.10, and Fact D.1, it suffices to show that the function $F_t := G_t - H \circ G_t$, is quasi-continuous on $\mathbb{R}_+^{Kt}$ to show that G_{t+1} is quasi-continuous. In particular, for any arrival sample path $\mathbf{a} = \mathbf{a}_{(0,t]} \in \mathbb{R}_+^{K\times t}$, we need to show that F_t is quasi-continuous at $\mathbf{a}_{(0,t]}$ with respect to the uniform topology. It suffices to show that it is possible to select a sequence $\hat{\mathbf{a}}^n \to \mathbf{a}$ for which

$$G_t(\hat{\mathbf{a}}^n_{(0,t]}) \quad \to G_t(\mathbf{a}_{(0,t]}), \tag{D.2}$$

$$H \circ G_t(\hat{\mathbf{a}}^n_{(0,t]}) \quad \to H \circ G_t(\mathbf{a}_{(0,t]}), \tag{D.3}$$

such that both $G_t(\cdot)$ and $H \circ G_t(\cdot)$ are continuous at every $\hat{\mathbf{a}}^n_{(0,t]}$.

We show this by first noting that the induction hypothesis, i.e., quasi-continuity of G_t, and the definition of quasi-continuity ensure that there exists a sequence $\{\mathbf{a}^n_{(0,t]}\}$ such that $\mathbf{a}^n_{(0,t]} \to \mathbf{a}$, in the uniform topology, such that $G_t(\mathbf{a}^n_{(0,t]}) \to G_t(\mathbf{a}_{(0,t]})$, and $G_t(\cdot)$ is continuous at $\mathbf{a}^n_{(0,t]}$ for all n. We will construct the desired sequence $\{\hat{\mathbf{a}}^n_{(0,t]}\}$ from this sequence $\{\mathbf{a}^n_{(0,t]}\}$. We proceed by considering the following two cases, depending on the value of $\mathbf{a}_1$.

Case 1: $\mathbf{a}_1 > \mathbf{0}$, i.e., every component of the $\mathbf{a}_1 \in \mathbb{R}^K$ is positive. Let $\epsilon > 0$ be the smallest component of $\mathbf{a}_1$. Since $H(\cdot)$ is quasi-continuous, it is possible to choose a sequence of (workload) vectors $\{w^n\}$ such that $w^n \to G_t(\mathbf{a}_{(0,t]})$, and H is continuous at w^n for all n. Now, we define

$$\begin{aligned} \tilde{\mathbf{a}}^n_1 &:= w^n - [F_{t-1}(\mathbf{a}^n_{(1,t]})]^+ && \text{(D.4)} \\ &= w^n - G_t(\mathbf{a}^n_{(0,t]}) + \mathbf{a}^n_1 && \text{(D.5)} \\ &= \big(w^n - G_t(\mathbf{a}_{(0,t]})\big) + \big(G_t(\mathbf{a}_{(0,t]}) - G_t(\mathbf{a}^n_{(0,t]})\big) + (\mathbf{a}^n_1 - \mathbf{a}_1) + \mathbf{a}_1. && \text{(D.6)} \end{aligned}$$

It is clear from the last equality that $\tilde{\mathbf{a}}^n_1 \to \mathbf{a}_1$ with respect to uniform topology. We need to ensure that $\tilde{\mathbf{a}}^n_1 \geq 0$ since negative quantities are involved in the definition. We do this by using the facts that every component of $\mathbf{a}_1 \in \mathbb{R}^K$ is greater or equal to $\epsilon > 0$, and that $w^n \to G_t(\mathbf{a}_{(0,t]})$, $G_t(\mathbf{a}^n_{(0,t]}) \to G_t(\mathbf{a}_{(0,t]})$, and $\mathbf{a}^n_1 \to \mathbf{a}_1$. These facts imply that there exists an n_ϵ such that for all $n > n_\epsilon$ we have $||w^n - G_t(\mathbf{a}_{(0,t]})|| < \epsilon/3$, $||G_t(\mathbf{a}_{(0,t]}) - G_t(\mathbf{a}^n_{(0,t]})|| < \epsilon/3$ and $||\mathbf{a}^n_1 - \mathbf{a}_1|| < \epsilon/3$ (with the L_1 norm) which then together with (D.6) imply that, for the sequence $\tilde{\mathbf{a}}^{m+n_\epsilon}_1$, we always have non-negativity of all components. Hence, we construct a new sequence $\{\hat{\mathbf{a}}^n_{(0,t]}\}$ where $\hat{\mathbf{a}}^n_1 = \tilde{\mathbf{a}}^{n+n_\epsilon}_1$ and $\hat{\mathbf{a}}^n_{(1,t]} = \mathbf{a}^{n+n_\epsilon}_{(1,t]}$.

This new sequence $\hat{\mathbf{a}}^n_{(0,t]}$ is the sequence we are after because using the induction hypothesis together with Observations D.2 and D.3, we have that $G_t(\hat{\mathbf{a}}^n) \to G_t(\mathbf{a})$, and G_t is continuous at $\hat{\mathbf{a}}^n$ for all n. Furthermore, by construction

$$G_t(\hat{\mathbf{a}}^n) = \hat{\mathbf{a}}^n_1 + [F_{t-1}(\hat{\mathbf{a}}^n_{(1,t]})]^+ = \tilde{\mathbf{a}}^{n+n_\epsilon}_1 + [F_{t-1}(\mathbf{a}^{n+n_\epsilon}_{(1,t]})]^+ = w^{n+n_\epsilon}. \tag{D.7}$$

Hence, we have shown that there exists a sequence $\hat{\mathbf{a}}^n_{(0,t]}$ satisfying (D.2) and (D.3). In addition, the continuity of $H \circ G_t$ at $\hat{\mathbf{a}}^n_{(0,t]}$ for all n is a direct consequence of continuity of G_t at $\hat{\mathbf{a}}^n_{(0,t]}$ and continuity of H at w^{n+n_ϵ}, which is equal to $G_t(\hat{\mathbf{a}}^n)$, for all n.

Case 2: $\mathbf{a}_1 \geq \mathbf{0}$. Without loss of generality by permuting the user labels, we can assume that the first k components of $\mathbf{a}_1$ are 0 while the rest of the $K-k$ components are positive. Now the sequence $\mathbf{a}^m_1$ with $1/m$ in the first k components and the non-zero values of $\mathbf{a}_1$ in the remaining coefficients converges to $\mathbf{a}_1$ such that for every m every component of $\mathbf{a}^m_1$ is positive. We construct a sequence $\{\mathbf{a}^m_{(0,t]}\}$ with this $\mathbf{a}^m_1$ and $\mathbf{a}^m_{(1,t]} = \mathbf{a}_{(1,t]}$. For ease of exposition we denote the vector with $1/m$ in the first k positions and 0s in the remaining $K-k$ positions by $[1/m]_k$. It is obvious that $G_t(\mathbf{a}^m_{(0,t]}) \to G_t(\mathbf{a}_{(0,t]})$ since

$$G_t(\mathbf{a}^m_{(0,t]}) = \mathbf{a}^m_1 + [F_{t-1}(\mathbf{a}^m_{(1,t]})]^+ = \mathbf{a}^m_1 + [F_{t-1}(\mathbf{a}_{(1,t]})]^+ = [1/m]_k + G_t(\mathbf{a}_{(0,t]}).$$

When $G_t(\mathbf{a}_{(0,t]}) \notin [0,C)^K$, for m large enough,[1] we have

$$H \circ G_t(\mathbf{a}^m_{(0,t]}) = \mathbf{e}\left(G_t(\mathbf{a}^m_{(0,t]})\right) = \mathbf{e}\left(G_t(\mathbf{a}_{(0,t]})\right) = H \circ G_t(\mathbf{a}_{(0,t]}),$$

where the function $\mathbf{e}$ is defined in the definition of H in (5.2). On the other hand, if $G_t(\mathbf{a}_{(0,t]}) \in [0,C)^K$, then the continuity of $\mathrm{Proj}_{\mathcal{R}}(\cdot)$ yields $H \circ G_t(\mathbf{a}^m_{(0,t]}) \to H \circ G_t(\mathbf{a}_{(0,t]})$.

Since for each m we have that $\mathbf{a}^m_1$ has all elements strictly positive, we can use the similar construction as in Case 1 but with $\mathbf{a}^m_{(0,t]}$ in place of $\mathbf{a}_{(0,t]}$. In particular, for each m, we can now generate a sequence $\{\tilde{\mathbf{a}}^{m,n}_{(0,t]}\}$ such that $\tilde{\mathbf{a}}^{m,n}_1 \to \mathbf{a}^m_1$ as $n \to +\infty$, $\tilde{\mathbf{a}}^{m,n}_{(1,t]} = \mathbf{a}^n_{(1,t]}$, and by using Observations D.2 and D.3, the following hold

$$\begin{aligned} G_t(\tilde{\mathbf{a}}^{m,n}_{(0,t]}) &\to G_t(\mathbf{a}^m_{(0,t]}), \qquad &(\text{D.8})\\ H \circ G_t(\tilde{\mathbf{a}}^{m,n}_{(0,t]}) &\to H \circ G_t(\mathbf{a}^m_{(0,t]}), \qquad &(\text{D.9}) \end{aligned}$$

with both $G_t(\cdot)$ and $H \circ G_t(\cdot)$ being continuous at $\tilde{\mathbf{a}}^{m,n}_{(0,t]}$ for all n.

Now we define the sequence $\hat{\mathbf{a}}^m_{(0,t]} = \tilde{\mathbf{a}}^{m,m}_{(0,t]}$ as the sequence we are after. By construction, we have $\hat{\mathbf{a}}^m_{(0,t]} \to \mathbf{a}_{(0,t]}$ and both $G_t(\cdot)$ and $H \circ G_t(\cdot)$ continuous at all $\hat{\mathbf{a}}^m_{(0,t]}$. Since $G_t(\mathbf{a}^m_{(0,t]}) \to G_t(\mathbf{a}_{(0,t]})$ and $H \circ G_t(\mathbf{a}^m_{(0,t]}) \to H \circ G_t(\mathbf{a}_{(0,t]})$, it follows from (D.8) and (D.9) that $G_t(\hat{\mathbf{a}}^m_{(0,t]}) \to G_t(\mathbf{a}_{(0,t]})$ and $H \circ G_t(\hat{\mathbf{a}}^m_{(0,t]}) \to H \circ G_t(\mathbf{a}_{(0,t]})$. □

[1] E.g., m being greater than the reciprocal of the maximum positive component of $G_t(\mathbf{a}_{(0,t]})$.

D.2 Proof of Lemma 5.13

Lemma 5.13. *If $K\mu < c$, the mapping G is quasi-continuous on $\mathcal{D}_{\mu}^{K}$ with respect to the scaled uniform topology.*

Proof. The proof follows the concept in [82]. Let $K\mu < c$ and $A \in \mathcal{D}_{\mu}^{K}$. Consider any sequence $\{A^n\}$ such that $A^n \to A$. The main step of the proof is based on the following claim:

Claim D.4. *There exists a $s^* = s^*(A) < \infty$ and n'_0 such that, when $n > n'_0$, the workloads at time $-s^*$ of the arrival sample paths A^n and A stay within the rate region $\mathcal{R}$, i.e., $Q_{s^*}(A^n) \in \mathcal{R}$ and $Q_{s^*}(A) \in \mathcal{R}$.*

With this claim and by the definition of G_{s^*}, the workloads at time zero for A^n and A are $G(A^n) = G_{s^*}(A^n_{(0,s^*]})$ and $G(A) = G_{s^*}(A_{(0,s^*]})$, respectively, when $n > n'_0$. In other words, we have transformed the infinite-horizon workload into the finite-horizon workload whose mapping is already known to be quasi-continuous by Lemma 5.12. The proof is now complete since G_{s^*} is quasi-continuous on $\mathbb{R}_{+}^{K\times s^*}$ and $A^n_{(0,s^*]} \to A_{(0,s^*]}$.

What is left is to show Claim D.4. To do this, we map the multi-queue problem into a single-queue problem with sum arrival processes, sum workload processes, and server capacity c (following the standard approach used, e.g., in [17]). We then follow the proof in [29,82] for the (aggregate) single-queue scenario. Given the definition of H and the simplex capacity region $\mathcal{R}$, the queue dynamics for the sum workload is that of a single queue whose arrivals are the sum of the arrivals, i.e.,

$$\hat{Q}_{t-1} = [\hat{Q}_t - c]^+ + \hat{A}_t, \tag{D.10}$$

where we define the hat $(\hat{\cdot})$ notation to mean the sum over all users, i.e. $\hat{A}_t = \sum_{k=1}^K A_t^k$ and $\hat{Q}_t = \sum_{k=1}^K Q_t^k$. Recursion of the queue dynamics (D.10) and letting $T \to \infty$ where $Q_T \in \mathcal{R}$ gives the standard expression for the stationary sum workload [29]:

$$\hat{Q}_0(A) = \sup_{t \in \mathbb{N}} \hat{A}(0,t] - c(t-1). \tag{D.11}$$

To prove the claim we use the fact that the rate region $\mathcal{R}$ is simplex, hence $\hat{Q}_s \leq c \Leftrightarrow Q_s \in \mathcal{R}$. That is, it suffices to show that there are a n_0' and a finite s such that, for $n \geq n_0'$, $\hat{Q}_s(A) \leq c$ and $\hat{Q}_s(A^n) \leq c$.

Since $A^n \to A$ under the scaled uniform topology, for any given $\epsilon > 0$, there exists a n_0 such that for $n \geq n_0$, $\max_{k \in K} \sup_{t \in \mathbb{N}} |\frac{A^{n,k}(0,t]}{t} - \frac{A^k(0,t]}{t}| < \epsilon$. Hence, $\sup_t |\frac{\hat{A}^n(0,t]}{t} - \frac{\hat{A}(0,t]}{t}| < K\epsilon$. Since $A \in \mathcal{D}_\mu^K$, there is a $t_0 < \infty$ such that for $t > t_0$ and $k \in K$, $\frac{A^k(0,t]}{t} \leq \mu + \epsilon$. Therefore, it follows that $\frac{\hat{A}(0,t]}{t} \leq K\mu + K\epsilon$ for $t > t_0$. Since $K\mu < c$, we choose $\epsilon = (c - K\mu)/4K$. We now have that for all $n \geq n_0$ and $t \geq t_0$, $\frac{\hat{A}^n(0,t]}{t} < K(\mu + 2\epsilon) = (c + K\mu)/2 < c$, and we also have that $\frac{\hat{A}(0,t]}{t} \leq K(\mu + \epsilon) = (c + 3K\mu)/4 < c$. In other words, for all $n \geq n_0$, the workload at time zero is a function of only the arrivals within time $(0, t_0]$ and hence,

$$\hat{Q}_0(A) = \sup_{1 \leq t \leq t_0} \hat{A}(0,t] - c(t-1) \quad \text{and} \quad \hat{Q}_0(A^n) = \sup_{1 \leq t \leq t_0} \hat{A}^n(0,t] - c(t-1).$$

Let $s \leq t_0$ and $s^n \leq t_0$ be the minimum values of the optimizing t's in the above equations, respectively. It can be shown as in [29, Lemma 5.4] that $\hat{Q}_s(A) \leq c$ and

$\hat{Q}_{s^n}(A^n) \leq c$ (and in addition, $\hat{Q}_v(A) > c$ and $\hat{Q}_{v^n}(A^n) > c$ for all $v \in (0, s)$ and $v^n \in (0, s^n)$).

Next we show that there exists n_1 such that for $n \geq n_1$, $s^n = s$. This is not difficult because it is known that $\hat{Q}_0$ is continuous on $\mathcal{D}_{K\mu}$ [82, Lemma 13]. Since $\hat{A}^n \to \hat{A}$ on $\mathcal{D}_{K\mu}$, we have $\hat{Q}_0(A^n) \to \hat{Q}_0(A)$ and $s^n \to s$. Since $s^n, s \in \mathbb{N}$, there exists a n_1 such that $s^n = s$ for $n \geq n_1$. The claim is now proved by taking $n_0' = \max(n_1, n_0)$. □

Bibliography

[1] E. Altman, B. Gaujal, and A. Hordijk, *Discrete-Event Control of Stochastic Networks: Multimodularity and Regularity.* Germany: Springer-Verlag, 2003.

[2] M. Andrews, A. Stolyar, K. Kumaran, R. Vijayakumar, K. Ramanan, and P. Whiting, "Scheduling in a queuing system with asynchronously varying service rates," *Probab. Eng. Inf. Sci.*, vol. 18, pp. 191–217, 2004.

[3] K. Azarian and H. El Gamal, "Cooperation in outage-limited multiple-access channels," in *2006 International Zurich Seminar on Communications*, Feb. 2006, pp. 150–153.

[4] K. Azarian, H. El Gamal, and P. Schniter, "On the achievable diversity-multiplexing tradeoff in half-duplex cooperative channels," *IEEE Trans. Inf. Theory*, vol. 51, no. 12, pp. 4152–4172, 2005.

[5] R. Berry, "Optimal power-delay trade-offs in fading channels: small delay asymptotics," in *Information Theory and Applications - Inaugural workshop*, San Diego, CA, Feb. 2006.

[6] R. Berry and R. Gallager, "Communication over fading channels with delay constraints," *IEEE Trans. Inf. Theory*, vol. 48, no. 5, pp. 1135–1149, 2002.

[7] R. Berry and E. Yeh, "Cross-layer wireless resource allocation," *IEEE Signal Processing Magazine*, vol. 21, no. 5, pp. 59–68, Sept. 2004.

[8] D. P. Bertsekas and R. Gallager, *Data Networks.* New Jersey: Prentice-Hall, Inc., 1992.

[9] D. Bertsimas, I. Paschalidis, and J. Tsitsiklis, "Asymptotic buffer overflow probabilities in multiclass multiplexers: an optimal control approach," *IEEE Trans. Autom. Control*, vol. 43, no. 3, pp. 315–335, 1998.

[10] I. Bettesh and S. Shamai, "Optimal power and rate control for minimal average delay: The single-user case," *IEEE Trans. Inf. Theory*, vol. 52, no. 9, pp. 4115–4141, 2006.

[11] D. D. Botvich and N. G. Duffield, “Large deviations, the shape of the loss curve, and economies of scale in large multiplexers,” *Queueing System*, vol. 20, pp. 293–320, 1995.

[12] E. Buffet and N. G. Duffield, “Exponential upper bounds via martingales for multiplexers with markovian arrivals,” *J. Appl. Prob.*, vol. 31, pp. 1049–1060, 1994.

[13] G. Caire, P. Elia, and K. R. Kumar, “Space-time coding: an overview,” *Journal of Communications Software and Systems*, Oct. 2006.

[14] G. L. Choudhury, D. M. Lucantoni, and W. Whitt, “Squeezing the most out of ATM,” *IEEE Trans. Comm.*, vol. 44, pp. 203–217, Feb 1996.

[15] C. Courcoubetis and R. Weber, “Buffer overflow asymptotics for a buffer handling many traffic sources,” *Journal of Applied Probability*, vol. 33, pp. 886–903, 1996.

[16] A. Czylwik, “Adaptive OFDM for wideband radio channels,” in *Proc. GLOBECOM ’96*, vol. 1, 1996, pp. 713–718.

[17] G. de Veciana and G. Kesidis, “Bandwidth allocation for multiple qualities of service using generalized processor sharing,” *IEEE Trans. Inf. Theory*, vol. 42, no. 1, pp. 268–272, Jan. 1996.

[18] A. Dembo and O. Zeitouni, *Large Deviations techniques and applications*, 2nd ed. Springer, 1998.

[19] N. G. Duffield, “Exponential bounds for queues with markovian arrivals,” *Queueing Systems*, vol. 17, pp. 413–430, 1994.

[20] N. Ehsan and T. Javidi, “Delay optimal transmission policy in a wireless multiaccess channel,” *IEEE Trans. Inf. Theory*, vol. 54, no. 8, pp. 3745–3751, Aug. 2008.

[21] N. Ehsan and M. Liu, “Optimal bandwidth allocation in a delay channel,” *IEEE J. Sel. Areas Commun.*, vol. 24, no. 8, pp. 1614–1626, Aug. 2006.

[22] H. El Gamal, G. Caire, and M. O. Damen, “The MIMO ARQ channel: diversity-multiplexing-delay tradeoff,” *IEEE Trans. Inf. Theory*, vol. 52, no. 8, pp. 3601–3621, Aug. 2006.

[23] H. El Gamal, G. Caire, and M. Damen, “Lattice coding and decoding achieve the optimal diversity-multiplexing tradeoff of MIMO channels,” *IEEE Trans. Inf. Theory*, vol. 50, no. 6, pp. 968–985, 2004.

[24] P. Elia, “Asymptotic universal optimality in wireless multi-antenna communications and wireless networks,” Ph.D. dissertation, USC, 2006.

[25] P. Elia, S. Kittipiyakul, and T. Javidi, "Cooperative diversity in wireless networks with stochastic and bursty traffic," in *IEEE Int. Symp. Information Theory*, Nice, France, June 2007.

[26] ——, "On the Responsiveness-Diversity-Multiplexing tradeoff," in *5th Intl. Symp. on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks*, Limassol, Cyprus, Apr. 2007.

[27] P. Elia, K. Kumar, S. Pawar, P. Kumar, and H.-F. Lu, "Explicit, minimum-delay space-time codes achieving the diversity-multiplexing gain tradeoff," *IEEE Trans. Inf. Theory*, vol. 52, no. 9, pp. 3869–3884, 2006.

[28] P. Elia, B. Sethuraman, and P. Vijay Kumar, "Perfect spacetime codes for any number of antennas," *IEEE Trans. Inf. Theory*, vol. 53, no. 11, pp. 3853–3868, 2007.

[29] A. Ganesh, N. O'Connell, and D. Wischik, *Big Queues.* Springer-Verlag, 2004.

[30] A. Ganti, "Transmission scheduling for multi-beam satellite systems," Ph.D. dissertation, Dept. of EECS, MIT, Cambridge, MA, 2003.

[31] A. Ganti, E. Modiano, and J. N. Tsitsiklis, "Optimal transmission scheduling in symmetric communication models with intermittent connectivity," *IEEE Trans. Inf. Theory*, vol. 53, no. 3, pp. 998–1008, Mar. 2007.

[32] J. Garcia, "An extension of the Contraction Principle," *Journal of Theoretical Probability*, vol. 17, no. 2, pp. 403–434, Apr. 2004.

[33] M. Goyal, A. Kumar, and V. Sharma, "Optimal cross-layer scheduling of transmissions over a fading multiaccess channel," *IEEE Trans. Inf. Theory*, vol. 54, no. 8, pp. 3518–3537, Aug. 2008.

[34] B. Hajek, "Optimal control of two interacting service stations," *IEEE Trans. Autom. Control*, pp. 491–499, 1984.

[35] N. Harvey, R. Ladner, L. Lovasz, and T. Tamir, "Semi-matchings for bipartite graphs and load balancing," in *Proc. of the Workshop on Algorithms and Data Structures (WADS '03)*, Ottawa, Canada, July 2003.

[36] T. Holliday and A. Goldsmith, "Joint source and channel coding for MIMO systems," in *Allerton Conf. on Comm., Control, and Computing*, 2004.

[37] ——, "Optimizing end-to-end distortion in MIMO systems," in *IEEE ISIT'05*, 2005.

[38] T. Holliday, A. Goldsmith, and H. Poor, "The impact of delay on the diversity, multiplexing, and ARQ tradeoff," in *IEEE International Conference on Communications (ICC '06)*, vol. 4, June 2006, pp. 1445–1449.

[39] T. Javidi, “Rate stable resource allocation in OFDM systems: from waterfilling to queue-balancing,” in *Allerton Conference on Communication, Control, and Computing*, Sept. 2004.

[40] T. Javidi, N. Song, and D. Teneketzis, “Expected makespan minimization on identical machines in two interconnected queues,” *Probab. Eng. Info. Sci.*, vol. 15, no. 4, pp. 409–443, 2001.

[41] F. P. Kelly, *Stochastic Networks: Theory and Applications.* Oxford University Press, 1996, ch. Notes on effective bandwidths, pp. 141–168.

[42] S. Kittipiyakul, P. Elia, and T. Javidi, “High-SNR analysis of outage-limited communications with bursty and delay-limited information,” *submitted to IEEE Trans. Inf. Theory*, 2007.

[43] S. Kittipiyakul and T. Javidi, “Subcarrier allocation in OFDMA systems: beyond water-filling,” in *2004 Asilomar Conference on Signals, Systems, and Computers*, Nov. 2004.

[44] ——, “Optimal operating point in MIMO channel for delay-sensitive and bursty traffic,” in *IEEE Int. Symp. Information Theory*, Seattle, Washington, USA, July 2006.

[45] ——, “Optimal operating point for MIMO multiple access channel with bursty traffic,” *IEEE Trans. Wireless Commun.*, vol. 6, no. 12, pp. 4464–4474, Dec. 2007.

[46] ——, “Relay scheduling and cooperative diversity for delay-sensitive and bursty traffic,” in *45th Annual Allerton Conference on Communication, Control, and Computing*, Monticello, Illinois, USA, Sept. 2007.

[47] E. Knightly and N. Shroff, “Admission control for statistical QoS: theory and practice,” *IEEE Netw.*, vol. 13, no. 2, pp. 20–29, Mar/Apr 1999.

[48] R. Knopp and P. A. Humblet, “Information capacity and power control in single-cell multiuser communications,” in *Proc. IEEE ICC 95*, Seattle, WA, June 1995.

[49] G. Koole, “Convexity in tandem queues,” *Prob. in Engr. and Info. Sciences*, vol. 18, no. 1, pp. 13–31, Jan. 2004.

[50] ——, “Monotonicity in Markov reward and decision chains: Theory and applications,” *Foundations and Trends in Stochastic Systems*, vol. 1, no. 1, 2006.

[51] R. Kumar and P. Varaiya, *Stochastic Control.* Prentice-Hall, 1986.

[52] J. Laneman, D. Tse, and G. Wornell, “Cooperative diversity in wireless networks: Efficient protocols and outage behavior,” *IEEE Trans. Inf. Theory*, vol. 50, no. 12, pp. 3062–3080, 2004.

[53] G. Li and H. Liu, "Dynamic resource allocation with finite buffer constraint in broadband OFDMA networks," *IEEE Trans. Wireless Commun.*, vol. 2, pp. 1037–1042, Mar. 2003.

[54] X. Lin, N. Shroff, and R. Srikant, "A tutorial on cross-layer optimization in wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 24, no. 8, pp. 1452–1463, 2006.

[55] L. Liu, P. Parag, J. Tang, W.-Y. Chen, and J.-F. Chamberland, "Resource allocation and quality of service evaluation for wireless communication systems using fluid models," *IEEE Trans. Inf. Theory*, vol. 53, no. 5, pp. 1767–1777, 2007.

[56] C. Lott and D. Teneketzis, "On the optimality of an index rule in multichannel allocation for single-hop mobile networks with multiple service classes," *Probab. Eng. Info. Sci.*, vol. 14, no. 3, pp. 259–297, July 2000.

[57] U. Manber, *Introduction to Algorithm: a creative approach.* Addison-Wesley Publishing Company, 1989.

[58] K. Mehlhorn and S. Naher, *The LEDA Platform of Combinatorial and Geometric Computing.* Cambridge University Press, 1999.

[59] J. Munkres, *Topology*, 2nd ed. Prentice Hall, 2000.

[60] S. Musy, "A delay optimal policy for symmetric broadcast channels," in *Inter. Conf. on Commun. Technology, 2006 (ICCT '06)*, Nov. 2006.

[61] R. Negi and S. Goel, "An information-theoretic approach to queuing in wireless channels with large delay bounds," in *IEEE Global Telecommunications Conference (GLOBECOM '04)*, vol. 1, 2004, pp. 116–122.

[62] L. Ozarow, S. Shamai, and A. Wyner, "Information theoretic considerations for cellular mobile radio," *IEEE Trans. Veh. Technol.*, vol. 43, no. 2, pp. 359–378, 1994.

[63] D. Rajan, A. Sabharwal, and B. Aazhang, "Delay-bounded packet scheduling of bursty traffic over wireless channels," *IEEE Trans. Inf. Theory*, vol. 50, no. 1, pp. 125–144, 2004.

[64] R. T. Rockafellar, *Convex Analysis.* Princeton Mathematical Series, No. 28, Princeton University, 1970.

[65] A. Sendonaris, E. Erkip, and B. Aazhang, "Increasing uplink capacity via user cooperation diversity," in *IEEE international symposium on information theory (ISIT'08)*, Aug. 1998.

[66] ——, “User cooperation diversity. part i. system description,” *IEEE Trans. Commun.*, vol. 51, no. 11, pp. 1927–1938, 2003.

[67] S. Shakkottai, “Effective Capacity and QoS for wireless scheduling,” *IEEE Trans. Autom. Control*, vol. 53, no. 3, pp. 749–761, Apr. 2008.

[68] S. Shakkottai and R. Srikant, “Many-sources delay asymptotics with applications to priority queues,” *Queueing Systems Theory and Applications (QUESTA)*, vol. 39, pp. 183–200, Oct. 2001.

[69] S. Shakkottai, R. Srikant, and A. L. Stolyar, “Pathwise optimality of the exponential scheduling rule for wireless channels,” *Adv. in Appl. Probab.*, vol. 36, no. 4, pp. 1021–1045, 2004.

[70] A. Simonian and J. Guibert, “Large deviations approximation for fluid queues fed by a large number of on/off sources,” *IEEE J. Sel. Areas Commun.*, vol. 13, no. 6, pp. 1017–1027, Aug. 1995.

[71] G. Song, Y. Li, J. Cimini, L.J., and H. Zheng, “Joint channel-aware and queue-aware data scheduling in multiple shared wireless channels,” in *2004 IEEE Wireless Communications and Networking Conference (WCNC)*, vol. 3, 2004, pp. 1939–1944.

[72] A. L. Stolyar and K. Ramanan, “Largest weighted delay first scheduling: Large deviations and optimality,” *Annals of Applied Probabilities*, vol. 11, no. 1, pp. 1–48, Feb. 2001.

[73] A. L. Stolyar, “Large deviations of queues sharing a randomly time-varying server,” *Queueing Systems*, vol. 59, no. 1, pp. 1–35, May 2008.

[74] V. G. Subramanian, “Large deviations of max-weight scheduling policies of convex rate regions,” in *2008 ITA*, 2008.

[75] ——, “Large deviations of Max-Weight scheduling policies on convex rate regions,” *submitted for publication*, 2008.

[76] L. Tassiulas and A. Ephremides, “Dynamic server allocation to parallel queues with randomly varying connectivity,” *IEEE Trans. Inf. Theory*, vol. 39, no. 2, pp. 466–478, Mar. 1993.

[77] S. Tavildar and P. Viswanath, “Approximately universal codes over slow-fading channels,” *IEEE Trans. Inf. Theory*, vol. 52, no. 7, pp. 3233–3258, 2006.

[78] D. Tse, P. Viswanath, and L. Zheng, “Diversity-multiplexing tradeoff in multiple-access channels,” *IEEE Trans. Inf. Theory*, vol. 50, no. 9, pp. 1859–1874, Sept. 2004.

[79] Y. Viniotis, *Probability and Random Processes for Electrical Engineers.* McGraw-Hill, 1998.

[80] J. Walrand and P. Varaiya, *High-Performance Communication Networks*, 2nd ed. Morgan Kaufmann Publishers, 2000.

[81] A. Weiss, "A new technique for analyzing large traffic systems," *Advances in Applied Probability*, vol. 18, pp. 506–532, 1986.

[82] D. J. Wischik, "Sample path large deviations for queues with many inputs," *Ann. Appl. Probab.*, vol. 11, no. 2, pp. 379–404, 2001.

[83] C. Y. Wong, R. Cheng, K. Lataief, and R. Murch, "Multiuser OFDM with adaptive subcarrier, bit, and power allocation," *IEEE J. Sel. Areas Commun.*, vol. 17, no. 10, pp. 1747–1758, Oct. 1999.

[84] D. Wu and R. Negi, "Effective capacity: a wireless link model for support of quality of service," *IEEE Trans. Wireless Commun.*, vol. 2, no. 4, pp. 630–643, July 2003.

[85] C.-W. Yang, A. Wierman, S. Shakkottai, and M. Harchol-Balter, "Tail asymptotics for policies favoring short jobs in a many-flows regime," *SIGMETRICS Perform. Eval. Rev.*, vol. 34, no. 1, pp. 97–108, 2006.

[86] S. Yang and J.-C. Belfiore, "Optimal SpaceTime codes for the MIMO amplify-and-forward cooperative channel," *IEEE Trans. Inf. Theory*, vol. 53, no. 2, pp. 647–663, 2007.

[87] E. M. Yeh and A. S. Cohen, "Delay optimal rate allocation in multiaccess fading communications," in *Proc. Allerton Conf. on Communication, Control, and Computing*, Monticello, IL, 2004.

[88] L. Ying, R. Srikant, A. Eryilmaz, and G. Dullerud, "A Large Deviations analysis of scheduling in wireless networks," *IEEE Trans. Inf. Theory*, vol. 52, no. 11, pp. 5088–5098, Nov. 2006.

[89] T. Zajic and A. Dembo, "Large deviations: From empirical mean and measure to partial sums process," *Stochastic Processes and their Applications*, vol. 57, no. 2, pp. 191–224, June 1995.

[90] L. Zheng and D. Tse, "Diversity-Multiplexing: a fundamental tradeoff in multiple-antenna channels," *IEEE Trans. Inf. Theory*, vol. 49, no. 5, pp. 1073–1096, May 2003.