June 2017

Delay QoS Provisioning and Optimal Resource Allocation for Wireless Networks

Yi Li
Syracuse University

Abstract

Recent years have witnessed a significant growth in wireless communication and networking due to the exponential growth in mobile applications and smart devices, fueling unprecedented increase in both mobile data traffic and energy demand. Among such data traffic, real-time data transmissions in wireless systems require certain quality of service (QoS) constraints e.g., in terms of delay, buffer overflow or packet drop/loss probabilities, so that acceptable performance levels can be guaranteed for the end-users, especially in delay sensitive scenarios, such as live video transmission, interactive video (e.g., teleconferencing), and mobile online gaming. With this motivation, statistical queuing constraints are considered in this thesis, imposed as limitations on the decay rate of buffer overflow probabilities. In particular, the throughput and energy efficiency of different types of wireless network models are analyzed under QoS constraints, and optimal resource allocation algorithms are proposed to maximize the throughput or minimize the delay.

In the first part of the thesis, the throughput and energy efficiency analysis for hybrid automatic repeat request (HARQ) protocols are conducted under QoS constraints. Approximations are employed for small QoS exponent values in order to obtain closed-form expressions for the throughput and energy efficiency metrics. Also, the impact of random arrivals, deadline constraints, outage probability and QoS constraints are studied. For the same system setting, the throughput of HARQ system is also analyzed using a recurrence approach, which provides more accurate results for any value of the QoS exponent. Similarly, random arrival models and deadline constraints are considered, and these results are further extended to the finite-blocklength coding regime.

Next, cooperative relay networks are considered under QoS constraints. Specifically, the throughput performance in the two-hop relay channel, two-way relay channel, and multi-source multi-destination relay networks is analyzed. Finite-blocklength codes are considered for the two-hop relay channel, and optimization over the error probabilities is investigated. For the multi-source multi-destination relay network model, the throughput for both cases of with and without CSI at the transmitter sides is studied. When there is perfect CSI at the transmitter, transmission rates can be varied according to instantaneous channel conditions. When CSI is not available at the transmitter side, transmissions are performed at fixed rates, and decoding failures lead to retransmission requests via an ARQ protocol.

Following the analysis of cooperative networks, the performance of both half-duplex and full-duplex operations is studied for the two-way multiple input multiple output (MIMO) system under QoS constraints. In full-duplex mode, the self-interference inflicted on the reception of a user due to simultaneous transmissions from the same user is taken into account. In this setting, the system throughput is formulated by considering the sum of the effective capacities of the users in both half-duplex and full-duplex modes. The low signal to noise ratio (SNR) regime is considered and the optimal transmission/power-allocation strategies are characterized by identifying the optimal input covariance matrices.

Next, mode selection and resource allocation for device-to-device (D2D) cellular networks are studied. As the starting point, transmission mode selection and resource allocation are analyzed for a time-division multiplexed (TDM) cellular network with one cellular user, one base station, and a pair of D2D users under rate and QoS constraints. For a more complicated setting with multiple cellular and D2D users, two joint mode selection and resource allocation algorithms are proposed. In the first algorithm, the channel allocation problem is formulated as a maximum-weight matching problem, which can be solved by employing the Hungarian algorithm. In

the second algorithm, the problem is divided into three subproblems, namely user partition, power allocation and channel assignment, and a novel three-step method is proposed by combining the algorithms designed for the three subproblems.

In the final part of the thesis, resource allocation algorithms are investigated for content delivery over wireless networks. Three different systems are considered. Initially, a caching algorithm is designed, which minimizes the average delay of a single-cell network. The proposed algorithm is applicable in settings with very general popularity models, with no assumptions on how file popularity varies among different users, and this algorithm is further extended to a more general setting, in which the system parameters and the distributions of channel fading change over time. Next, for D2D cellular networks operating under deadline constraints, a scheduling algorithm is designed, which manages mode selection, channel allocation and power maximization with acceptable complexity. This proposed scheduling algorithm is designed based on the convex delay cost method for a D2D cellular network with deadline constraints in an OFDMA setting. Power optimization algorithms are proposed for all possible modes, based on our utility definition. Finally, a two-step intercell interference (ICI)-aware scheduling algorithm is proposed for cloud radio access networks (C-RANs), which performs user grouping and resource allocation with the goal of minimizing delay violation probability. A novel user grouping algorithm is developed for the user grouping step, which controls the interference among the users in the same group, and the channel assignment problem is formulated as a maximum-weight matching problem in the second step, which can be solved using standard algorithms in graph theory.

DELAY QOS PROVISIONING AND OPTIMAL RESOURCE ALLOCATION FOR WIRELESS NETWORKS

By

Yi Li

B.E., University of Science and Technology of China, Hefei, China, 2011

DISSERTATION

Submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Electrical and Computer Engineering

Syracuse University

June 2017

Acknowledgements

First, I would like to thank my advisors, Prof. Mustafa Cenk Gursoy and Prof. Senem Velipasalar, for their help and guidance during my entire Ph.D. study. In this period, I have gained much experience in research, and completed many studies in the area of wireless communication.

I would like to thank all my Ph.D. defense committee members Prof. Lixin Shen, Prof. Biao Chen, Prof. Yingbin Liang, Prof. Makan Fardad, and Prof. Pramod Varshney for carefully reading my thesis.

Finally, I would like to thank all my labmates and friends for their support during these years at Syracuse. I had really wonderful time studying and working with them.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Quality of Service (QoS) Requirements

Recent years have witnessed significant growth in wireless communication and networking due to the exponential growth in mobile applications and smart devices, fueling unprecedented increase in both mobile data traffic and energy demand. Among such data traffic, real-time data transmissions in wireless systems require certain QoS constraints e.g., in terms of delay, buffer overflow or packet drop/loss probabilities, so that acceptable performance levels can be guaranteed for the end-users, especially in delay sensitive scenarios, such as live video transmission, interactive video (e.g., teleconferencing), and mobile online gaming. With this motivation, we consider statistical queuing constraints in this thesis, imposed as limitations on the decay rate of buffer overflow probabilities. In [1], effective bandwidth was introduced as a measure of the system throughput under such statistical queuing or QoS constraints. More specifically, effective bandwidth has been defined as the minimum constant transmission rate required to support time-varying arrivals while the buffer overflow probability decays exponentially with increasing overflow threshold. In [2], effective bandwidths of departure processes with time-varying service rates were investigated,

and the theory of effective bandwidth was employed to analyze the performance of high speed networks in [3]. Later, effective capacity was defined in [4] as a dual concept to characterize the maximum constant arrival rates that can be supported by time-varying wireless transmission rates again under statistical queuing constraints.

1.2 Literature Review

1.2.1 Throughput and Energy Efficiency of Hybrid Automatic Repeat Request (HARQ) Protocols under QoS Constraints

In wireless communications, higher throughput and better energy efficiency are two key considerations and have become critical performance metrics due to the exponential growth in mobile applications and smart devices, fueling unprecedented increase in both mobile data traffic and energy demand. More specifically, with this growth coupled with the availability of only limited battery power for mobile devices, rising energy costs and growing concerns on environmental impact, the analysis of the energy efficiency and green operation in wireless systems have become increasingly more important in recent years. At the same time, throughput and energy efficiency are not the only considerations. In a wireless propagation environment in which noise, fading, path loss, multipath propagation and Doppler frequency shift are being experienced, reliability is equally important with strong implications on energy efficiency.

Due to the challenges in wireless systems, many advanced techniques have been developed to address these concerns, and automatic repeat request (ARQ) and forward error correction (FEC) are two types of widely used schemes applied in order to ensure reliable delivery of data in such challenging wireless channel conditions. While ARQ facilitates the retransmission of erroneously received data packets with

feedback from the receiver to the transmitter, FEC schemes enable the correction of transmission errors without retransmission by adding redundancy to the data. In order to provide better error correction performance and lower implementation cost, ARQ and FEC schemes are combined to develop hybrid ARQ (HARQ) [5].

HARQ protocols have the ability to increase the probability of successful transmission and adapt the transmission rate to time-varying channel conditions with limited channel side information (CSI) at the transmitter [6]. In HARQ with chase combining (HARQ-CC) and HARQ with incremental redundancy (HARQ-IR) schemes, the corrupted packets are not deleted but rather stored and combined in the next transmission period. In particular, better adaptation to channel conditions and higher throughput can be achieved by employing HARQ-IR. A detailed study on the performance of HARQ-CC and HARQ-IR protocols was provided in [7], in which the throughput was characterized following an outage probability analysis. The throughput of HARQ protocols was studied from an information-theoretic perspective and it was shown that the throughput of HARQ-IR could approach the ergodic capacity for large transmission rates with only limited CSI. More recently, performance of HARQ in Rayleigh block fading channels was investigated via a mutual information-based analysis in [8], and long-term average rates achieved with HARQ were characterized under constraints on the outage probability and the maximum number of HARQ rounds. A similar throughput analysis of HARQ schemes subject to an outage constraint was also conducted in [9]. And in [10], the tradeoff between energy efficiency and transmission delay in wireless multiuser systems employing HARQ-IR was studied. The energy efficiency of HARQ protocols has been addressed recently. For instance, the energy efficiency of HARQ-CC and HARQ-IR schemes for delay insensitive systems was studied in [11], and the energy efficiency achievable by HARQ schemes with optimized code rate is studied in [12].

In the presence of QoS constraints, it is critical to evaluate the performance of

HARQ schemes since they involve retransmissions. With this motivation, the authors in [13] analyzed the impact of different power allocation schemes on energy per bit and effective transmission delay of HARQ-IR in a multiuser downlink channel. Moreover, the recent work in [14] mainly focused on the performance comparison between adaptive modulation and coding (AMC) and HARQ-IR in terms of energy efficiency under QoS constraints. More recently, the authors of [15] characterized the effective capacity of retransmission schemes through a recurrence relation approach.

1.2.2 Throughput of Cooperative Relay Networks under QoS Constraints

In recent years, the traffic load of wireless networks has grown significantly, primarily due to the exponential increase in multimedia traffic driven by applications on smart mobile devices. Many wireless communication schemes have been proposed to boost the transmission speed, reduce latency, enhance reliability and improve energy efficiency. One such strategy is cooperative relaying, which can greatly enhance the performance for long distance transmissions among users and improve resource efficiency.

Two-hop relay channel is a basic cooperative relay network structures. Several studies have applied the effective capacity analysis to two-hop wireless relay channels. In [16], the power allocation policies in relay networks under QoS constraints were analyzed. However, no buffer constraints were imposed at the relay node in this work. In [17], the effective capacity of two-hop relay channel was investigated under queueing constraints at both the source and relay nodes. In [16] and [17], it was assumed that the instantaneous transmission rates were given by the Shannon capacity achieved with channel codes with blocklengths growing without bound. In [18] and [19], the coding rate expressions in finite blocklength regimes were studied. This has led to interest in the analysis of performance attained with finite blocklength codes.

For instance, the effective capacity of single-hop channels was characterized in [20] under finite blocklength assumption. Recently, performance of relaying in the finite blocklength regime was studied in [21] and [22] without considering any queueing constraints.

Multiple-access relay channel is a more general model, in which multiple users transmit to a destination node with the help of a single relay node. The throughput of multiple-access relay networks have been analyzed by several studies. In [23], the achievable rates of Gaussian orthogonal multi-access relay channels were investigated, which were also proved to have a max-flow min-cut interpretation. The throughput region of the same system model was also given in [24] with superposition block Markov encoding and multiple access encoding. Further analysis was also provided in [25], in which the optimal resource allocation strategy was studied to achieve the maximum sum rate. In [26], the system throughput region of a generalized multiple access relay network, which includes multiple transmitters, multiple relays, and a single destination, was studied. In all cases, with the help of relay nodes, the channel conditions effectively improve for long distance wireless communication, and performance enhancements are realized.

A further generalization of multiple-access relay channels is to introduce multiple destination nodes. These models are referred to as multi-source multi-destination relay networks. Multi-source multi-destination relay network model can be seen as a combination of multiple-access, broadcast, and two-hop relay channels, and it can be used to address scenarios in which multiple pairs of users simultaneously communicate with the help of a relay node. A basic practical example of these models is cellular operation in which multiple mobile users within a cell communicate with each other through a base station, which essentially acts as a relay unit between the source and destination nodes[1]. Such networks have been analyzed in several recent studies. In

[1]Moreover, in LTE-Advanced cellular standards, relaying and coordinated multi point (CoMP) operation are introduced to provide enhanced coverage and capacity at cell edges, and multi-user

[27], the throughput of the amplify-and-forward multi-source multi-destination relay network was studied, when the relay was equipped with multiple antennas. Based on this work, the same authors studied the impact of imperfect CSI in [28], and proposed an antenna selection algorithm to improve the performance. In [29], the joint power optimization was investigated for the multi-source multi-destination relay network, and in [30], network coding was applied to this type of network, and the system performance was evaluated.

Recently, effective capacity analysis has been applied to multiuser and cooperative relay systems. For cooperative relay systems, the authors in [16] studied efficient resource allocation strategies over wireless relay channels under statistical QoS constraints by employing effective capacity as the throughput metric. However, in this work, either no buffer was needed at the relay if amplify-and-forward (AF) strategy was employed or no relay buffer constraints were imposed when decode-and-forward (DF) was used. In [31], queueing analysis was conducted for a butterfly network when the arrivals were modeled as a two-state Markov-modulated fluid process, and network coding or classical routing was performed by the intermediate relay node. In this study, all links were assumed to be time-invariant. Therefore, static rather than fading channels were considered. For multi-user systems, in [32], the effective capacity region of the multiple-access fading channel under queueing constraints was analyzed, and this result was extended to characterize the throughput region of the multiple-access channel with Markov arrivals in [33]. The effective capacity region of the fading broadcast channel and optimal power allocation policies were studied in [34]. These multi-user studies addressed single-hop channels and did not consider cooperative schemes.

relay models can be realized in these operation modes as well.

1.2.3 Throughput and Mode Selection in Two-way Multiple-Input Multiple-Output (MIMO) Systems under Queuing Constraints

Recently, full-duplex two-way communication has attracted much attention due to its potential to significantly improve the spectrum efficiency by allowing two users to communicate simultaneously over a single channel. A survey of the key concepts, challenges and opportunities in full-duplex wireless communications was provided in [35]. In particular, self-interference cancelation is a critical concern in full-duplex wireless systems, which has been addressed by several recent studies. For instance, in [36], multiple schemes in full-duplex MIMO relays have been studied for loopback self-interference cancelation, including natural isolation, time-domain cancelation, and spatial suppression. It is important to note that self-interference cancelation is generally imperfect and the residual interference is considered to be proportional to the transmitted signal of the same node. In addition to self-interference management, there are numerous works analyzing the performance in full-duplex two-way channels in various settings. Among different models, two-way MIMO systems have been highlighted as they can further boost the system throughput by employing multiple antennas for transmitting and receiving. In [37], the influence of spatial fading correlation was investigated in two-way MIMO systems, and strategies were proposed to reduce the sensitivity to spatial fading correlation. In [38], the impact of channel estimation error was studied for full-duplex two-way networks, and closed-form expressions were derived for the ergodic capacity. Also, Effective capacity has been studied extensively for various different wireless channel models in order to determine the system throughput under such statistical queuing constraints. For instance, in [39], the effective capacity of a one-directional point-to-point MIMO systems was investigated.

1.2.4 Mode Selection and Resource Allocation for Device-to-Device (D2D) Cellular Networks

The concept of D2D communication underlaid with cellular networks has attracted much interest recently. D2D communication enables users to communicate directly without going through the base station, and reuse the same spectral resources with cellular users. In [40], the advantages of D2D communications were studied, and it was shown that it could greatly enhance the spectral efficiency and lower the latency. A comprehensive overview was provided in [41], where different modeling assumptions and key considerations in D2D communications were detailed.

Mode selection and resource allocation are two key problems in D2D communication, which has attracted much interest. In mode selection, each D2D user has to decide whether to transmit directly using a dedicated spectrum or by sharing the spectrum with cellular users in the D2D mode, or transmit in the same way as cellular users via a D2D two-hop channel through the base station in the cellular mode (which is essentially a non-D2D mode). In resource allocation, the system has to assign a channel resource to each user, and users have to optimize their transmission power. The resource allocation problem in D2D cellular networks is rather complicated because D2D users can reuse (i.e., share) the same channel resources with cellular users and inflict interference to them. Due to this reusing mechanism, the number of possible solutions for channel assignment increases exponentially with the number of D2D users, and the power optimization problem becomes high-dimensional and non-convex. Therefore, the analysis becomes even more complicated when mode selection and resource allocation problems are considered jointly for improved performance.

In the literature, many studies have been conducted to address the mode selection and resource allocation problems for D2D cellular networks. In the literature, many studies have been conducted to address the mode selection and resource allocation problems for D2D cellular networks. For instance, the authors of [42] considered

the mode selection problem in a cell with one D2D pair and one cellular user. The channel allocation problem was solved using the Hungarian algorithm in [43], just for the uplink reuse mode. Later in [44], the resource allocation in the D2D cellular network was investigated, for the uplink and downlink reuse mode, and in [45], a resource allocation strategy was proposed, which also included the D2D dedicated mode. More recently, the joint mode selection and resource allocation in a general cellular network with multiple D2D pairs were addressed in [46]. In order to reduce the complexity in analysis, most of these studies were based on the instantaneous channel conditions. In such cases, the system may have to perform the mode selection and resource allocation very frequently, resulting in high computational load and significant cost.

In order to achieve improved results with lower time consumption, several algorithms were proposed via game-theoretic approaches. For example, the resource allocation problem was considered in [47] via the reverse iterative combinatorial auction game, and the authors of [48] solved a similar problem using the coalitional game theory. Besides the game-theoretic techniques, vertex coloring is another method that can efficiently divide D2D users into groups in which interference constraints are satisfied. In [49], vertex coloring algorithm was used to group D2D users with the goal of avoiding interference. A similar approach was used in [49] and [50] to maximize the instantaneous sum rate while satisfying the instantaneous signal to interference plus noise ratio (SINR) constraints, and a frequency band assignment process was also included after dividing D2D users into groups in [51].

1.2.5 Delay-Aware Scheduling Algorithms for Content Delivery over Wireless Networks

Scheduling is one of the key design considerations in cellular networks. Scheduling algorithms allocate limited channel resources to large amount of users, while also

guaranteeing the performance requirements of the system. For delay sensitive applications, packet delay is an important criterion of the performance. Numerous studies have been conducted to design and analyze delay optimal scheduling algorithms. A series of works considered this problem from the perspective of minimizing the queue length at transmitters. In [52], the MaxWeight rule was proposed, and the throughput optimality was shown for MaxWeight type algorithms. These type of algorithms stabilize the queueing system, and minimize the queue length. However, small queue length can not always guarantee good delay performance.

The second line of work considered minimizing the decay rate of the delay violation probability as the scale of the system (which is the numbers of the users and available channels) increases. In [53], the Delay Weighted Matching (DWM) algorithm was proposed, which can provide good delay performance. In order to reduce the complexity of the DWM algorithm, the authors of [54] proposed a hybrid algorithm that reduced the asymptotic complexity of the scheduling algorithm. Although this approach guarantees good delay performance, the asymptotic analysis of these algorithms are complicated, making them difficult to be extended for a more general system setting. In [53] and [54], only downlink transmission with binary rate was considered.

Yet another line of studies considered constructing convex delay costs. In [55] and [56], convex delay cost approach was proposed and developed to deal with systems operating under deadline constraints. This type of algorithm minimizes the delay violation probability under heavy traffic assumption.

1.2.6 Wireless D2D Caching Networks

Recently, many studies have been conducted to analyze caching strategies in wireless networks in order to satisfy the throughput, energy efficiency and latency requirements in next-generation 5G wireless systems. By storing parts of the popular files

at the base station and users' devices, network traffic load can be managed/balanced effectively, and traffic delay can be greatly reduced. It has been pointed out that 60% of the content is cacheable in the network traffic [57], which can be transmitted and stored close to the users before receiving the requests. A brief overview of wireless caching was provided in [58], which introduced the key notions, challenges, and research topics in this area. In order to improve the performance effectively, the system needs to estimate and track the popularity of those cacheable contents, and predict the popularity variations, helping to guarantee that the most popular contents are cached and the outdated contents are removed. In [59], popularity matrix estimation algorithms were studied for wireless networks with proactive caching.

Multiple caching strategies have been investigated in the literature, which improve the performance in different ways. When contents are cached at the base stations, the energy consumption, traffic load and delay of the backhaul can be reduced [60], and the base stations in different cells can cooperate to improve the spectral efficiency gain [61]. When contents are cached at the users' devices, the base station can combine different files together and multicast to multiple users, and the users can decode their desired files using their cached files. A content distribution algorithm for this approach was given in [62], and the analysis of the coded multicasting gain was provided in [63].

In the literature, several studies have been performed to combine content caching with D2D wireless networks. In such cases a user can receive from its neighbors if these have cached the requested content. An overview on wireless D2D caching networks was provided in [64], in which the key results for different D2D caching strategies were presented. To design caching policies for the wireless D2D network, the authors of [65] proposed a caching policy that maximizes the probability that requests can be served via D2D communications. For a similar system setting, a caching policy that maximizes the average number of active D2D links was obtained

in [66]. Most of these works were based on stochastic geometry models, in which nodes/users were distributed randomly. However, these types of models mainly focus on the path loss, and do not fully address the effects of channel fading. Without the characterization of the channel fading, an accurate analysis on the throughput and delay is not viable. Moreover, many works only tackle a simple case in which users have identical popularity vectors.

1.2.7 Intercell Interference (ICI) Control in Cloud Radio Access Network (C-RAN)

In order to satisfy the growing demands and provide the required QoS guarantees and high reliability in next-generation 5G wireless systems, several advanced techniques have been proposed, and C-RAN is one novel mobile network architecture that improves the performance of cellular networks. By centralizing the baseband processing resources of multiple cells in a virtualized baseband unit (BBU) pool, C-RAN can achieve cooperative processing among different cells and utilize the BBUs more efficiently [67] [68]. Remote radio heads (RRHs) and BBU are separated geographically and connected via optical fibers in the C-RAN architecture. BBU pool is shared between cells as a virtualized cluster which operates baseband processing. Compared with the conventional architectures in which BBUs of different cells are not shared, C-RAN can achieve information exchange and cooperative processing between cells more easily with low latency, and it has high adaptability to nonuniform traffic. A comprehensive survey on C-RAN and its implementation is provided in [69].

For most orthogonal frequency division multiple access (OFDMA)-based cellular networks, ICI is a significant interference source because of the frequency reuse among multiple neighbouring cells. Many advanced methods have been studied to control ICI. For instance, the soft frequency reuse (SFR) scheme is proposed in [70] and [71], in which cell edge users transmit with high power in non-overlapping cell

edge bands allocated to adjacent cells, and center users use the cell center bands with limited transmission power. The authors in [72] further compared the performance of SFR with partial frequency reuse scheme. In these conventional ICI control schemes, cooperation between neighbouring cells are not considered, which limits their performances. In C-RAN, cooperative processing among the cells sharing the same BBU pool becomes easier and more efficient, which helps to improve ICI control. In [73], a resource allocation and RRH association algorithm was proposed for ICI coordination in a long term evolution (LTE) heterogeneous network setting with C-RAN architecture. However, optimization over multiple cells greatly increases the complexity, which causes problems in delay sensitive applications. In addition, packet delay is an important performance criterion for delay sensitive applications such as live video streaming and online gaming. In most of the related studies considering ICI control, the objectives are interference minimization, SINR maximization and throughput maximization, and hence delay minimization is not addressed.

1.3 Outline and Main Contributions

In Chapter 2, we provide a detailed review on the formulation of our statistical queuing constraints, and describe the methods to characterize the throughput under queuing constraints for different channel and arrival models. More specifically, both single-hop and two-hop channels are considered with constant-rate arrivals at the source nodes, and random arrival models are considered for the single-hop channel.

In Chapter 3, we conduct the throughput and energy efficiency analysis for HARQ protocols under QoS constraints. Approximations are employed for small QoS exponent values in order to obtain closed-form expressions for the throughput and energy efficiency metrics.

- In Section 3.1, the throughput of HARQ-IR with fixed transmission rate is studied in the presence of queuing constraints imposed as limitations on buffer overflow probabilities. In particular, tools from the theory of renewal processes and stochastic network calculus are employed to characterize the maximum arrival rates that can be supported by the wireless channel when HARQ-IR is adopted. Effective capacity formulation is employed as the throughput metric and a closed-form expression for the effective capacity of HARQ-IR is determined for small values of the QoS exponent. The impact of the fixed transmission rate, queuing constraints, and hard deadline limitations on the throughput is investigated and comparisons with regular ARQ operation are provided.

- In Section 3.2, energy efficiency of HARQ schemes with statistical queuing constraints is studied for both constant-rate and random Markov arrivals by characterizing the minimum energy per bit and wideband slope. In particular, two queuing models are considered. Specifically, when outage occurs, the transmitter keeps the packet, lowers its priority, and attempts to retransmit it later in the first queue model while the packet is discarded and removed from the buffer in the second queue model. For both models, energy efficiency is investigated when outage constraints, statistical queuing constraints and deadline constraints are imposed. The deadline constraint provides a limitation on the number of retransmissions or equivalently the number of HARQ rounds. Under these assumptions, closed-form expressions are obtained for the energy efficiency metrics of HARQ-CC, and comparisons among different arrival models are made. For instance, it is shown that stricter queuing constraints and more bursty sources degrade the energy efficiency by lowering the wideband slope.

In Chapter 4, we conduct throughput analysis for HARQ protocols under QoS

constraints via the recurrence approach proposed in [15]. Also, deadline and outage constraints, random arrival models and finite blocklength codes are considered.

- In Section 4.1, we characterize the throughput of HARQ-CC for three typical Markov sources in the presence of statistical queuing constraints while satisfying a target outage probability. In most of the related works investigating the throughput of HARQ schemes under statistical queuing constraints, the occurrence of packet drops from the buffer due to deadline violation have generally not been explicitly addressed. From the perspective of the buffer, packets dropped/discarded from the buffer should contribute to the departure rate in the queuing analysis. However, when characterizing the throughput, the discarded packets should not be taken into account, since the receiver does not receive them. In this section, the impact of such packet drops is explicitly considered. Also, we identify the impact of source randomness on the throughput of HARQ-CC systems under statistical QoS constraints.

- In Section 4.2, we study the throughput of HARQ-IR with finite-blocklength codes, deadline limits, and statistical queuing constraints by employing the notions of effective capacity and effective bandwidth from stochastic network calculus. Two different arrival models, namely the constant-rate and ON-OFF discrete time Markov arrivals, are studied, and throughput characterizations are obtained for both arrival models. In prior works focusing on HARQ under queuing constraints, it was assumed that the instantaneous transmission rates were given by the Shannon capacity achieved with channel codes with blocklengths growing without bound. However, it is practically more appealing to address the performance with finite-blocklength codes and more explicitly take into account decoding error probabilities in the analysis of HARQ especially in the presence of deadline and queuing constraints. With this motivation, we leverage recent advances in the characterization of coding rates in the finite

blocklength regime [18], [19], and study the throughput of HARQ-IR schemes achieved with finite-blocklength codes subject to statistical queuing constraints at the transmitter buffer.

In Chapter 5, the throughput of three cooperative relay network models are studied under QoS constraints.

- In Section 5.1, we characterize the throughput of the two-hop relay channel in the finite blocklength regime when statistical queueing constraints are imposed at both the source and relay. We first identify the throughput by determining the effective capacity of the two-hop relay system in the finite blocklength regime, and then establish several key properties of the throughput in terms of the target error rates. Based on these properties, we develop an efficient search algorithm to solve the throughput maximization problem and obtain the corresponding optimal parameter values.

- In Section 5.2, we extend the effective capacity analysis of one-directional two-hop relay channel to a two-way relay channel setting. More specifically, we study the throughput of two-way relay channels in the presence of queueing constraints at both the source nodes and the relay node. Note that the two-way relay model has significant differences from that of the one-way relay considered in [17]. We consider half-duplex, decode-and-forward relaying. In this setting, our main goal is to identify the pair of maximum arrival rates at the sources while the statistical queuing constraints at the source nodes and relay are satisfied.

- In Section 5.3, the throughput of relay networks with multiple source-destination pairs under queuing constraints has been investigated for both variable-rate and fixed-rate schemes. When CSI is available at the transmitter side, transmitter-

s can adapt their transmission rates according to the channel conditions, and achieve the instantaneous channel capacities. In this case, the departure rates at each node have been characterized for different system parameters, which control the power allocation, time allocation and decoding order. In the other case of no CSI at the transmitters, a simple ARQ protocol with fixed rate transmission is used to provide reliable communication. Under this ARQ assumption, the instantaneous departure rates at each node can be modeled as an ON-OFF process, and the probabilities of ON and OFF states are identified. With the characterization of the arrival and departure rates at each buffer, stability conditions are identified and effective capacity analysis is conducted for both cases to determine the system throughput under statistical queuing constraints. In addition, for the variable-rate scheme, the concavity of the sum rate is shown for certain parameters, helping to improve the efficiency of parameter optimization.

In Chapter 6, we extend the analysis conducted in [39] to two-way MIMO systems. We consider both half-duplex and full-duplex operations. In half-duplex mode, we can have time-division multiplexing or frequency-division multiplexing, which are essentially equivalent in terms of their performances. In full-duplex mode, we take into account the self-interference inflicted on the reception of a user due to simultaneous transmissions from the same user. In this setting, we initially formulate the system throughput by considering the sum of the effective capacities of the users in both half-duplex and full-duplex modes. Subsequently, we consider the low signal to noise ratio (SNR) regime and characterize the optimal transmission/power-allocation strategies by identifying the optimal input covariance matrices. Finally, via numerical results, we address mode selection by determining which mode yields higher through-

put under different SNR levels and distances. For fair comparison, we assume that the number of transmitting and receiving antennas are the same at each node in both half-duplex and full-duplex modes.

In Chapter 7, we study the mode selection and resource allocation algorithms for D2D cellular networks.

- In Section 7.1, transmission mode selection and resource allocation in a time-division multiplexed (TDM) cellular network with one cellular user, one base station, and a pair of D2D users is investigated under rate and queuing constraints. Using tools from stochastic network calculus, the system throughput under statistical queuing constraints is formulated, efficient resource allocation algorithms for all possible modes are proposed, and the influence of the positions of each node and the queuing constraints is analyzed. Scenarios and conditions for different modes to be optimal in the sense of maximizing the sum-throughput are identified.

- In Section 7.2, we propose a novel channel matching algorithm for joint mode selection and channel allocation with the goal of maximizing the system throughput under statistical queuing constraints. Seven possible modes are considered, namely the D2D cellular mode, D2D dedicated mode, uplink dedicated mode, downlink dedicated mode, uplink reuse mode, downlink reuse mode, and D2D reuse mode. Using tools from stochastic network calculus, the throughput is characterized by determining the effective capacity. We formulate the channel allocation problem as a maximum-weight matching problem, which can be solved by employing the Hungarian algorithm.

- In Section 7.3, we propose a novel joint mode selection and channel resource allocation algorithm via the vertex coloring approach. In our analysis, we divide

the problem into three subproblems, namely user partition, power allocation and channel assignment. Different from prior works, the power allocation, mode selection and channel assignment are considered after grouping D2D users via vertex coloring. Algorithms are designed for each subproblem, and we propose a novel three-step joint mode selection and resource allocation method by combining these algorithms designed for the three subproblems. We also incorporate the adaptation of the interference constraints in the grouping step when the given interference constraints are relatively loose, and fairness among the users in the same group is considered in the power allocation step.

In Chapter 8, we investigate resource allocation algorithms for content delivery over wireless networks.

- In Section 8.1, we design a caching algorithm that minimizes the average delay of a single-cell cellular networks. We first provide a characterization of the average delay for both cellular and D2D modes, and then we propose a very efficient and robust algorithm to solve the delay minimization problem. Our algorithm is applicable in settings with very general popularity models, with no assumptions on how file popularity varies among different users, and we further extend our algorithm to a more general setting, in which the system parameters and the distributions of channel fading change over time.

- In Section 8.2, we propose, for D2D cellular networks operating under deadline constraints, a scheduling algorithm that manages mode selection, channel allocation and power maximization with acceptable complexity. Our scheduling algorithm is designed based on the convex delay cost method for a D2D cellular network with deadline constraints in an OFDMA setting. All seven

possible modes are included into our scheduling decisions, namely the D2D cellular mode, D2D dedicated mode, uplink dedicated mode, downlink dedicated mode, uplink reuse mode, downlink reuse mode, and D2D reuse mode. Power optimization algorithms are proposed for all possible modes, based on our utility definition.

- In Section 8.3, we propose a two-step ICI-aware scheduling algorithm for C-RAN, which performs user grouping and resource allocation with the goal of minimizing delay violation probability. A novel user grouping algorithm is developed for the user grouping step, which controls the interference among the users in the same group, and we formulate the channel assignment problem in the second step as a maximum-weight matching problem, which can be solved using standard algorithms in graph theory. The performance of our algorithm is verified via simulations, and we compare our algorithm with a conventional SFR algorithm. Also, the influence of the system parameters is investigated with the help of numerical results.

1.4 Bibliographic Note

- The results in Section 3.1 appeared in the journal paper:
  - Y. Li, M. C. Gursoy and S. Velipasalar, "On the Throughput of Hybrid-ARQ Under Statistical Queuing Constraints," in *IEEE Trans. Veh. Technol.*, vol. 64, no. 6, pp. 2725-2732, June 2015.
- The results in Section 3.2 appeared in the journal paper:
  - Y. Li, G. Ozcan, M. C. Gursoy and S. Velipasalar, "Energy Efficiency of Hybrid-ARQ Under Statistical Queuing Constraints," in *IEEE Trans. Commun.*, vol. 64, no. 10, pp. 4253-4267, Oct. 2016

and in the conference paper:

  - Y. Li, G. Ozcan, M. C. Gursoy and S. Velipasalar, "Energy efficiency of hybrid-ARQ systems under QoS constraints," in *Proc. of the 48th Annual Conference on Information Sciences and Systems (CISS)*, Princeton, NJ, 2014, pp. 1-6.

- The results in Section 4.1 appeared in the conference paper:

  - Y. Li, M. C. Gursoy and S. Velipasalar, "Throughput of Hybrid-ARQ Chase Combining with ON-OFF Markov Arrivals under QoS Constraints," in *Proc. of the IEEE Global Communications Conference (GLOBECOM)*, Washington, DC, 2016, pp. 1-6.

- The results in Section 4.2 will appear in the accepted conference paper:

  - Y. Li, M. C. Gursoy and S. Velipasalar, "Throughput of HARQ-IR with Finite Blocklength Codes and QoS Constraints," in *Proc. of the IEEE International Symposium on Information Theory (ISIT)*, Aachen, 2017.

- The results in Section 5.1 appeared in the conference paper:

  - Y. Li, M. C. Gursoy and S. Velipasalar, "Throughput of two-hop wireless channels with queueing constraints and finite blocklength codes," in *Proc. of the IEEE International Symposium on Information Theory (ISIT)*, Barcelona, 2016, pp. 2599-2603.

- The results in Section 5.2 appeared in the conference paper:

  - Yi Li, D. Qiao, M. C. Gursoy and S. Velipasalar, "On the throughput of two-way relay systems under queueing constraints," in *Proc. of the IEEE Global Communications Conference (GLOBECOM)*, Atlanta, GA, 2013, pp. 2003-2008.

- The results in Section 5.3 appeared in the journal paper:
  - Y. Li, M. C. Gursoy and S. Velipasalar, "On the Throughput of Multi-Source Multi-Destination Relay Networks With Queueing Constraints," in *IEEE Trans. Wireless Commun.*, vol. 15, no. 8, pp. 5368-5383, Aug. 2016

  and the conference paper:

  - Y. Li, M. C. Gursoy and S. Velipasalar, "On the throughput of ARQ over multiple-access relay fading channels with queueing constraints," in *Proc. of the IEEE Wireless Communications and Networking Conference (WCNC)*, New Orleans, LA, 2015, pp. 741-746.
- The results in Chapter 6 appeared in the conference paper:
  - Y. Li, M. C. Gursoy and S. Velipasalar, "Throughput and mode selection in two-way MIMO systems under queuing constraints," in *Proc. of the IEEE International Conference on Communications (ICC)*, London, 2015, pp. 2271-2276.
- The results in Section 7.1 appeared in the conference paper:
  - Y. Li, M. C. Gursoy and S. Velipasalar, "Device-to-device communication in cellular networks under statistical queueing constraints," in *Proc. of the IEEE International Conference on Communications (ICC)*, Kuala Lumpur, 2016, pp. 1-6.
- The results in Section 7.2 appeared in the conference paper:
  - Y. Li, M. C. Gursoy and S. Velipasalar, "Joint mode selection and resource allocation for D2D communications under queueing constraints," in

Proc. of the IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS), San Francisco, CA, 2016, pp. 490-495.

- The results in Section 7.3 are reported in the submitted paper:
  - Y. Li, M. C. Gursoy and S. Velipasalar, "Joint Mode Selection and Resource Allocation for D2D Communications via Vertex Coloring," in *Proc. of the IEEE Global Communications Conference (GLOBECOM)*, Singapore, 2017.
- The results in Section 8.1 appeared in the conference paper:
  - Y. Li, M. C. Gursoy and S. Velipasalar, "A Delay-Aware Caching Algorithm for Wireless D2D Caching Networks," in *Proc. of the IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, Atlanta, GA, 2017.
- The results in Section 8.2 appeared in the conference paper:
  - Y. Li, M. C. Gursoy and S. Velipasalar, "Scheduling in D2D Underlaid Cellular Networks with Deadline Constraints," in *Proc. of the IEEE 84th Vehicular Technology Conference (VTC-Fall)*, Montreal, QC, 2016, pp. 1-5.
- The results in Section 8.3 will appear in the accepted conference paper:
  - Y. Li, M. C. Gursoy and S. Velipasalar, "Intercell Interference-Aware Scheduling for Delay Sensitive Applications in C-RAN," in *Proc. of the IEEE 86th Vehicular Technology Conference (VTC-Fall)*, Toronto, ON, 2017.

Chapter 2

Preliminaries of Statistical Queuing Constraints

2.1 Statistical Queuing Constraints

In this thesis, the statistical queuing constraints require the buffer overflow probability to decay exponentially fast, i.e.,[4] [74]

$$\Pr\{Q \geq q\} \approx \varsigma e^{-\theta q}, \tag{2.1}$$

for sufficiently large q, where Q is the stationary queue length, q is the overflow threshold, $\varsigma = \Pr\{Q > 0\}$ is the probability of non-empty buffer, and the non-negative scalar θ is called the QoS exponent. More rigorously, QoS exponent θ is defined as [1] [1]

$$\theta = \lim_{q \to \infty} \frac{-\log \Pr\{Q \geq q\}}{q}. \tag{2.2}$$

[1]Throughout the text, logarithm expressed without a base, i.e., $\log(\cdot)$, refers to the natural logarithm $\log_e(\cdot)$.

Note that θ is a factor that controls the exponential decay rate of the buffer overflow probability. From (2.1), we notice that higher values of θ indicate stricter limitations on the buffer overflow probability, leading to more stringent QoS constraints whereas lower values of θ represent looser QoS requirements. Conversely, for a given buffer threshold q and overflow probability limit $\Pr\{Q \geq q\} = \delta$, the desired value of θ can be determined as

$$\theta = -\frac{1}{q}\log\delta + \frac{1}{q}\log\varsigma. \tag{2.3}$$

As $q \to \infty$, the term $\frac{1}{q}\log\varsigma$ in (2.3) vanishes, which leads to (2.2).

2.2 Throughput of Single-hop Channels under Statistical Queuing Constraints

At the buffer, the arrival rates a_i (bits/block) and the departure rates c_i (bits/block) form the arrival and departure processes, respectively, where i is the time index. According to the effective bandwidth and effective capacity formulations provided in [1] and [4], respectively, in the presence of queuing constraints with QoS exponent θ, the arrival process and departure process at the buffer should satisfy

$$\Lambda_a(\theta) + \Lambda_c(-\theta) = 0, \tag{2.4}$$

where $\Lambda_p(\theta) = \lim_{t\to\infty}\frac{1}{t}\log_e\ \mathbb{E}\{e^{\theta\sum_{i=1}^{t}p_i}\}$ is the asymptotic logarithmic moment generating function (LMGF) of the random process p_i.

2.2.1 Effective Capacity

When the arrival rate is constant i.e., $a_i = a$ for all i, it can be easily seen that

$$\Lambda_a(\theta) = a\theta. \tag{2.5}$$

Then, from (2.4), we have

$$a = -\frac{1}{\theta}\Lambda_c(-\theta). \tag{2.6}$$

Indeed, the right-hand side of (2.6) is defined as the effective capacity of the wireless link [4]

$$C_E(\theta, \text{SNR}) = -\frac{1}{\theta}\Lambda_c(-\theta), \tag{2.7}$$

characterizing the maximum constant arrival rate that can be supported by the time-varying wireless transmission rates while satisfying the statistical queuing constraint in (2.1). Therefore, under the constant-rate arrival assumption, the maximum average arrival rate is given by the effective capacity:

$$r_{\text{avg}}(\theta, \text{SNR}) = \mathbb{E}\{a_i\} = a = C_E(\theta, \text{SNR}) = -\frac{1}{\theta}\Lambda_c(-\theta). \tag{2.8}$$

2.2.2 Average Arrival Rates of Random Arrival Sources under Statistical Queuing Constraints

In effective capacity analysis, constant-rate arrivals are often assumed at the transmitter. On the other hand, randomly time-varying arrivals are frequent in real applications. For instance, the data traffic can be regarded as an ON-OFF process in voice communications (e.g., in VoIP) and variable bit-rate video traffic is statistical-

ly characterized as autoregressive, Markovian, or Markov-modulated processes [75]. With this motivation, the authors in [76] studied the impact of source burstiness on the energy efficiency under statistical queuing constraints, and they further developed energy-efficient power control policies in [77] considering Markov arrivals.

When the arrival rate is not constant, the computation of the system throughput is more complicated. In general, we need to formulate the LMGF of the arrival process as a function of the average arrival rate, and obtain the throughput by solving (2.4). In this thesis, three random arrival sources are considered, which are the ON-OFF discrete Markov and Markov fluid sources, and ON-OFF Markov modulated Poisson sources (MMPS), in Chapters 3 and 4. After characterizing their LMGF of the arrival process, the average arrival rates can be obtained as functions of the effective capacities by solving (2.4).

2.2.2.1 Average Arrival Rates of ON-OFF Discrete-Time Markov Source under Statistical Queuing Constraints

ON-OFF discrete-time Markov source only has two states, namely, ON and OFF states. We define state 1 as the OFF state, in which the source keeps silent. When the source is in ON state, or equivalently state 2, the arrival rate is $a_i = r$. The state transition probability matrix of this Markov source can be written as

$$\mathbf{G} = \begin{pmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{pmatrix}, \tag{2.9}$$

where p_{11} and p_{22} denote the probabilities that the source remains in the same state (OFF and ON states, respectively) in the next time block, and p_{12} and p_{21} are the probabilities that source will transition to a different state in the next time block. Using the properties of Markov processes, we can express the probability of the ON

state as

$$P_{ON} = \frac{1 - p_{11}}{2 - p_{11} - p_{22}}. \tag{2.10}$$

Then, the average arrival rate of this ON-OFF Markov source is

$$r_{\text{avg}} = rP_{ON} = r\frac{1 - p_{11}}{2 - p_{11} - p_{22}}. \tag{2.11}$$

From [78], the LMGFs of this arrival process is given by

$$\Lambda_a(\theta) = \log_e \left(\frac{p_{11} + p_{22}e^{r\theta} + \sqrt{(p_{11} + p_{22}e^{r\theta})^2 - 4(p_{11} + p_{22} - 1)e^{r\theta}}}{2} \right). \tag{2.12}$$

Plugging the characterizations in (2.12) and (2.7) into (2.4), we obtain the arrival rate in the ON state as

$$r = \frac{1}{\theta} \log \left(\frac{e^{2\theta C_E} - p_{11}e^{\theta C_E}}{1 - p_{11} - p_{22} + p_{22}e^{\theta C_E}} \right). \tag{2.13}$$

Further inserting this result into (2.11), we find the maximum average arrival rate as

$$r_{\text{avg}} = \frac{P_{ON}}{\theta} \log \left(\frac{e^{2\theta C_E} - p_{11}e^{\theta C_E}}{1 - p_{11} - p_{22} + p_{22}e^{\theta C_E}} \right), \tag{2.14}$$

where P_{ON} is given in (2.10).

2.2.2.2 Average Arrival Rates of ON-OFF Fluid Markov Source under Statistical Queuing Constraints

Different from the discrete-time Markov source whose state does not change in a given time block and state transitions occur in discrete time steps, fluid Markov source may stay in a state over a continuous duration of time. In other words, the source can change its state at any time. Here, the definitions of ON and OFF states are the same as for the ON-OFF discrete-time source. The generating matrix of this

continuous-time Markov process is given by

$$\mathbf{G} = \begin{pmatrix} -\alpha & \alpha \\ \beta & -\beta \end{pmatrix}, \tag{2.15}$$

and the ON state probability is

$$P_{ON} = \frac{\alpha}{\alpha + \beta}. \tag{2.16}$$

In this case, the average arrival rate is

$$r_{\text{avg}} = rP_{ON} = r\frac{\alpha}{\alpha + \beta}. \tag{2.17}$$

Using a similar approach as for the discrete-time Markov source, we can find the maximum average arrival rate for the ON-OFF fluid Markov source, under statistical queuing constraints. From [79], the LMGF of the arrival process of the ON-OFF Markov fluid source is given by

$$\Lambda_a(\theta) = \frac{1}{2}\left(\theta r - \alpha - \beta + \sqrt{(\theta r - \alpha - \beta)^2 + 4\alpha\theta r}\right). \tag{2.18}$$

Plugging (2.18) into (2.4), we can find that

$$r = C_E\frac{\theta C_E + \alpha + \beta}{\theta C_E + \alpha}. \tag{2.19}$$

Inserting this result into (2.17), we determine the maximum average arrival rate as

$$r_{\text{avg}} = P_{ON}C_E\frac{\theta C_E + \alpha + \beta}{\theta C_E + \alpha}, \tag{2.20}$$

where P_{ON} is given in (2.16).

2.2.2.3 Average Arrival Rates of ON-OFF MMPS under Statistical Queuing Constraints

The arrival rates of ON-OFF MMPS models are described as a Poisson process with intensity ρ in the ON state while there is no arrival in the OFF state. State transitions are governed by a continuous-time Markov chain as in the Markov fluid model. However, compared to the ON-OFF Markov fluid source analyzed in Section 2.2.2.2, MMPS can be seen to have a higher degree of burstiness since its arrival rate, rather than being a constant, is random in the ON state. Here, the expressions of the generating matrix and ON state probability are the same as in Section 2.2.2.2. In this case, the average arrival rate is

$$r_{\text{avg}} = \rho P_{ON} = \rho \frac{\alpha}{\alpha + \beta}. \tag{2.21}$$

From [79], the LMGF of the arrival process of the ON-OFF MMPS is given by

$$\Lambda_a(\theta) = \frac{1}{2}\left[(e^{\theta}-1)\rho - (\alpha+\beta)\right] + \frac{1}{2}\sqrt{\left[(e^{\theta}-1)\rho - (\alpha+\beta)\right]^2 + 4\alpha\rho(e^{\theta}-1)}. \tag{2.22}$$

Plugging (2.22) into (2.4), we can find

$$\rho = \frac{\theta\left[\theta C_E + \alpha + \beta\right]}{(e^{\theta}-1)\left[\theta C_E + \alpha\right]} C_E. \tag{2.23}$$

Inserting (2.23) into (2.21), we find the maximum average arrival rate as

$$r_{\text{avg}} = P_{ON} C_{E,Q_1} \frac{\theta}{e^{\theta}-1} \frac{\theta C_{E,Q_1} + \alpha + \beta}{\theta C_{E,Q_1} + \alpha}, \tag{2.24}$$

where P_{ON} is given in (2.16).

2.3 Throughput of Two-hop Channels under Statistical Queuing Constraints

In a two-hop channel, source node **S** sends information to the receiver **D** with the help of the intermediate relay node **R**, and there is no direct link between **S** and **D** (which, for instance, holds, if these nodes are sufficiently far apart in distance). The system model of two-hop relay channels is shown in Figure 5.1 in Section 5.1. Both the source and the intermediate relay nodes operate under statistical queuing constraints with QoS exponents θ_1 and θ_2, resectively. In such cases, we assume a constant arrival rate at the source node, and we characterize the maximum constant arrival rate at the source node that can be supported by the two-hop link under queuing constraints as the throughput. From the theory of effective bandwidth and effective capacity [1], [2], [4], the queuing constraint can be satisfied at the source node when the constant arrival rate R satisfies

$$\mathrm{R} = -\frac{\Lambda_{\mathbf{S},\mathbf{R}}(-\tilde{\theta})}{\tilde{\theta}} \tag{2.25}$$

for some $\tilde{\theta} \geq \theta_1$, where $\Lambda_{\mathbf{S},\mathbf{R}}$ is the LMGF of the service (or equivalently transmission) rate at the source. The above arrival rate formulation considers only the queuing constraints at the source node. However, we need to address the constraints at the relay buffer as well. It was shown in [2] that the queuing constraint can be satisfied at the relay if we have

$$\Lambda_{\mathbf{R}}(\hat{\theta}) + \Lambda_{\mathbf{R},\mathbf{D}}(-\hat{\theta}) = 0 \tag{2.26}$$

for some $\hat{\theta} \geq \theta_2$. Above, $\Lambda_{\mathbf{R},\mathbf{D}}$ is the LMGF of the service rate at relay. In (2.26), $\Lambda_{\mathbf{R}}$ is the LGMF of the arrival process to $\mathbf{R}$ and is formulated as [2, equation (18)]

$$\Lambda_{\mathbf{R}}(\theta) = \begin{cases} \mathrm{R}\theta, & 0 \leq \theta \leq \tilde{\theta} \\ \mathrm{R}\tilde{\theta} + \Lambda_{\mathbf{S},\mathbf{R}}(\theta - \tilde{\theta}), & \theta > \tilde{\theta} \ . \end{cases} \tag{2.27}$$

Hence, in order to satisfy the queuing constraints at both the source and relay nodes, the arrival rate R should satisfy (2.25) and (2.26) simultaneously, which implies that R should be the minimum of the rates obtained from (2.25) and (2.26).

Chapter 3

Throughput and Energy Efficiency of Hybrid-ARQ under Statistical Queuing Constraints with Low QoS Exponents

HARQ is a high performance communication protocol, leading to effective use of the wireless channel and the resources with only limited feedback about the CSI to the transmitter. In this chapter, we analyze the throughput and energy efficiency of HARQ protocols in the presence of statistical queuing requirements when the QoS exponent θ is sufficiently small via Taylor expansion.

In Section 3.1, the throughput of HARQ with fixed transmission rate is studied. In particular, tools from the theory of renewal processes and stochastic network calculus are employed to characterize the maximum arrival rates that can be supported by the wireless channel when HARQ is adopted. Effective capacity is employed as the throughput metric and a closed-form expression for the effective capacity of HARQ is determined for small values of the QoS exponent. The impact of the fixed trans-

mission rate, queuing constraints, and hard deadline limitations on the throughput is investigated.

In Section 3.2, the energy efficiency of the HARQ chase combining scheme under outage, deadline, and statistical queuing constraints in the low-power and low-θ regimes is studied for both constant-rate and random Markov arrivals by characterizing the minimum energy per bit and wideband slope. In particular, two queuing models are considered. Specifically, when outage occurs, the transmitter keeps the packet, lowers its priority, and attempts to retransmit it later in the first queue model while the packet is discarded and removed from the buffer in the second queue model. For both models, energy efficiency is investigated when outage constraints, statistical queuing constraints and deadline constraints are imposed. The deadline constraint provides a limitation on the number of retransmissions or equivalently the number of HARQ rounds. Under these assumptions, closed-form expressions are obtained for the minimum energy per bit and wideband slope for HARQ-CC, and comparisons among different arrival models are made. For instance, it is shown that stricter queuing constraints and more bursty sources degrade the energy efficiency by lowering the wideband slope. In the numerical results, analytical characterizations are verified through simulations. Moreover, the impact of source variations/burstiness, deadline constraints, outage probability, queuing constraints on the energy efficiency is analyzed.

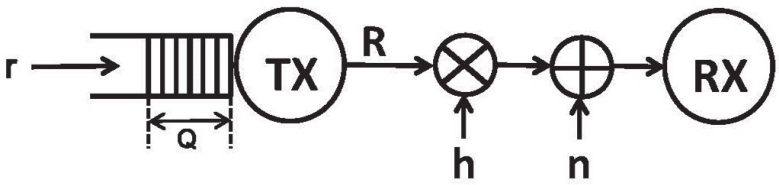

Figure 3.1: System Model

3.1 Throughput of HARQ under Statistical Queuing Constraints

3.1.1 System Model

3.1.1.1 Fading Channel

We consider a point-to-point wireless link shown in Fig. 3.1 and assume that block fading is experienced in the channel. More specifically, in each block duration of T_s seconds, fading is assumed to stay fixed and then change independently in the subsequent block. We assume that transmissions occur in blocks and one fading block is our basic time unit throughout the paper. In each block duration, transmitter sends l symbols to the receiver. In the i^{th} block, the transmitter sends the l-dimensional signal vector $\mathbf{x}_i$ with average energy $\mathbb{E}\{\|\mathbf{x}_i\|^2\} = l\mathcal{E}$, and the received discrete time signal can be expressed as

$$\mathbf{y}_i = h_i\mathbf{x}_i + \mathbf{n}_i \quad i = 1, 2, \ldots \tag{3.1}$$

where h_i is the channel fading coefficient in this time block, and $\mathbf{n}_i$ denotes the noise vector with i.i.d. complex, circularly-symmetric Gaussian components with zero-mean and variance N_0. Then, the instantaneous capacity in each fading block can be expressed as

$$C_i = T_s B \log_2(1 + \text{SNR} z_i) \quad \text{bits/block} \tag{3.2}$$

where B is the system bandwidth, T_s is the block duration, $\text{SNR} = \frac{\mathbb{E}\{\|\mathbf{x}\|^2\}}{E\{\|\mathbf{n}\|^2\}} = \frac{l\mathcal{E}}{lN_0} = \frac{\mathcal{E}}{N_0}$ represents the transmitted average signal-to-noise ratio, and $z_i = |h_i|^2$ denotes the magnitude-square of the fading coefficient.

3.1.1.2 HARQ-IR with Fixed-Rate Transmissions

We assume that the transmitter sends information at the constant rate of R bits/block[1] and a Type-II HARQ-IR protocol is employed for reliable reception. In this scheme, the messages at the transmitter are encoded according to a certain codebook and the codewords are divided into a number of subblocks of the same length. During each fading block, only one subblock is sent to the receiver. At the receiver side, the transmitted message is decoded according to the current received subblock combined with the previously received subblocks related to the current transmitted message. In this case, information accumulates at the receiver side. According to information-theoretical results [80], the receiver can decode the transmitted message at the end of the N^{th} subblock without error only if R satisfies

$$R < T_s B \sum_{i=1}^{N} \log_2(1 + \text{SNR} z_i). \tag{3.3}$$

Hence, with the above characterization, we consider the maximum achievable rates of HARQ with an information-theoretic perspective as in [80]. Indeed, a coding strategy for HARQ-IR is described in detail in [80] for messages to be decoded successfully when (3.3) is satisfied. Hence, if (3.3) is satisfied, the receiver gets R bits of information, an ACK feedback signal is sent, and the first subblock of a new message is transmitted in the next interval. We assume that the decoder at the receiver has the ability to detect transmission errors reliably. Therefore, if R does not satisfy (3.3),

[1]More accurately, we assume that the transmitter, after each successful transmission, attempts to send R bits within the next transmission duration. If the transmitted codeword is successfully decoded in the first fading block, the received data rate is R bits/block. If successful decoding occurs at the end of N^{th} fading block, the received data rate is R/N bits/block.

receiver detects the error and sends NACK feedback to the transmitter, triggering the transmission of the next subblock of the same message in the subsequent transmission interval.

We define the random transmission time T of a message as

$$T = \min \left\{ N : R < T_s B \sum_{i=1}^{N} \log_2(1 + \text{SNR} z_i) \right\}. \tag{3.4}$$

Hence, T denotes the number of block-fading channel uses needed to successfully send a message. In our HARQ-IR model, if a renewal event occurs when the receiver decodes the transmitted message correctly. Therefore, T describes the interarrival time (in terms of number of fading blocks) between consequent renewal events.

It is shown in [80] that the throughput of this HARQ-IR scheme is given by

$$\gamma = \frac{R}{\mathbb{E}\{T\}} = \frac{R}{\mu_1} \tag{3.5}$$

where μ_1 denotes the expected value of T. Additionally, it is proven in [80] that as $R \to \infty$, the throughput approaches the ergodic capacity, i.e.,

$$\lim_{R \to \infty} \gamma = \mathbb{E}\{T_s B \log_2(1 + \text{SNR} z_i)\} = \mathbb{E}\{C_i\}. \tag{3.6}$$

3.1.2 Effective Capacity of the HARQ-IR Scheme

We assume that the transmitter operates under queuing constraints imposed as limitations on the buffer overflow probability, which is introduced in Chapter 2. In [81] the notion of effective capacity was proposed to analyse the system throughput under such constraints, which provides the maximum constant arrival rates that can be supported while satisfying (2.1) in Chapter 2. From (2.8) in Chapter 2, the maximum arrival rate or equivalently the effective capacity under this queuing constraint is given by

$$C_e = -\lim_{t\to\infty} \frac{1}{\theta t} \log_e \mathbb{E}\{\mathrm{e}^{-\theta S_t}\} \tag{3.7}$$

$$= -\lim_{t\to\infty} \frac{1}{\theta t} \log_e \mathbb{E}\{\mathrm{e}^{-\theta R N_t}\} \tag{3.8}$$

where $S_t = \sum_{i=1}^{i=t} s[i]$ is the time-accumulated service process representing the total number of bits sent until time t. In our system setting, if we denote the number of successful message transmissions until time t by N_t, then $S_t = RN_t$. Note that N_t is the number of renewals made by time t and hence $\{N_t\}$ can be regarded as the renewal counting process with i.i.d. interarrival intervals. More explicitly, we can define N_t as

$$N_t = \max\left\{k : \sum_{j=1}^{k} T_j < t\right\} \tag{3.9}$$

where $\{T_j\}$ is the i.i.d. sequence of durations of successful transmissions of consecutive messages. Note that T_j is the number of fading blocks needed to successfully decode the j^{th} message. Renewals occur when the receiver decodes a packet successfully and N_t is number of renewal events (or equivalently the number of instances R bits of information has been successfully received) up until time t.

Using the properties of renewal processes, we obtain the following closed-form expression of the effective capacity for small θ values in terms of the statistical averages of the random transmission time T.

Theorem 1 *For the HARQ-IR scheme with fixed rate transmissions, the effective capacity in (3.8) has the following first order expansion with respect to the QoS exponent θ around $\theta = 0$:*

$$C_e = \frac{R}{\mu_1} - \frac{R^2\sigma^2}{2\mu_1^3}\theta + o(\theta), \tag{3.10}$$

where R is the fixed transmission rate, μ_1 and σ^2 are the mean and variance of the random transmission time T, and θ is the QoS exponent. Note that μ_1 and σ^2 are also functions of R. Finally, $o(\theta)$ represents the terms that vanish faster than θ as $\theta \to 0$, i.e., $\lim_{\theta\to 0} \frac{o(\theta)}{\theta} = 0$.

Proof: See Appendix A.1.

To calculate effective capacity, we need the mean and variance of the transmission time. For the HARQ-IR scheme, the distribution of T is not available in closed-form and hence we resort to Monte-Carlo simulations to obtain μ_1 and σ^2.

Remark 1 *We note that if no queuing constraints are imposed and hence $\theta = 0$, the effective capacity expression in (3.10) specializes to $C_e = \frac{R}{\mu_1}$ and therefore we recover the throughput formulation obtained in [80]. Additionally, we notice in (3.10) that since $R \geq 0$, $\mu_1 \geq 0$, and $\sigma^2 \geq 0$, the introduction of the queuing constraints even with small QoS exponent θ leads to a loss in the throughput, which was quantified by the term $-\frac{R^2\sigma^2}{2\mu_1^3}\theta$. Finally, another observation is that while depending only on μ_1 when $\theta = 0$, the throughput starts also depending on the variance, σ^2, of the random transmission time in the presence of queuing requirements. Indeed, the larger the variance, the smaller the throughput becomes in the regime of small θ.*

Remark 2 *By the Central Limit Theorem for renewal counting processes [82], if the inter-renewal intervals have finite variance σ^2, then we have the following convergence in distribution*

$$\frac{N_t - \frac{t}{\mu_1}}{\sigma \mu_1^{-\frac{3}{2}} t^{1/2}} \longrightarrow \mathcal{N}(0,1) \quad \text{as } t \to \infty. \tag{3.11}$$

Hence, the distribution of N_t tends to a Gaussian distribution with mean $\frac{t}{\mu_1}$ and variance $\frac{\sigma^2 t}{\mu_1^3}$ for large t. Now, if we approximate the distribution of N_t as

$$f_{N_t}(x) \approx \frac{1}{\sqrt{2\pi \frac{\sigma^2 t}{\mu_1^3}}} \exp\left(-\frac{\left(x - \frac{t}{\mu}\right)^2}{\frac{\sigma^2 t}{\mu_1^3}} \right) \quad \text{for large } t, \tag{3.12}$$

then plugging the parameters into the moment generating function of Gaussian distribution, we can obtain

$$\mathbb{E}\{\mathrm{e}^{-\theta R N_t}\} \approx \exp\left(-\frac{R}{\mu}\theta t + \frac{R^2\sigma^2}{2\mu_1^3}\theta^2 t\right) \tag{3.13}$$

which implies that

$$C_e = -\lim_{t\to\infty}\frac{1}{\theta t}\log_e \mathbb{E}\{\mathrm{e}^{-\theta R N_t}\} \approx \frac{R}{\mu_1} - \frac{R^2\sigma^2}{2\mu_1^3}\theta. \tag{3.14}$$

Remark 3 *Theorem 1 can also be applied to Type-II HARQ-CC protocol. In HARQ-CC protocol, the transmitter just sends the same coded data in each retransmission, and the receiver uses maximum-ratio combining for decoding. Therefore, the packet can be decoded within N fading blocks only if R satisfies*

$$R \le T_s B \log_2\left(1 + \sum_{i=1}^{N} SNRz_i\right). \tag{3.15}$$

Since the number of successful message transmissions, N_t, can still be regarded as a renewal counting process, and all moments of the transmission time T are also finite in Chase Combining, the characterization in (3.10) applies to this type of HARQ as well.

3.1.2.1 Hard Deadline Constraints

Heretofore, we have not considered any restrictions on the random transmission time T. Hence, the number of block-fading channel uses needed to successfully send a message can be arbitrarily large especially if the transmission rate R is also large. Indeed, as will be evidenced in the numerical results, throughput improves as R increases but this comes at the cost of increased transmission time. On the other hand, practical systems can require hard deadline constraints for the messages and it

is of interest to have bounds on T. For instance, we can impose

$$T \leq T_u, \tag{3.16}$$

and hence limit the number of HARQ rounds to send a message by T_u. More specifically, if $R > T_s B \sum_{i=1}^{T_u} \log_2(1+\mathrm{SNR} z_i)$ and hence the message is not correctly decoded at the end of the T_u^{th} transmission, the transmitter initiates the transmission of the new message. Detailed discussions about dealing with outage events is provided in Section 3.2. Here we consider a case, in which the transmitted packet is not removed from the buffer when an outage happens. The transmitter reduces its priority, and starts transmitting the packet with the highest priority in the next time block. This scenario corresponds to the queue model I with deadline constraints in Section 3.2. Under this situation, the system has to operate under queuing constraint and deadline constraint.

We can easily see that the characterization in Theorem 1 applies in the presence of hard-deadline constraints as well, once we adopt the following approach. We define $\hat{T}$ as the total duration of time that has taken to successfully send one message, including the periods of failed transmissions due to imposing the upper bound T_u. In this case, the count starts from 0 after a successful decoding, and transmission time $\hat{T}$ increases until the next successful decoding. Again, the renewal events happen only when the receiver can decode the packet successfully. Now, the probability that $\hat{T} = n + kT_u$, i.e., the probability that the transmission of first k messages have ended in failure due to the deadline constraint and $(k+1)^{\text{th}}$ message is successfully transmitted after $n \leq T_u$ HARQ transmissions, can be expressed as

$$\Pr\{\hat{T} = n + kT_u\} = (\Pr\{T > T_u\})^k \Pr\{T = n\} \quad \text{for } n = 1, 2, \ldots, T_u \text{ and } k = 0, 1, 2, \ldots \tag{3.17}$$

where T is as defined in (3.4). Under the upper bound constraint T_u, the new inter-renewal time between successful message transmissions is $\hat{T}$. Hence, only the statistical description of inter-renewal time changes and the throughput formulation in (3.10) still applies but now with $\mu_1 = \mathbb{E}\{\hat{T}\}$ and $\sigma^2 = \text{var}(\hat{T})$.

Note that the inter-renewal time $\hat{T}$ can grow very fast on average with increasing rate R. This is due to the fact that the likelihood to complete the message transmission within T_u intervals becomes small for large R. Hence, many message transmissions can fail before a successful transmission. More specifically, as R increases, $\Pr\{T > T_u\}$ grows, increasing the probability of large values of $\hat{T}$ and also increasing $\mu_1 = \mathbb{E}\{\hat{T}\}$. This growth is faster than what would be experienced in the absence of hard-deadline constraints and it can lower the throughput significantly if R is larger than a threshold.

3.1.3 Numerical Results

In this subsection, we provide our numerical results. In particular, we focus on the relationship between the transmission rate R and our throughput metric C_e. In our results, we both compute the first-order expansion of the effective capacity given in (3.10) and also simulate the HARQ-IR transmissions and estimate the effective capacity by computing $-\frac{1}{\theta t}\log_e \mathbb{E}\{\mathrm{e}^{-\theta R N_t}\}$ for large t. More specifically, $\mathbb{E}\{\mathrm{e}^{-\theta R N_t}\}$, the expected value and variance of the transmission time are determined via Monte-Carlo simulations. In the numerical analysis, we assume the fading coefficient h_i has a circularly symmetric complex Gaussian distribution with zero mean and variance 1. Hence, we consider a Rayleigh fading environment.

In Fig. 3.2, we plot the effective capacity C_e as a function of the transmission rate R for Type-I HARQ, HARQ Chase Combining and HARQ-IR schemes. The throughput curves of HARQ-IR and HARQ Chase Combining are plotted both by computing the first-order expansion in (3.10) and also via simulation. We immediately

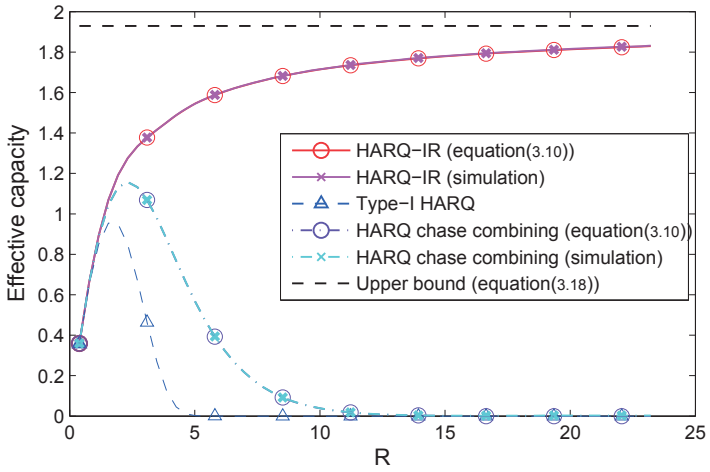

Figure 3.2: Effective capacity C_e vs. transmission rate R at SNR $= 6$ dB and $\theta = 0.01$ for both ARQ and HARQ-IR.

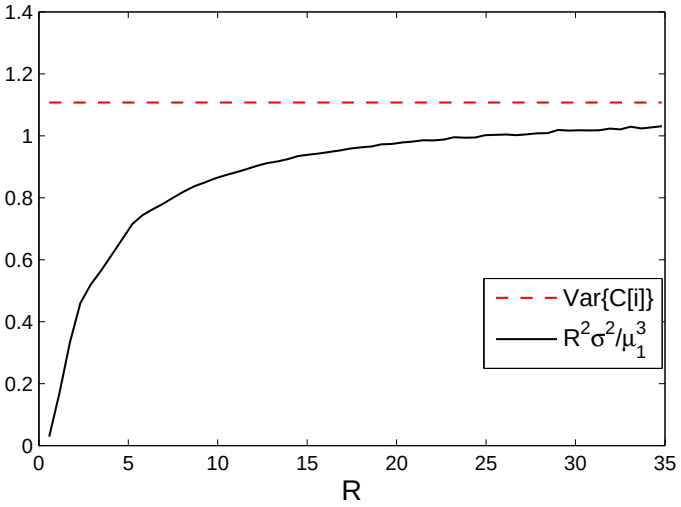

Figure 3.3: $\text{var}(\log_2(1 + \text{SNR}z))$ and $\frac{R^2\sigma^2}{\mu_1^3}$ vs. transmission rate R. SNR $= 6$ dB and $\theta = 0.01$.

notice that for both HARQ-IR and HARQ Chase Combining, the effective capacity approximation provided by the first-order expansion is very close to that obtained by simulation for $\theta = 0.01$. Hence, as predicted in Section 3.1.2, first-order expansion gives an accurate characterization of the throughput of HARQ-IR and HARQ Chase Combining. In the figure, we further observe that HARQ-IR significantly outperforms Type-I HARQ and HARQ Chase Combining. Throughput of Type-I HARQ initially increases and reaches its peak value at an optimal value R^* beyond which it starts to diminish. Hence, in Type-I HARQ, rates higher than the optimal R^* are leading to a large number of retransmissions and resulting in lower throughput. Similar behavior is shown by HARQ Chase Combining. On the other hand, the throughput of HARQ-IR interestingly improves with increasing R and approaches

$$\begin{aligned} C_{e,\text{perfect CSI}} &= -\frac{1}{\theta}\log_e \mathbb{E}\{e^{-\theta C}\} \\ &= -\frac{1}{\theta}\log_e \mathbb{E}\{e^{-\theta T_s B \log_2(1+\text{SNR}z)}\} \end{aligned} \tag{3.18}$$

which is the effective capacity of a system in which the transmitter knows the channel fading coefficients perfectly and transmits the data at the time-varying rate of $B\log_2(1+\text{SNR}z)$ in each block. Note that this observation can be seen as the extension of (3.6) to the case with queuing constraints. Furthermore, it can be easily verified that the first-order expansion of $C_{e,\text{perfect CSI}}$ is given by

$$C_{e,\text{perfect CSI}} = \mathbb{E}\{T_s B\log_2(1+\text{SNR}z)\} - \text{var}(T_s B\log_2(1+\text{SNR}z))\frac{\theta}{2} + o(\theta) \tag{3.19}$$

where $\text{var}(T_s B\log_2(1+\text{SNR}z))$ denotes the variance of $T_s B\log_2(1+\text{SNR}z)$. Comparing this expansion with (3.10) and noting the limiting result in (3.6) and the observation in Fig. 3.2, we expect that $\frac{R^2\sigma^2}{\mu_1^3}$ approaches $\text{var}(T_s B\log_2(1+\text{SNR}z))$ as R increases, which is verified numerically in Fig. 3.3.

The improvement in the throughput of HARQ-IR with increasing R comes at the

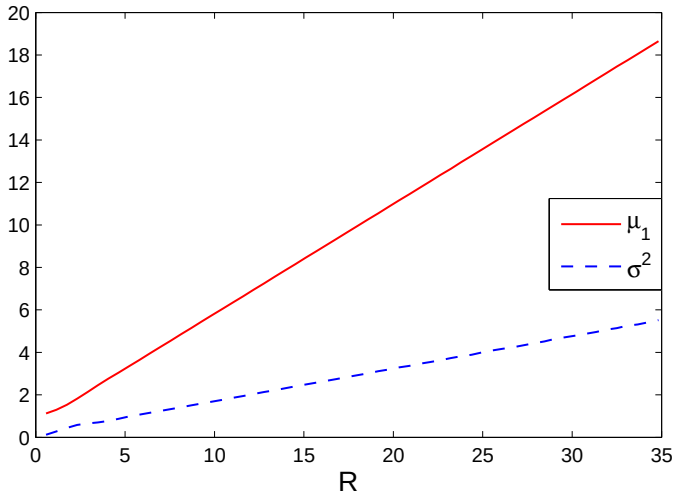

Figure 3.4: Mean μ_1 and the variance σ^2 of the transmission time T vs. transmission rate R. SNR = 6 dB and $\theta = 0.01$.

cost of increased transmission time. This is demonstrated in Fig. 3.4 which shows that both the mean $\mu_1 = \mathbb{E}\{T\}$ and the variance $\sigma^2 = \text{var}(T)$ of the random transmission time T increases with increasing R. Two curves in Fig. 3.4 are obtained using simulation results. It is interesting to note that this increased transmission time in HARQ-IR does not have detrimental impact on the throughput under queuing constraints, which is a testament to the efficient utilization of the channel and resources by HARQ-IR. Indeed, it takes more time to send the data but proportionally a large amount of data is sent successfully with HARQ-IR over this extended period of time. Another observation in Fig. 3.4 is at the other end of the line. As R diminishes, μ_1 and σ^2 approach 1 and 0, respectively. This implies from (3.10) that $C_e \approx R$ for very small R, explaining the linear growth of the effective capacity curve of HARQ-IR for small R values in Fig. 3.2.

In Fig. 3.5, we plot the effective capacity vs. R curve for different values of the QoS exponent θ. We see that larger θ values (and hence stricter queuing constraints) expectedly lead to lower throughput. Equivalently, as θ increases, the same effective capacity is achieved by transmitting at higher rates R and hence by potentially ex-

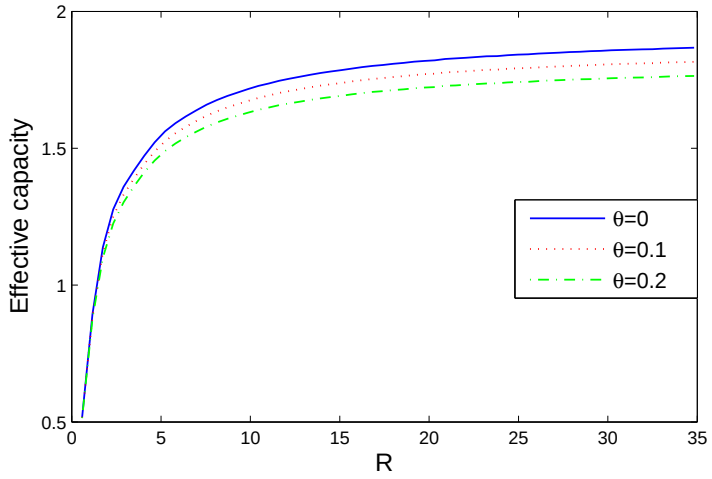

Figure 3.5: Effective capacity C_e of HARQ-IR vs. transmission rate R at SNR = 6 dB for different θ values.

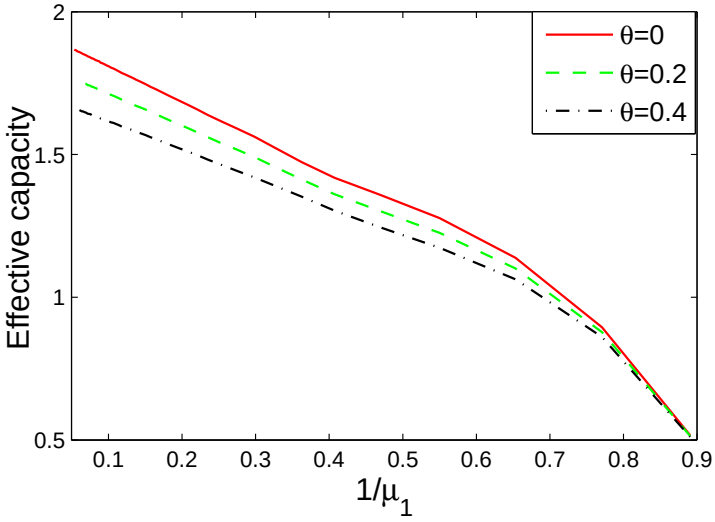

Figure 3.6: Effective capacity C_e vs. $\frac{1}{\mu_1}$ at SNR = 6 dB for different θ values.

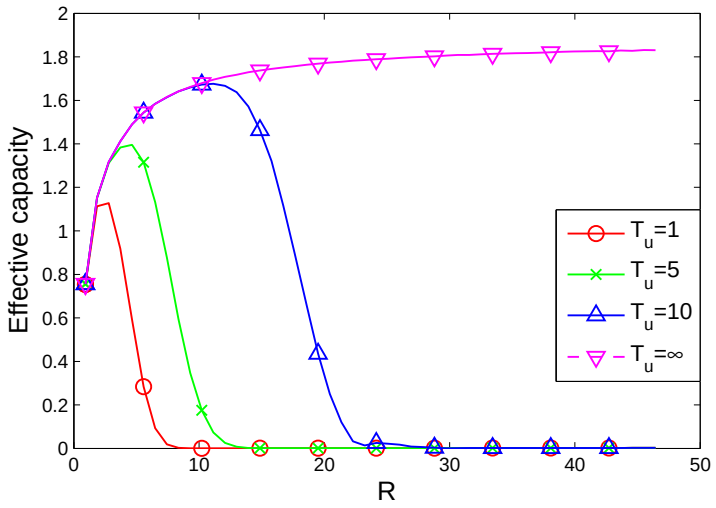

Figure 3.7: Effective capacity C_e vs. transmission rate R for different hard deadline constraints. SNR = 6 dB and $\theta = 0.1$

periencing larger transmission time as depicted in Fig. 3.6. We note in Fig. 3.6 that especially for high effective capacities, when θ is increased, the same effective capacity is achieved at smaller values of $\frac{1}{\mu_1} = \frac{1}{\mathbb{E}\{T\}}$.

Finally, we address the impact of hard-deadline constraints in Fig. 3.7. We plot C_e vs. R curves for different values of the upper bound T_u on the transmission time T (or equivalently the number of HARQ rounds). Again, the expected value and variance of the transmission time are obtained via simulation, and effective capacities are calculated using our first-order approximation. We readily observe that when hard deadline constraints are imposed, there exists an optimal transmission rate $R^*(T_u)$ at which the throughput is maximized and beyond which the throughput starts diminishing. The optimal $R^*(T_u)$ and the achieved maximum throughput get larger for larger T_u while the throughput monotonically increases with increasing R when no deadline constraints are imposed, i.e., when $T_u = \infty$.

3.2 Energy Efficiency of Hybrid-ARQ under Statistical Queuing Constraints

3.2.1 System Model and Preliminaries

In this subsection, we introduce our system model, preliminaries on the HARQ-CC scheme, and energy efficiency metrics in the low-SNR regime. First, we describe our system and channel model. In order to enhance the reliability, the system employs HARQ-CC scheme with fixed transmission rate. A brief introduction on HARQ-CC is provided following the introduction on the system model. A detailed discussion regarding dealing with outage events is given in Section 3.2.1.3. Finally, we introduce the two energy efficiency metrics, namely minimum energy per bit and wideband slope, in the low-SNR regime.

3.2.1.1 System Model

In this section, as depicted in Fig. 3.1, the same point-to-point wireless communication system is considered, in which data packets arriving from the source are initially stored in a buffer at the transmitter before being sent over a fading channel to a receiver. We assume a block flat-fading model in which the fading coefficients stay the same within one block, but change independently across blocks. Each fading block is assumed to have a duration of l symbols. Throughout this section, we use subscript i as the discrete time index. The received signal in the i^{th} block can be written as

$$\mathbf{y}_i = h_i \mathbf{x}_i + \mathbf{n}_i \quad i = 1, 2, \dots \tag{3.20}$$

Above, $\mathbf{x}_i$ and $\mathbf{y}_i$ are the transmitted and received signal vectors of length l, respectively, and h_i denotes the channel fading coefficient in the i^{th} block. Also, $\mathbf{n}_i$ represents the noise vector with i.i.d. circularly-symmetric, zero-mean Gaussian com-

ponents, each with variance N_0 [2]. Then, the instantaneous capacity (bits/symbol) in the i^{th} block is given by

$$C_i = \log_2(1 + \text{SNR} z_i), \tag{3.21}$$

where $z_i = |h_i|^2$ is the magnitude-square of the fading coefficient, $\text{SNR} = \frac{\mathcal{E}}{N_0}$ denotes the signal-to-noise ratio, and $\mathcal{E} = \frac{1}{l}\mathbb{E}\{\|\mathbf{x}_i\|^2\}$ is the average energy per transmitted symbol.

3.2.1.2 HARQ-CC

To guarantee the reliability of the system, we assume that the system employs HARQ-CC scheme with fixed transmission rate R (bits/symbol), and the size of each packet is fixed at lR bits, where l is the number of symbols in each fading block. If the receiver decodes the received packet correctly, it sends an ACK feedback to the transmitter through an error-free feedback link, and a new packet will be sent in the next time block. If the receiver cannot decode the packet, a retransmission request is sent through the feedback link, and another codeword block of the same packet will be sent in the next time block. For simplicity, we assume an ideal ARQ protocol in our analysis, in which the transmitter gets the feedback immediately at the end of each time block without any delay.

In this section, deadline constraint is incorporated to control the average packet delay. More specifically, the deadline constraint limits the the maximum number of successive retransmission attempts of a packet (or equivalently the number of HARQ rounds for a packet). We define the HARQ period as the duration of successive time blocks used to transmit a single packet. Then, the deadline constraint limits the

[2]Our model considers block fading with independent fading coefficients across different blocks and also a white noise process. In practical scenarios in which fading is correlated, our model assumptions will be applicable if frame-level interleaving and deinterleaving are performed at the transmitter and receiver, respectively, potentially introducing more delay compared with symbol-level interleaving. Also, if non-white noise is experienced, a whitening filter can be employed at the receiver.

maximum duration of HARQ periods. In this section, we assume that the deadline constraint is M time blocks, and the packets that cannot be received correctly by the receiver in the first HARQ period become outdated or their transmission priority is lowered. Therefore, the retransmission of a packet continues until the receiver gets the packet without error or if the limit on the number of retransmissions is reached, and then the transmitter starts with another packet in the next HARQ period. The receiver starts combining the received signal from the beginning in each HARQ period. Whether the previous packet (which has experienced deadline violation) is kept in the buffer for transmission later or is discarded from the buffer depends on the queue models described in the next subsection.

In the HARQ-CC scheme, the same coded data is transmitted in each retransmission. The receiver employs maximum-ratio combining and decodes the data packet error-free after the N^{th} round only if R satisfies [7]

$$R \leq \log_2 \left(1 + \text{SNR} \sum_{i=1}^{N} z_i \right). \tag{3.22}$$

Outage happens when a packet cannot be received correctly within one HARQ period, and we denote the value of outage probability by ε. More specifically, the outage probability is expressed as

$$\Pr \left\{ \log_2 \left(1 + \text{SNR} \sum_{i=1}^{M} z_i \right) < R \right\} = \varepsilon. \tag{3.23}$$

Although the transmitter always sends information at a fixed rate, HARQ-CC has the ability to adapt the average transmission rate to the channel conditions without requiring perfect CSI at the transmitter. According to (3.22), if the data packet is successfully decoded in the N^{th} round, the average transmission rate is bounded as $\frac{1}{N} \log_2 \left(1 + \text{SNR} \sum_{i=1}^{N-1} z_i \right) < \frac{R}{N} \leq \frac{1}{N} \log_2 \left(1 + \text{SNR} \sum_{i=1}^{N} z_i \right)$. For instance, when the channel conditions are favorable, the transmission of a single packet can be completed

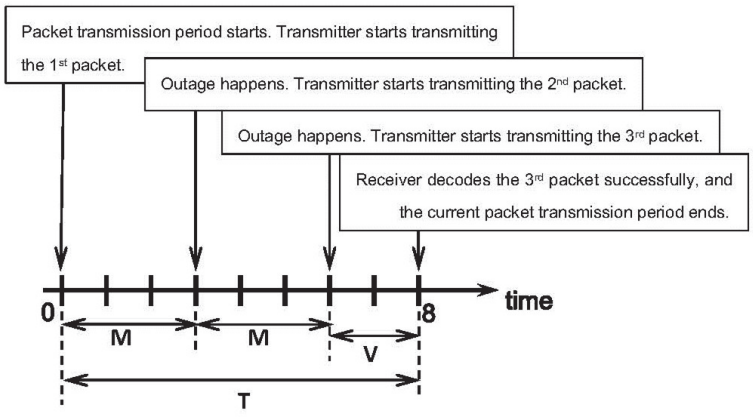

Figure 3.8: Structure of a packet transmission period in queue model I

within a few blocks, resulting in a relatively large average transmission rate, and vice versa if the channel conditions are poor. In [7], a detailed discussion about the throughput of different types of HARQ is provided.

3.2.1.3 Queue Models

As we have mentioned in Section 3.2.1.1, the transmitter is equipped with a buffer that is used to store the packets before transmission to the receiver. In this section, we consider two typical queue models.

1. **Queue Model I**: Packets are removed from the buffer only when they are received by the receiver correctly. If a packet is not received by the receiver correctly within M successive time blocks, the transmitter reduces its priority, and starts transmitting the packet with the highest priority in the next time block.

2. **Queue Model II**: Packets are removed from the buffer when they are received by the receiver correctly or if the duration of its HARQ period reaches the upper limit M.

In queue model I, there is no limitation on the overall number of time blocks used for a packet, and no packet is discarded. Instead, the packets are sorted according to

their priorities[3], and the transmitter transmits the packet with the highest priority in each HARQ period. The priority can be determined by the urgency of the data packet, and the priority level of a packet is reduced every time the deadline constraint is violated during its transmission. Once the priority of a packet is reduced, it can be transmitted again when it has the highest priority in the queue. In other words, a packet can occupy multiple HARQ periods in this model. In this case, the meaning of the deadline constraint is to control the average packet delay[4]. By reducing the priority of an outdated packet, the packets waiting behind this outdated packet can have smaller packet delay.

We define the packet transmission period as the duration of successive time blocks until one packet is removed from the buffer (either due to successful transmission, or for instance, in queue model II, due to deadline violation), and denote the packet transmission period by T.

When a successful transmission of the packet occurs, we define V as the number of time blocks used to successfully transmit the packet within the last HARQ period. In queue model I, if there are multiple HARQ periods, transmission in the last HARQ period of a packet transmission period is successful, and all other HARQ periods in the same packet period have ended up with outages. Therefore, the duration of the final HARQ period in a packet transmission period can be represented by V, and we can have the relationship $T = kM+V$, where k represents the number of failed HARQ periods or equivalently the number of outage events within the packet transmission period. Fig. 3.8 shows an example of a packet transmission period with $T = 8$, $k = 2$, $M = 3$, and $V = 2$ in queue model I. A detailed discussion about the distributions of T and V is provided in Section 3.2.2.

[3]The packets with the same priority level are sorted according to their arrival order.

[4]In this section, packet delay is defined as the duration a packet waits starting from its entrance to the buffer until its successful transmission.

Queue model I is suitable for applications, in which the receiver insists on getting every data packet it requests from the transmitter. For instance, in a cache-aided system, the receiver updates its cached data using the outdated data packets for future use [83].

In the second queue model, a packet transmission period only contains one HARQ period, and we have $T \leq M$. Hence, there is always a departure from the buffer within one HARQ period due to either successful transmission or packet drop because of deadline violation. Queue model II is suitable for applications, in which outdated data is useless, such as live video transmission.

In the buffer analysis, the instantaneous departure rate takes only two values. The departure rate is $c_i = R$ (bits/symbol) when a packet is removed from the queue in the i^{th} time block, otherwise $c_i = 0$. More specifically, in queue model I, $c_i = R$ when a packet is received by the receiver correctly; in queue model II, $c_i = R$ when a packet is received by the receiver correctly or a packet is discarded because of the deadline constraint.

For the queue model I, the throughput is characterized by the maximum average arrival rate r_{avg} that can be supported under queuing constraints, described by (2.1) in Chapter 2. For the queue model II, the throughput is given by the maximum average arrival rate r_{avg} multiplied by $(1-\varepsilon)$, because only $(1-\varepsilon)$ of the packets are received by the receiver, and ε of the packets are discarded on average due to deadline violations.

3.2.1.4 Energy Efficiency Metrics

As mentioned in the previous subsection, the system throughput is characterized by the average arrival rate r_{avg} and $(1-\varepsilon)r_{\text{avg}}$ in the queue models I and II, respectively.

Moreover, we choose energy per bit[5], defined as SNR over the throughput r_{TH}, i.e.,

$$\frac{E_b}{N_0} = \frac{\text{SNR}}{r_{TH}} = \begin{cases} \frac{\text{SNR}}{r_{\text{avg}}(\theta,\text{SNR})} & \text{in queue model I} \\ \frac{\text{SNR}}{(1-\varepsilon)r_{\text{avg}}(\theta,\text{SNR})} & \text{in queue model II} \end{cases}, \tag{3.24}$$

as the metric for energy efficiency under statistical QoS constraints. For a given throughput requirement, the system with smaller energy per bit has better energy efficiency.

In this section, we focus on the low-SNR regime, in which the throughput r_{TH} can be approximated as a linear function of the bit energy in dB scale [84]:

$$r_{TH} = \frac{S_0}{10\log_{10}2}\left(\frac{E_b}{N_{0\,\text{dB}}} - \frac{E_b}{N_{0\,\text{min,dB}}}\right) + o\left(\frac{E_b}{N_0} - \frac{E_b}{N_{0\,\text{min}}}\right), \tag{3.25}$$

where $\frac{E_b}{N_{0\,\text{dB}}} = 10\log_{10}\frac{E_b}{N_0}$, $\frac{E_b}{N_{0\,\text{min,dB}}}$ is the minimum energy per bit in dB scale, which is achieved as SNR and throughput approach 0 in our system setting, and $\frac{S_0}{10\log_{10}2}$ is the slope of the throughput curve at $\frac{E_b}{N_{0\,\text{min}}}$. Because of the linear behavior of the throughput curve in the low-SNR regime, energy efficiency can be described by the minimum energy per bit and wideband slope, and we have the following characterizations:

1. The system with smaller $\frac{E_b}{N_{0\,\text{min}}}$ has better energy efficiency for sufficiently small SNR.

2. Among the systems with the same $\frac{E_b}{N_{0\,\text{min}}}$, the system with higher S_0 value has better energy efficiency in the low-SNR regime, because higher slope provides higher throughput increment as $\frac{E_b}{N_0}$ increases, or equivalently the same throughput is achieved at a smaller value of $\frac{E_b}{N_0}$.

Therefore, the minimum energy per bit $\frac{E_b}{N_{0\,\text{min}}}$ and wideband slope S_0 are the key

[5]In the literature, $\frac{E_b}{N_0}$ is also referred to as the *signal-to-noise ratio per bit*. In this section, we denote signal-to-noise ratio per symbol as SNR, and throughput is in bits/symbol. Therefore, SNR divided by the throughput provides the SNR per bit. In (3.24), we also assume that the circuit power consumption is negligible and the transmission power is the dominant factor.

metrics of energy efficiency in the low-SNR regime.

In queue model I, the minimum energy per bit is obtained from [84]

$$\frac{E_b}{N_0}_{\min} = \lim_{\mathrm{SNR}\to 0} \frac{\mathrm{SNR}}{r_{\mathrm{avg}}(\theta, \mathrm{SNR})} = \frac{1}{\dot{r}_{\mathrm{avg}}(\theta, 0)} \tag{3.26}$$

where $\dot{r}_{\mathrm{avg}}(\theta, 0)$ denotes the first derivative of the system throughput $r_{\mathrm{avg}}(\theta, \mathrm{SNR})$ with respect to SNR at zero SNR. The wideband slope S_0 is given by

$$S_0 = \frac{-2(\dot{r}_{\mathrm{avg}}(\theta, 0))^2}{\ddot{r}_{\mathrm{avg}}(\theta, 0)} \log_e 2. \tag{3.27}$$

Above, $\ddot{r}_{\mathrm{avg}}(\theta, 0)$ denotes the second derivative of $r_{\mathrm{avg}}(\theta, \mathrm{SNR})$[6] with respect to SNR at zero SNR.

Correspondingly, the minimum energy per bit and wideband slope in queue model II are given by

$$\frac{E_b}{N_0}_{\min} = \frac{1}{(1-\varepsilon)\dot{r}_{\mathrm{avg}}(\theta, 0)} \tag{3.28}$$

and

$$S_0 = (1-\varepsilon)\frac{-2(\dot{r}_{\mathrm{avg}}(\theta, 0))^2}{\ddot{r}_{\mathrm{avg}}(\theta, 0)} \log_e 2, \tag{3.29}$$

respectively.

3.2.2 Energy Efficiency of HARQ-CC scheme with Fixed Outage Probability

In this subsection, we study the energy efficiency of HARQ-CC scheme with fixed outage probability. Initially, we consider constant-rate arrivals, characterize throughput by employing the effective capacity formulation, and derive the minimum energy

[6] In the remainder of this section, especially when θ is fixed and derivatives with respect to SNR are considered, we generally express average arrival rate and effective capacity only as a function of SNR explicitly as $r_{\mathrm{avg}}(\mathrm{SNR})$ and $C_E(\mathrm{SNR})$, respectively, and suppress θ in order to avoid cumbersome expressions.

per bit and wideband slope. Subsequently, we incorporate random arrival models by considering discrete-time Markov, Markov fluid, and Markov modulated Poisson sources and determine the system throughput and analyze the energy efficiency again by determining the minimum energy per bit and wideband slope. Based on these results, a comparison of the energy efficiency with different arrival models is given in the next subsection.

3.2.2.1 Statistical Distribution of T

Before obtaining the minimum energy per bit and wideband slope expressions for HARQ-CC, we first characterize the system throughput of HARQ-CC scheme subject to an outage constraint. Recall that an outage event happens if the receiver does not correctly decode the message within an HARQ period with a maximum duration of M time blocks. The formulation of the outage probability is given in (3.23). Correspondingly, the transmission rate that guarantees an outage probability of ϵ can be expressed as [9]

$$R = \log_2\left(1 + \mathrm{F}_M^{-1}(\varepsilon)\mathrm{SNR}\right) \tag{3.30}$$

for both queue model I and II, where F_M^{-1} is the inverse cumulative distribution function (CDF) of $\sum_{i=1}^{M} z_i$. Specifically, for Rayleigh fading, $\frac{2}{\mathbb{E}\{z\}}\sum_{i=1}^{M} z_i$ follows a chi-square distribution with $2M$ degrees of freedom; for Nakagami-m fading, $\sum_{i=1}^{M} z_i$ follows a Gamma distribution with shape parameter Mm and scale parameter $\mathbb{E}\{z\}/m$.

Hence, using the above rate expression and the formulation (3.10) in Section 3.1, we can express, for small θ, the throughput of the HARQ-CC scheme subject to an outage constraint ϵ as

$$r_{\mathrm{avg}}(\mathrm{SNR}) = \frac{\log_2(1 + \mathrm{F}_M^{-1}(\varepsilon)\ \mathrm{SNR})}{\mu} - \frac{\left[\log_2(1 + \mathrm{F}_M^{-1}(\varepsilon)\ \mathrm{SNR})\right]^2 \sigma^2\theta}{2\mu^3}, \tag{3.31}$$

for both queue model I and II. The only difference between the average arrival rates

in these two queue models lies in the expressions of μ and σ^2. In order to obtain the expressions of μ and σ^2, we first find the probability mass function (pmf) of T, which represents the duration of a packet period. Recall that in Section 3.2.1.3, we denote the duration of the last HARQ period in a packet transmission period by V, and we have characterized the relationship $T = kM + V$ in queue model I. We denote the values of the random variables T and V by t and v, respectively. In the rest of this section, we use subscript $Q1$ and $Q2$ to distinguish the notations (T, μ, σ, C_E, $\frac{E_b}{N_0\,\min}$ and S_0) for queue models I and II, respectively.

Queue model I:

The probability that the transmission of the first k packets have ended in failure due to the the deadline constraint M, and the $(k+1)^{\text{th}}$ packet is successfully transmitted after $v \leq M$ time blocks is given as follows:

$$\Pr\{T_{Q1} = kM + v\} = \varepsilon^k \Pr\{V = v\} \tag{3.32}$$

where ε is the outage probability. According to the condition given in (3.22), $\Pr\{V = v\}$ for $v \leq M$ can be expressed as

$$\Pr\{V = v\} = \Pr\{V \leq v\} - \Pr\{V \leq v-1\} \tag{3.33}$$

$$= \Pr\left\{\log_2\left(1 + \text{SNR}\sum_{i=1}^{v} z_i\right) \geqslant R\right\} - \Pr\left\{\log_2\left(1 + \text{SNR}\sum_{i=1}^{v-1} z_i\right) \geqslant R\right\} \tag{3.34}$$

$$= \Pr\left\{\sum_{i=1}^{v} z_i \geqslant \text{F}_M^{-1}(\varepsilon)\right\} - \Pr\left\{\sum_{i=1}^{v-1} z_i \geqslant \text{F}_M^{-1}(\varepsilon)\right\} \tag{3.35}$$

$$= \text{F}_{v-1}\left(\text{F}_M^{-1}(\varepsilon)\right) - \text{F}_v\left(\text{F}_M^{-1}(\varepsilon)\right) \tag{3.36}$$

where F_v is the CDF of $\sum_{i=1}^{v} z_i$. Now, (3.32) can be expressed as

$$\Pr\{T_{Q1} = kM + v\} = \varepsilon^k \left(\text{F}_{v-1}\left(\text{F}_M^{-1}(\varepsilon)\right) - \text{F}_v\left(\text{F}_M^{-1}(\varepsilon)\right)\right). \tag{3.37}$$

Queue model II:

Recall that the value of T_{Q2} cannot exceed M in queue model II. $T_{Q2} < M$ corresponds to successful transmission, and $T_{Q2} = M$ corresponds to either successful transmission using M time blocks or an outage event due to deadline violation, which leads to packet being removed from the buffer and discarded. Therefore, we can express the pmf of T_{Q2} as

$$\Pr\{T_{Q2} = t\} = \begin{cases} \Pr\{V = t\} = \mathrm{F}_{t-1}\left(\mathrm{F}_M^{-1}(\varepsilon)\right) - \mathrm{F}_t\left(\mathrm{F}_M^{-1}(\varepsilon)\right), & t < M \\ \Pr\{V = M\} + \varepsilon = \mathrm{F}_{M-1}\left(\mathrm{F}_M^{-1}(\varepsilon)\right), & t = M \end{cases} \tag{3.38}$$

where V has the same pmf as in queue model I. Recall that V is defined only for successful transmission, and thus $V = M$ in (3.38) represents that the packet is received successfully using M time blocks.

Theorem 2 *For queue model I, the expected value and variance of T are given by*

$$\mu_{Q1} = \frac{1}{1-\varepsilon} \sum_{v=1}^{M} v \ \Pr\{V = v\} + \frac{M\varepsilon}{1-\varepsilon} \tag{3.39}$$

$$\sigma_{Q1}^2 = \frac{1}{1-\varepsilon} \sum_{v=1}^{M} v^2 \ \Pr\{V = v\} + \frac{2M\varepsilon}{(1-\varepsilon)^2} \sum_{v=1}^{M} v \ \Pr\{V = v\} + \frac{M^2\varepsilon(1+\varepsilon)}{(1-\varepsilon)^2} - \mu_{Q1}^2, \tag{3.40}$$

respectively. And for queue model II, the expected value and variance of T are given by

$$\mu_{Q2} = \sum_{t=1}^{M} t \ \Pr\{V = t\} + M\varepsilon, \tag{3.41}$$

$$\sigma_{Q2}^2 = \sum_{t=1}^{M} t^2 \ \Pr\{V = t\} + M^2\varepsilon - \mu_{Q2}^2, \tag{3.42}$$

respectively. In the above expressions, the pmf of random variable V is given by (3.36)

for both queue models I and II.

Proof: See Appendix A.2.

3.2.2.2 Energy Efficiency of HARQ-CC with Constant-Rate Arrivals

Note that the expressions of μ and σ^2 do not depend on SNR in both queue models I and II. In the following result, we characterize the energy efficiency in the low SNR regime for small θ.

Theorem 3 *For small QoS exponent θ, the minimum energy per bit and wideband slope of the HARQ-CC scheme with the outage constraint ϵ are given, respectively, by*

$$\frac{E_b}{N_{0\,\min\,\mathrm{Q1}}} = \frac{\mu_{Q1}\ \log_e 2}{F_M^{-1}(\varepsilon)}, \tag{3.43}$$

$$S_{0\,Q1} = \frac{2\mu_{Q1}\log_e 2}{\sigma_{Q1}^2\theta + \mu_{Q1}^2\log_e 2}, \tag{3.44}$$

for queue model I, where μ_{Q1} and σ_{Q1}^2 are given by (3.39) and (3.40), respectively. For queue model II, the minimum energy per bit and wideband slope are given, respectively, by

$$\frac{E_b}{N_{0\,\min\,\mathrm{Q2}}} = \frac{\mu_{Q2}\ \log_e 2}{(1-\varepsilon)F_M^{-1}(\varepsilon)}, \tag{3.45}$$

$$S_{0\,Q2} = (1-\varepsilon)\frac{2\mu_{Q2}\log_e 2}{\sigma_{Q2}^2\theta + \mu_{Q2}^2\log_e 2}, \tag{3.46}$$

respectively, where μ_{Q2} and σ_{Q2}^2 are given by (3.41) and (3.42).

Proof: See Appendix A.3.

We immediately notice that for both queue models I and II, the minimum energy per bit $\frac{E_b}{N_{0\,\min}}$ does not depend on the QoS exponent θ, and hence is not affected by the presence of QoS constraints. On the other hand, via μ and $\mathrm{F}_M^{-1}(\varepsilon)$, $\frac{E_b}{N_{0\,\min}}$ is a function of the deadline constraint M and the outage limit ϵ. This dependence will

be explored in the numerical results. We further notice that the wideband slope S_0 diminishes with increasing θ. Hence, stricter QoS constraints lead to smaller slopes, increasing the energy per bit requirements at the same throughput level.

Proposition 1 *For the same SNR, θ and channel fading, queue models I and II lead to the same minimum energy per bit. On the other hand, the system operating with queue model II achieves a higher wideband slope.*

Proof: See Appendix A.4.

A detailed discussion of the characterization in Proposition 1 is provided in the numerical results.

3.2.2.3 Energy Efficiency of HARQ-CC with ON-OFF Discrete-Time Markov Source

As mentioned in Section 2.2.2, when the arrival rate a_i is not constant, the computation of the throughput is more involved. Generally, we need to express the LMGFs of the random arrival processes and random departure processes (or equivalently random wireless transmissions), and then solve (2.4) in order to determine the maximum average arrival rate r_{avg} that can be supported by the wireless transmissions under statistical queuing constraints. In these cases, derivation of the minimum bit energy and wideband slope only involves the first and second order derivatives of r_{avg} evaluated at $\text{SNR} = 0$, which can be obtained easily by taking the derivatives of both sides of (2.4) and letting $\text{SNR} \to 0$. In this subsection, we analyze the energy efficiency of HARQ-CC with fixed outage probability when we have ON-OFF discrete-time Markov sources.

A detailed study about the throughput of the ON-OFF discrete-time Markov source under statistical queuing constraints is provided in Section 2.2.2.1. Since the departure and arrival processes at the transmitter are independent, for both queue

models I and II, the expressions of μ and σ^2 in (3.39), (3.40), (3.41) and (3.42) are still valid.

Theorem 4 *For small QoS exponent θ and ON-OFF discrete-time Markov source, the minimum energy per bit and wideband slope of the HARQ-CC scheme with the outage constraint ϵ are given, respectively, by*

$$\frac{E_b}{N_0}_{\min \mathrm{Q1}} = \frac{\mu_{Q1} \log_e 2}{F_M^{-1}(\varepsilon)}, \tag{3.47}$$

$$S_{0\,Q1} = \frac{2\log_e 2}{\frac{\sigma_{Q1}^2\theta+\mu_{Q1}^2\log_e 2}{\mu_{Q1}} + \theta\zeta} \tag{3.48}$$

for queue model I, where μ_{Q1} and σ_{Q1}^2 are given by (3.39) and (3.40), respectively, and ζ is defined as

$$\zeta = \frac{(1-p_{22})(p_{11}+p_{22})}{(1-p_{11})(2-p_{11}-p_{22})}. \tag{3.49}$$

For queue model II, the minimum energy per bit and wideband slope are given by

$$\frac{E_b}{N_0}_{\min \mathrm{Q2}} = \frac{\mu_{Q2} \log_e 2}{(1-\varepsilon)F_M^{-1}(\varepsilon)}, \tag{3.50}$$

$$S_{0\,Q2} = \frac{2(1-\varepsilon)\log_e 2}{\frac{\sigma_{Q2}^2\theta+\mu_{Q2}^2\log_e 2}{\mu_{Q2}} + \theta\zeta}, \tag{3.51}$$

respectively, where μ_{Q2} and σ_{Q2}^2 are given by (3.41) and (3.42).

Proof: See Appendix A.5.

3.2.2.4 Energy Efficiency of HARQ-CC with ON-OFF Fluid Markov Source

In this subsection, we consider the ON-OFF fluid Markov sources, which is studied in detail in Section 2.2.2.2. Using a similar approach as for the discrete-time Markov source, we can find the minimum energy per bit and wideband slope for the ON-OFF fluid Markov source as in the following result.

Theorem 5 *For small QoS exponent θ and ON-OFF fluid Markov source, the minimum energy per bit and wideband slope of the HARQ-CC scheme with the outage constraint ϵ are given, respectively, by*

$$\frac{E_b}{N_{0\,\min\,\mathrm{Q1}}} = \frac{\mu_{Q1}\,\log_e 2}{F_M^{-1}(\varepsilon)}, \tag{3.52}$$

$$S_{0\,Q1} = \frac{2\log_e 2}{\frac{\sigma_{Q1}^2\theta+\mu_{Q1}^2\log_e 2}{\mu_{Q1}} + \frac{2\theta\beta}{\alpha(\alpha+\beta)}} \tag{3.53}$$

for queue model I, where μ_{Q1} and σ_{Q1}^2 are given by (3.39) and (3.40), respectively. For queue model II, the minimum energy per bit and wideband slope are given, respectively, by

$$\frac{E_b}{N_{0\,\min\,\mathrm{Q2}}} = \frac{\mu_{Q2}\,\log_e 2}{(1-\varepsilon)F_M^{-1}(\varepsilon)}, \tag{3.54}$$

$$S_{0\,Q2} = \frac{2(1-\varepsilon)\log_e 2}{\frac{\sigma_{Q2}^2\theta+\mu_{Q2}^2\log_e 2}{\mu_{Q2}} + \frac{2\theta\beta}{\alpha(\alpha+\beta)}} \tag{3.55}$$

respectively, where μ_{Q2} and σ_{Q2}^2 are given by (3.41) and (3.42).

Proof: See Appendix A.6.

3.2.2.5 Energy Efficiency of HARQ-CC with ON-OFF MMPS

In this subsection, we investigate the energy efficiency of ON-OFF MMPS models. The throughput of MMPS is investigated in Section 2.2.2.3, and the following result identifies the the minimum energy per bit and wideband slope for the ON-OFF MMPS models.

Theorem 6 *For small QoS exponent θ and ON-OFF MMPS, the minimum energy per bit and wideband slope of the HARQ-CC scheme with the outage constraint ϵ are*

given, respectively, by

$$\frac{E_b}{N_{0\,\min\,\mathrm{Q1}}} = \frac{e^\theta - 1}{\theta}\,\frac{\mu_{Q1}\,\log_e 2}{F_M^{-1}(\varepsilon)}, \tag{3.56}$$

$$S_{0\,Q1} = \frac{\theta}{e^\theta - 1}\,\frac{2\log_e 2}{\frac{\sigma_{Q1}^2\theta + \mu_{Q1}^2 \log_e 2}{\mu_{Q1}} + \frac{2\theta\beta}{\alpha(\alpha+\beta)}} \tag{3.57}$$

for queue model I, where μ and σ^2 are given by (3.39) and (A.37), respectively. For queue model II, the minimum energy per bit and wideband slope are given, respectively, by

$$\frac{E_b}{N_{0\,\min\,\mathrm{Q2}}} = \frac{e^\theta - 1}{\theta}\,\frac{\mu_{Q2}\,\log_e 2}{(1-\varepsilon)F_M^{-1}(\varepsilon)}, \tag{3.58}$$

$$S_{0\,Q2} = \frac{\theta}{e^\theta - 1}\,\frac{2(1-\varepsilon)\log_e 2}{\frac{\sigma_{Q2}^2\theta + \mu_{Q2}^2 \log_e 2}{\mu_{Q2}} + \frac{2\theta\beta}{\alpha(\alpha+\beta)}} \tag{3.59}$$

respectively, where μ_{Q2} and σ_{Q2}^2 are given by (3.41) and (3.42).

Proof: See Appendix A.7.

A comparison of the results in Theorems 3–6 is provided in the following subsection.

3.2.3 Comparison of the Energy Efficiency for Different Arrival Models

In this subsection, we compare the results obtained in the previous subsection for constant-rate arrivals, ON-OFF discrete-time and fluid Markov sources, and ON-OFF MMPS. In the first part below, we compare the results between constant-rate arrivals and ON-OFF discrete-time Markov sources. In the second part, we provide a comparison among constant-rate arrivals, ON-OFF fluid Markov sources and MMPS. Our analysis shows that source burstiness makes it difficult to satisfy the queuing constraint, which leads to degraded energy efficiency. The key parameters that have

significant impact on the energy efficiency of these random arrival models are the QoS exponent θ, ON state probability P_{ON} and the state transition parameters p_{21} and β.

3.2.3.1 Comparison Between Constant Arrival and ON-OFF Discrete-Time Markov Source

The results on the minimum energy per bit and wideband slope for constant-rate arrivals and ON-OFF discrete-time Markov sources are given in Theorems 3 and 4, respectively. Since the results for queue model II are very similar to the results for queue model I with the only difference being the additional factor $(1-\varepsilon)$, the discussion in this subsection is applicable to both queue models.

We first observe that source randomness does not have any influence on the minimum energy per bit, and the results of minimum energy per bit shown in Theorem 4 are the same as in the case of constant-rate arrivals. On the other hand, source burstiness has an impact on the wideband slope. Compared with the case of constant-rate arrivals, there is an additional term $\theta\zeta$ in the denominator, and this additional term is only related to the arrival process. Since both of p_{11} and p_{22} are between 0 and 1, it is easy to verify that $\theta\zeta \geq 0$, which means that random arrivals always degrade the wideband slope and make the system less energy-efficient compared with constant-rate arrivals.

For the ON-OFF discrete-time Markov source, source burstiness is described by P_{ON} and p_{21}. P_{ON} represents the probability that the source is in ON state, and p_{21} denotes the probability that the source transitions from ON state to the OFF state. When $P_{ON} = 1$, ON-OFF discrete-time Markov source becomes a constant-rate source, and $p_{11} = 0$, $p_{22} = 1$. Under this situation, we have $\zeta = 0$, and the results in Theorem 4 specialize to those obtained in the case of the constant-rate arrivals.

For any P_{ON} in the open interval $(0,1)$, we can rewrite the expression of ζ as

$$\zeta = \frac{(1-P_{ON})(1-p_{21}-2P_{ON})}{p_{21}P_{ON}}, \tag{3.60}$$

by applying the facts $p_{22} = 1 - p_{21}$, $p_{12} = \frac{p_{21}P_{ON}}{1-P_{ON}}$ and $p_{11} = 1 - p_{12} = \frac{1-(1+p_{21})P_{ON}}{1-P_{ON}}$ to (3.49). It can be easily verified that ζ is an decreasing function of both P_{ON} and p_{21}, which means that higher P_{ON} and p_{21} values improve the energy efficiency by increasing the wideband slope.

Also, we notice that when $\theta = 0$, the additional term is zero, and the parameters of the arrival process do not have any influence on the energy efficiency. When θ becomes larger, the influence of source burstiness becomes more significant. An intuitive description for this is provided in the numerical results subsection.

3.2.3.2 Comparison Among Constant Arrivals, ON-OFF Fluid Markov Source and MMPS

The results on the minimum energy per bit and wideband slope for constant-rate arrivals, ON-OFF fluid Markov sources and MMPS are given in Theorems 3, 5 and 6, respectively. Similarly as in the previous subsection, our following remarks are applicable to both queue models.

From the comparison between Theorems 3 and 5, we notice that burstiness/randomness of the ON-OFF fluid Markov sources does not affect the minimum energy per bit, and it only results in the addition of the positive term $\frac{2\theta\beta}{\alpha(\alpha+\beta)}$ in the denominator of the wideband slope expressions in (3.53) and (3.55). Therefore, constant-rate arrival sources have better energy efficiency, compared with ON-OFF fluid Markov sources. Similar to ON-OFF discrete-time Markov sources, the burstiness of the ON-OFF fluid Markov sources is described by P_{ON} and β. When $P_{ON} = 1$, arrival rates become constant, and this additional term vanishes. For any P_{ON} in the open interval $(0,1)$,

we can rewrite the expression of the additional term as

$$\frac{2\theta\beta}{\alpha(\alpha+\beta)} = 2\theta\frac{(1-P_{ON})^2}{P_{ON}\beta}, \tag{3.61}$$

by applying $\alpha = \frac{P_{ON}\beta}{1-P_{ON}}$. It can be easily verified that this additional term is an decreasing function of both P_{ON} and β, which means that higher P_{ON} and β values improve the energy efficiency by increasing the wideband slope.

As in the case of ON-OFF discrete-time Markov sources, we notice that as θ increases, the effect of source burstiness becomes more pronounced, while the parameters of the arrival process do not have any influence on the energy efficiency when $\theta = 0$.

When comparing the results of Theorems 5 and 6, we assume that these two kinds of Markovian sources share the same α and β values. From the comparison, we notice that Poisson arrival model only leads to an additional factor of $\frac{\theta}{e^{\theta}-1}$ in the expressions of the minimum energy per bit and wideband slope. Therefore, β has the same impact as in the case of ON-OFF fluid Markov sources. For $\theta \geq 0$, we have $\frac{\theta}{e^{\theta}-1} \leq 1$, resulting in a larger minimum energy per bit and smaller wideband slope for the ON-OFF MMPS compared to those for the ON-OFF Markov fluid source. Since the factor $\frac{\theta}{e^{\theta}-1}$ is a decreasing function of θ, the performance gap grows further as the queuing constraint gets stricter. Moreover, as a stark contrast to the observations in Sections 3.2.2.3 and 3.2.2.4, the minimum energy per bit depends on θ when ON-OFF MMPS arrival model is considered.

Therefore, we can conclude that among these three arrival models, highest energy efficiency is achieved in the case of constant-rate arrivals while ON-OFF MMPS leads to the worst levels of energy efficiency.

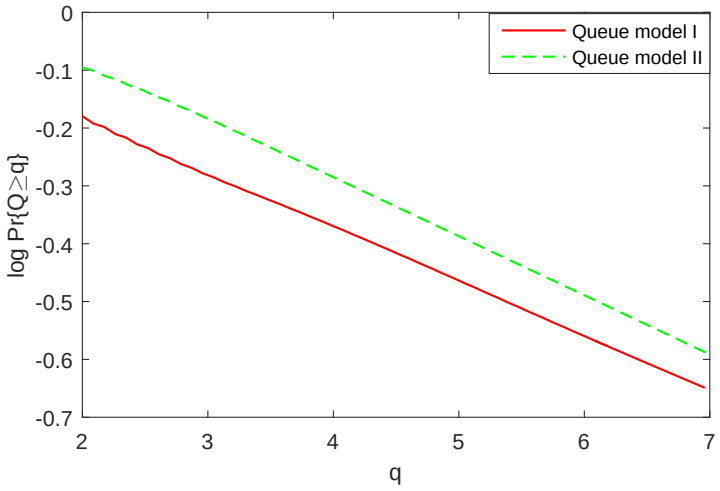

Figure 3.9: Logarithmic buffer overflow probability vs. buffer overflow threshold.

3.2.4 Numerical Results

In this subsection, we present numerical results to illustrate the energy efficiency of HARQ-CC in the presence of QoS constraints. In the first part, numerical results for the constant-rate arrival model are provided to demonstrate the influence of the deadline constraint M and outage probability ε. In the second part, we concentrate on the impact of random arrivals and source burstiness. Via Monte Carlo simulation, we verify the analytical results provided in our theorems. A comparison between queue models I and II is also provided in the first subsection, verifying Proposition 1.

3.2.4.1 Constant-Rate Arrival Models

In this part, we analyze the energy efficiency of the HARQ-CC scheme with fixed transmission rate and constant arrival rate, and we assume Rayleigh fading channel with exponentially distributed fading power having a mean value of $\mathbb{E}\{z\} = 1$ within this subsection.

First, we have performed Monte Carlo simulations to verify that the constant arrival rate, or equivalently the effective capacity, given by (3.10) in Section 3.1 satisfies

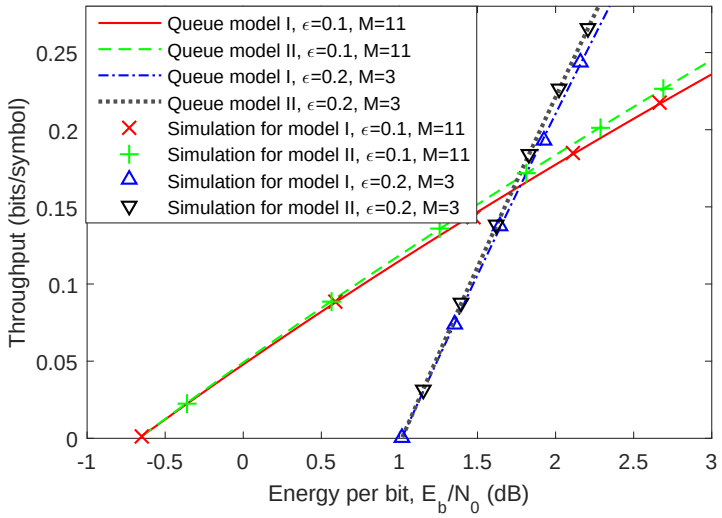

Figure 3.10: Throughput vs. energy per bit $\frac{E_b}{N_0}$.

the statistical queuing constraint in both queue models I and II[7]. In Fig. 3.9, we set the queuing constraint as $\theta = 0.1$, choose the outage probability as $\varepsilon = 0.1$, and we plot the logarithmic buffer overflow probabilities $\log \Pr\{Q \geq q\}$ as functions of the buffer overflow threshold q for both queue models I and II. For each queue model, we have repeated the simulations 100 times, in each of which the simulation is conducted over 1×10^7 time blocks. In the simulation, the values of μ_{Q1} and σ_{Q1}^2 are computed using (3.39) and (3.40), respectively, and μ_{Q2} and σ_{Q2}^2 are computed using (3.41) and (3.42), respectively. From Fig. 3.9, we observe that the logarithmic buffer overflow probabilities decrease linearly even starting from relatively small q values, agreeing with the characterizations in (2.1) and (2.2) [8]. We have estimated the slopes via linear regression, and the estimated slopes for queue models I and II are -0.096 and

[7]In queue model I, when a deadline violation occurs and the packet is not successfully sent within an HARQ period of M time blocks, we keep the packet in the buffer, clear the accumulated signal at the receiver, and initiate the transmission anew in the next HARQ period. Under these assumptions, since the sole goal of the simulations is to keep track of the queue length, there is no distinction regarding whether a new packet is transmitted in the next HARQ period or the same packet is repeated. On the other hand, in queue model II, we discard the outdated packet from the buffer after violating the deadline constraint, and start transmitting a new packet.

[8]Note that if (2.1) holds, then $\log(\Pr\{Q \geq q\}) \approx -\theta q + \log \varsigma$ and hence the logarithmic overflow probability decays linearly with slope $-\theta$ as threshold q increases.

-0.103 respectively, which are indeed very close to the desired value -0.1.

In Fig. 3.10, we plot the throughput, which is $C_E(\theta, \mathrm{SNR})$ for queue model I, and $(1-\varepsilon)C_E(\theta, \mathrm{SNR})$ for queue model II, as a function of the energy per bit $\frac{E_b}{N_0}$, under two different outage constraints ϵ and deadline constraints M. The results in Fig. 3.10 are also validated via Monte Carlo simulations using (2.8). To determine the LMGF of the departure process, we have conducted simulations over 1×10^4 time blocks and repeated this for 1×10^4 times. We notice that analytical and simulation results agree perfectly for both queue models I and II. Note that the throughput for both queue models cannot exceed $\mathbb{E}\{\log_2(1+\mathrm{SNR}z)\}$, which is the Shannon capacity achieved in the absence of queuing constraints. Since this throughput upper bound is an increasing concave function of SNR, the minimum energy per bit is achieved as SNR approaches 0, i.e., $\lim_{\mathrm{SNR}\to 0} \frac{\mathrm{SNR}}{\mathbb{E}\{\log_2(1+\mathrm{SNR}z)\}} = \frac{\log_e 2}{\mathbb{E}\{z\}}$. Since we set $\mathbb{E}\{z\} = 1$, the minimum energy per bit cannot be smaller than $\log_e 2$, which is equal to -1.59 dB.

Comparing the curves of these two queue models, we find that queue model II has better energy efficiency. According to our results in Proposition 1, queue model I and II should achieve the same minimum energy per bit, while queue model II achieves a higher wideband slope in the constant-rate arrival model. In order to verify Proposition 1, we have computed the $\frac{E_b}{N_0}_{\min}$ and S_0 for both queue models I and II in Figs. 3.11 and 3.12.

All $\frac{E_b}{N_0}_{\min}$ and S_0 values in Figs. 3.11 and 3.12 are verified via simulations. For each pair of $\frac{E_b}{N_0}_{\min}$ and S_0, we have obtained 50 points of the throughput curves (throughput vs. $\frac{E_b}{N_0}_{\mathrm{dB}}$) in the low-SNR regime ($\mathrm{SNR} \leq -33$ dB) from simulations, and estimated $\frac{E_b}{N_0}_{\min}$ and S_0 via linear regression according to (3.25). The maximum errors of $\frac{E_b}{N_0}_{\min}$ and S_0 are 0.0007% and 0.3%, respectively. From Figs. 3.11 and 3.12, we observe that queue model I and II have the same minimum energy per bit, and the wideband slopes of queue model II are slightly greater than the wideband slopes of queue model I. This observation agrees with Proposition 1, and an intuitive

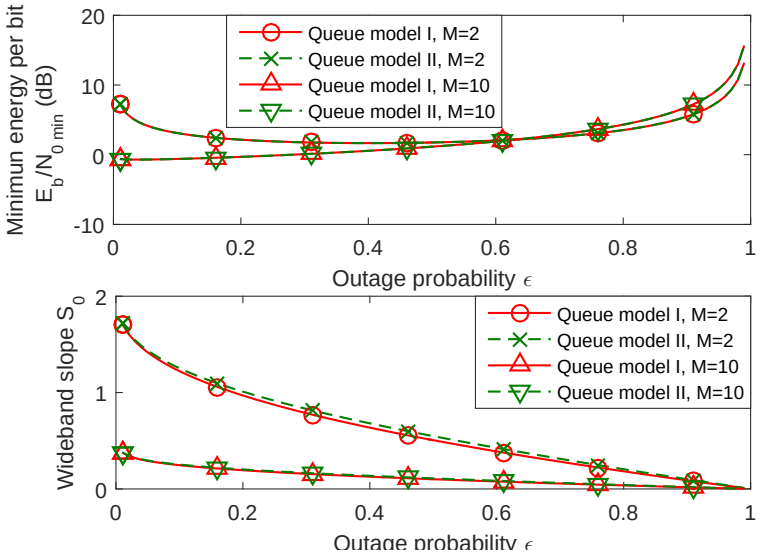

Figure 3.11: Minimum energy per bit $\frac{E_b}{N_{0\,\min}}$ and wideband slope S_0 vs. outage probability ϵ.

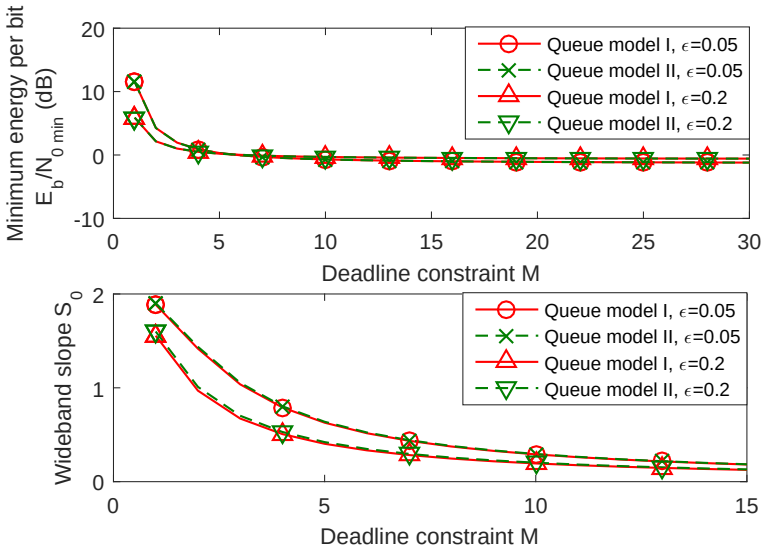

Figure 3.12: Minimum energy per bit $\frac{E_b}{N_{0\,\min}}$ and wideband slope S_0 vs. deadline constraint M.

explanation is given as follows for the minimum energy per bit.

The throughput of queue model I and II are given by $r_{TH\,Q1} = R/\mu_{Q1}$ and $r_{TH\,Q2} = (1-\varepsilon)R/\mu_{Q2}$, respectively, when $\theta = 0$. From (A.46) in Appendix A.4, we have $r_{TH\,Q1} = r_{TH\,Q2}$, which means that the throughput curves of these two queue models are exactly the same. This implies that $\frac{E_b}{N_0\,\min\,\mathrm{Q1}} = \frac{E_b}{N_0\,\min\,\mathrm{Q2}}$ and $S_{0\,Q1} = S_{0\,Q2}$, when $\theta = 0$. Since minimum energy per bit does not depend on the queuing constraints, $\frac{E_b}{N_0\,\min\,\mathrm{Q1}} = \frac{E_b}{N_0\,\min\,\mathrm{Q2}}$ is valid for any θ value.

In Fig. 3.11, we display the minimum energy per bit $\frac{E_b}{N_0\,\min}$ and wideband slope S_0 as functions of the outage probability constraint ϵ for two different values of the deadline constraint M. It is observed from the figure that the minimum energy per bit first decreases with increasing ϵ and then starts increasing after a certain threshold point. When the outage probability is small, the fixed transmission rate R is small, which leads to small departure rates for both queue models I and II. On the other hand, when the outage probability is large, the average transmission rate is small for both queue models I and II, because the transmitter wastes a whole HARQ period when an outage happens. Also, we observe that the wideband slope always decreases with increasing ϵ.

In Fig. 3.12, the minimum energy per bit and wideband slope are plotted as functions of the deadline constraint M for both queue models I and II. It is seen that both the minimum energy per bit and wideband slope decrease with increasing M. Hence, by reducing the minimum energy per bit, relaxed deadline constraints lead to improvements in energy efficiency in the vicinity of $\frac{E_b}{N_0\,\min}$.

3.2.4.2 Random Arrival Models

In this part, we investigate the impact of source randomness/burstiness on the energy efficiency. Within this subsection, we assume a Nakagami-m fading channel with $m = 2$, and $\mathbb{E}\{z\} = 1$. Unless mentioned explicitly, QoS exponent is set to $\theta = 0.1$.

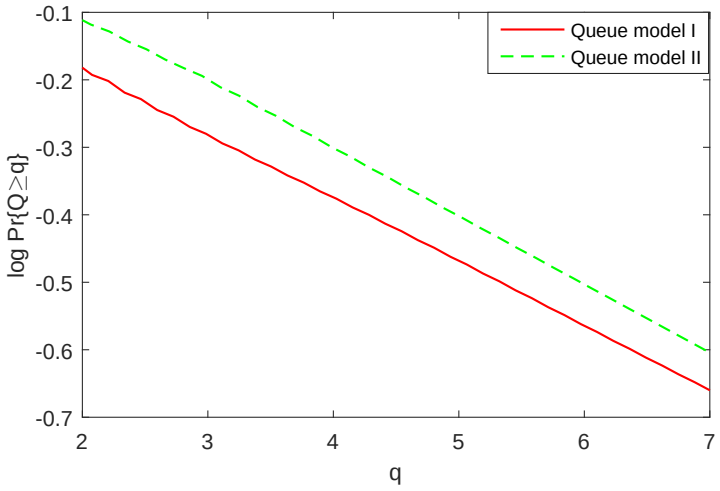

Figure 3.13: Logarithmic buffer overflow probability vs. buffer overflow threshold for the ON-OFF discrete-time Markov source.

For all fixed outage probability results, we fix $\varepsilon = 0.1$. In Fig. 3.13, we provide the results of buffer simulations for the ON-OFF discrete-time Markov source, similarly as depicted in Fig. 3.9. The arrival rates in the ON state are given by (2.13) in Section 2.2.2.1 for both queue models. We set $p_{11} = 0.4$, $p_{22} = 0.7$, and $\theta = 0.1$ for both queue models. All other parameters are the same as in Fig. 3.9. The estimated slopes of queue models I and II are -0.096 and -0.102, respectively, which are again very close to the desired value -0.1. As we have mentioned in Section 3.2.3, since source burstiness has similar impacts on queue models I and II, we only consider queue model I in the following discussion on the impacts of source burstiness.

Figs. 3.14 and 3.15 demonstrate the influence of source burstiness considering both ON-OFF discrete-time Markov and Markov fluid sources for queue model I. For the ON-OFF discrete-time Markov source, the source burstiness is described by P_{ON} and p_{21}, and the source burstiness is described by P_{ON} and β for the ON-OFF Markov fluid source. As discussed in Section 3.2.3, larger P_{ON}, p_{21} and β values improve the energy efficiency for both queue models I and II. In Fig. 3.14, when fix $p_{21} = 0.3$, the slopes of the throughput curves increase as P_{ON} increases from 0.1 to 0.75. Also,

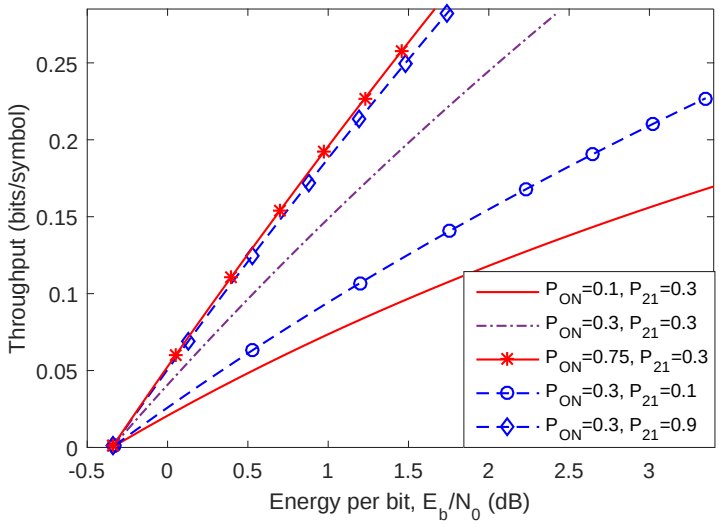

Figure 3.14: Throughput vs. energy per bit $\frac{E_b}{N_0}$ for ON-OFF discrete-time Markov source with fixed outage probability $\varepsilon = 0.1$.

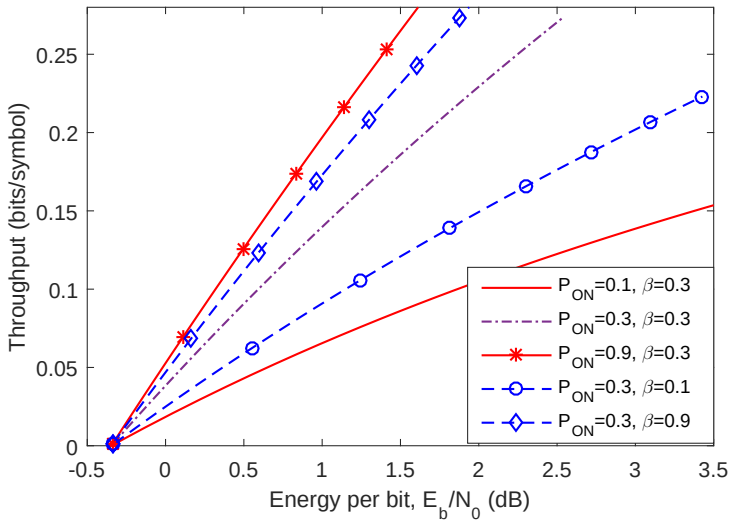

Figure 3.15: Throughput vs. energy per bit $\frac{E_b}{N_0}$ for ON-OFF Markov fluid source with fixed outage probability $\varepsilon = 0.1$.

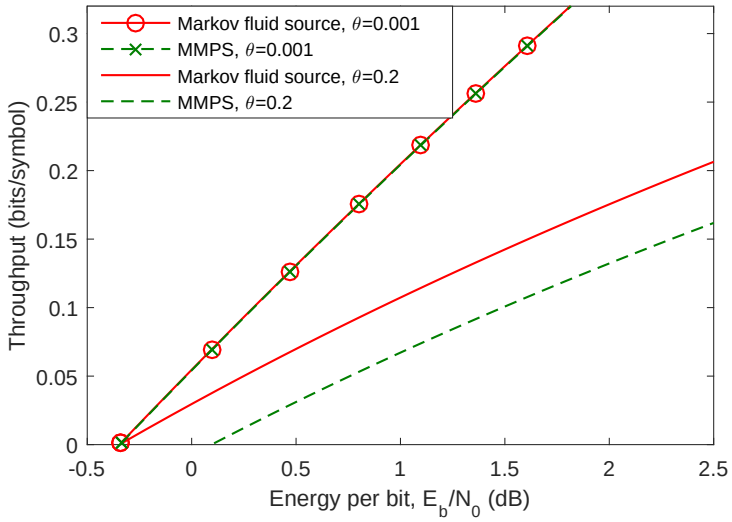

Figure 3.16: Throughput vs. energy per bit $\frac{E_b}{N_0}$ for ON-OFF Markov fluid source and MMPS with fixed outage probability $\varepsilon = 0.1$

when $P_{ON} = 0.3$ is fixed, the wideband slope increases as p_{21} increases from 0.1 to 0.9. Since source burstiness does not affect the minimum energy per bit for both ON-OFF discrete-time Markov and Markov fluid sources, we observe that the throughput curves in Fig. 3.14 converge to the same minimum energy per bit. Similarly in Fig. 3.15, for the ON-OFF Markov fluid source, we can observe that larger P_{ON} and β values increase the slope of the throughput curve, and all throughput curves in Fig. 3.15 again approach the same minimum energy per bit.

When the average arrival rate is fixed, the arrival rate in the ON state increases as P_{ON} decreases because $r_{\text{avg}} = rP_{ON}$, and large arrival rates make it difficult to satisfy the queuing constraint. Hence, larger P_{ON} improves the energy efficiency for both ON-OFF discrete-time Markov and Markov fluid sources. When P_{ON} is fixed, higher p_{21} and β values make the source transition between two states more frequently. For the same P_{ON}, more frequent state transitions make the queuing constraints to be satisfied more easily, because the change from ON state to OFF state gives the source the chance to clear its buffer. As we have mentioned in Section

3.2.3 that as θ increases, the influence of source burstiness becomes more significant, and the parameters of the arrival process do not have any influence on the energy efficiency when $\theta = 0$, for both ON-OFF discrete-time Markov and Markov fluid sources. Therefore, larger P_{ON}, p_{21} and β values improve the energy efficiency by helping the system to satisfy queuing constraints more effectively, and this impact becomes more striking when the queuing constraints become stricter. If the system is not restricted by the queuing constraint, then the source burstiness does not affect the energy efficiency for the ON-OFF discrete-time Markov and Markov fluid sources.

Finally, in Fig. 3.16, we compare the performances of ON-OFF Markov fluid source and MMPS for queue model I. As mentioned in Section 3.2.3, compared to the the minimum energy per bit and wideband slope of the ON-OFF Markov fluid source, the corresponding results for MMPS are scaled by the factor $\frac{e^{\theta}-1}{\theta}$ and its reciprocal, respectively. When θ is close to 0, both $\frac{e^{\theta}-1}{\theta}$ and its reciprocal approach 1. For this reason, the throughput curves of ON-OFF Markov fluid source and MMPS stay very close to each other in both figures when $\theta = 0.001$. As θ increases, the factor $\frac{e^{\theta}-1}{\theta}$ grows, which leads to larger gap between the throughput curves of these two types of Markov sources. For instance, we can easily observe from Fig. 3.16 that there is a 0.44 dB difference between the corresponding minimum energy per bit values when $\theta = 0.2$.

Chapter 4

Throughput of Hybrid-ARQ under Statistical Queuing Constraints Using Recurrence Approach

In this chapter, throughput of HARQ under statistical queuing constraints is studied via recurrence approach. Compared with the low-θ approximation used in Chapter 3, recurrence approach is more accurate for any QoS exponent value. However, it is difficult to obtain closed-form expression via recurrence approach. Therefore, we cannot conduct a similar energy efficiency analysis as we have done in Section 3.2.

In Section 4.1, throughput of HARQ-CC schemes is studied in the presence of Markovian data arrivals and statistical queuing constraints. In particular, two queuing models are considered. Specifically, when outage occurs, the transmitter keeps the packet, lowers its priority, and attempts to retransmit it later in the first queue model while the packet is discarded and removed from the buffer in the second queue model. The throughput is investigated when outage constraints, statistical queuing constraints and deadline constraints are imposed. The deadline constraint provides a limitation on the number of retransmissions. Under these assumptions, throughput

characterizations are obtained for HARQ-CC scheme with three types of Markovian sources, namely the ON-OFF discrete-time and fluid Markov sources and MMPS.

In Section 4.2, throughput of HARQ-IR schemes with finite blocklength codes is studied for both constant-rate and ON-OFF discrete-time Markov arrivals under statistical queuing constraints and deadline limits. After analyzing the decoding error probability and outage probability, the distribution of transmission period is characterized, and the throughput expressions are obtained for both arrival models. The analytical results are verified via Monte Carlo simulations.

4.1 Throughput of Hybrid-ARQ Chase Combining with ON-OFF Markov Arrivals under Statistical Queuing Constraints

In this section, we study the throughput of HARQ-CC protocol under both statistical queuing constraints and deadline constraints. The system model is the same with in Section 3.2, and the discussions about HARQ-CC scheme, deadline constraints, outage probability, two queue models and the throughput metrics in Section 3.2.1 are valid in this section as well. Different from the analysis in Section 3.2, we characterize the throughput via a recurrence approach, which provides sufficiently accurate results for any QoS exponent value.

4.1.1 Throughput of HARQ-CC Scheme with Queuing Constraints

In this subsection, we study the throughput of HARQ-CC scheme under queuing constraints. Initially, we consider constant-rate arrivals, characterize throughput by employing the effective capacity formulation. Subsequently, we incorporate random

arrival models by considering three types of Markovian sources and determine the system throughput using the results obtained in the constant-rate arrival model.

4.1.1.1 Throughput of HARQ-CC Scheme with Constant-rate Arrivals

Recall that an outage event happens if the receiver does not correctly decode the message within an HARQ period with a maximum duration of M time blocks. The formulation of the outage probability is given in (3.23). Correspondingly, the transmission rate that guarantees an outage probability of ϵ can be expressed as [9]

$$R = \log_2\left(1 + \mathrm{F}_M^{-1}(\varepsilon)\mathrm{SNR}\right) \tag{4.1}$$

for both queue models I and II, where F_M^{-1} is the inverse cumulative distribution function (CDF) of $\sum_{i=1}^{M} z_i$.

In this section, we define P_{t,v,Q_j} as the probability that the duration of an HARQ period is t, and the number of packets removed from the queue in this HARQ period is v, for queue model j. From the discussion in Section 3.2.1.2, we have $1 \leq t \leq M$, and v only takes two values, 1 or 0.

In queue model I, $v = 0$ when outage occurs. In such cases, we have $t = M$, because outage happens only after the transmitter's unsuccessful M transmission attempts. $P_{t,1,Q_1}$ is the probability that a transmission period ends up with success in the t^{th} time block. From the above discussion, we have

$$P_{t,0,Q_1} = \begin{cases} 0 & t < M \\ \varepsilon & t = M \end{cases} \tag{4.2}$$

and

$$P_{t,1,Q_1}=\Pr\left\{\log_2\left(1+\mathrm{SNR}\sum_{i=1}^{t}z_i\right)\geqslant R\right\}-\Pr\left\{\log_2\left(1+\mathrm{SNR}\sum_{i=1}^{t-1}z_i\right)\geqslant R\right\} \tag{4.3}$$

$$=\Pr\left\{\sum_{i=1}^{t}z_i\geqslant \mathrm{F}_M^{-1}(\varepsilon)\right\}-\Pr\left\{\sum_{i=1}^{t-1}z_i\geqslant \mathrm{F}_M^{-1}(\varepsilon)\right\} \tag{4.4}$$

$$=\mathrm{F}_{t-1}\left(\mathrm{F}_M^{-1}(\varepsilon)\right)-\mathrm{F}_t\left(\mathrm{F}_M^{-1}(\varepsilon)\right) \tag{4.5}$$

where F_t is the CDF of $\sum_{i=1}^{t}z_i$. In (4.3), we use the fact that the probability that the receiver decodes the packet successfully in the t^{th} time block is equal to the probability that the receiver decodes the packet within t time blocks minus the probability that the receiver decodes the packet within $t-1$ time blocks.

In queue model II, v can only be 1, because a packet definitely leaves the queue at the end of each HARQ period due to either successful transmission or packet drop. Similar to the discussion in queue model I, $t<M$ only corresponds to successful transmission, and $t=M$ corresponds to two cases, in which the receiver gets the packet in the M^{th} time block, or an outage happens and the packet is dropped. Therefore, we have

$$P_{t,1,Q_2}=\begin{cases}\mathrm{F}_{t-1}\left(\mathrm{F}_M^{-1}(\varepsilon)\right)-\mathrm{F}_t\left(\mathrm{F}_M^{-1}(\varepsilon)\right) & t<M\\ \mathrm{F}_{M-1}\left(\mathrm{F}_M^{-1}(\varepsilon)\right) & t=M.\end{cases} \tag{4.6}$$

For the case of $t=M$, we use the fact the $P_{t,1,Q_2}=\mathrm{F}_{M-1}\left(\mathrm{F}_M^{-1}(\varepsilon)\right)-\mathrm{F}_M\left(\mathrm{F}_M^{-1}(\varepsilon)\right)+\varepsilon$, and $\mathrm{F}_M\left(\mathrm{F}_M^{-1}(\varepsilon)\right)=\varepsilon$.

Proposition 2 *The throughput of HARQ-CC scheme with fixed outage probability ε,*

deadline constraint M, and constant-rate arrivals is given by

$$r_{th} = \begin{cases} C_{E,Q_1} & \text{for queue model I} \\ (1-\varepsilon)C_{E,Q_2} & \text{for queue model II,} \end{cases} \tag{4.7}$$

where the effective capacity for queue model j is given by

$$C_{E,Q_j} = -\frac{1}{\theta}\log\left(\max\{|\lambda_{1,Q_j}|,\cdots,|\lambda_{M,Q_j}|\}\right). \tag{4.8}$$

$\{\lambda_{1,Q_j},\cdots,\lambda_{M,Q_j}\}$ are the eigenvalues of the matrix $\mathbf{A_{Q_j}}$ expressed as

$$\mathbf{A_{Q_j}} = \begin{pmatrix} a_{1,Q_j} & a_{2,Q_j} & \cdots & a_{M-1,Q_j} & a_{M,Q_j} \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{pmatrix} \tag{4.9}$$

where

$$a_{k,Q_j} = \begin{cases} \left[F_{k-1}\left(F_M^{-1}(\varepsilon)\right) - F_k\left(F_M^{-1}(\varepsilon)\right)\right]e^{-\theta R} & 1 \le k \le M-1 \\ \left[F_{M-1}\left(F_M^{-1}(\varepsilon)\right) - \varepsilon\right]e^{-\theta R} + \varepsilon & k = M \text{ and } j = 1 \\ F_{M-1}\left(F_M^{-1}(\varepsilon)\right)e^{-\theta R} & k = M \text{ and } j = 2. \end{cases} \tag{4.10}$$

Proposition 2 can be directly obtained by inserting our characterization of P_{t,v,Q_j} into Theorem 1 in [15]. Since the authors of [15] did not consider packet drop, we need to redefine the variable ν in [15] as the number of packets leaving the queue in a transmission period, in order to apply their theorem to our queue models. In the Section 4.1.2, simulation results are provided to verify our effective capacity characterization.

4.1.1.2 Throughput of HARQ-CC with ON-OFF Discrete-Time Markov Source

Since the departure and arrival processes at the transmitter are independent, for both queue models I and II, the effective capacity characterizations in Proposition 2 are still valid.

Theorem 7 *For the ON-OFF discrete-time Markov source with fixed outage probability ε and deadline constraint M, the throughput in queue model I is given by*

$$r_{th} = \frac{P_{ON}}{\theta} \log \left(\frac{e^{2\theta C_{E,Q_1}} - p_{11} e^{\theta C_{E,Q_1}}}{1 - p_{11} - p_{22} + p_{22} e^{\theta C_{E,Q_1}}} \right), \tag{4.11}$$

and the throughput in queue model II is given by

$$r_{th} = (1-\varepsilon) \frac{P_{ON}}{\theta} \log \left(\frac{e^{2\theta C_{E,Q_2}} - p_{11} e^{\theta C_{E,Q_2}}}{1 - p_{11} - p_{22} + p_{22} e^{\theta C_{E,Q_2}}} \right), \tag{4.12}$$

where the effective capacities C_{E,Q_1} and C_{E,Q_2} are given in (4.8).

According to our discussion in Section 3.2.1.3, the throughput of queue model I is given by r_{avg}, and the throughput of queue model II is given by $(1-\varepsilon) r_{\text{avg}}$. With the result in (2.14), we readily obtain the throughput expressions given in Theorem 7.

4.1.1.3 Throughput of HARQ-CC with ON-OFF Fluid Markov Source

Theorem 8 *For ON-OFF fluid Markov source with fixed outage probability ε and deadline constraint M, the throughput in queue model I is given by*

$$r_{th} = P_{ON} C_{E,Q_1} \frac{\theta C_{E,Q_1} + \alpha + \beta}{\theta C_{E,Q_1} + \alpha}, \tag{4.13}$$

and the throughput in queue model II is given by

$$r_{th} = (1-\varepsilon) P_{ON} C_{E,Q_2} \frac{\theta C_{E,Q_2} + \alpha + \beta}{\theta C_{E,Q_2} + \alpha}, \tag{4.14}$$

where the effective capacities C_{E,Q_1} and C_{E,Q_2} are given in (4.8).

The throughput expressions in Theorem 8 can be determined directly using the maximum average arrival rate given in (2.20).

4.1.1.4 Throughput of HARQ-CC with ON-OFF MMPS

Theorem 9 *For ON-OFF MMPS with fixed outage probability ε and deadline constraint M, the throughput in queue model I is given by*

$$r_{th} = P_{ON} C_{E,Q_1} \frac{\theta}{e^{\theta} - 1} \frac{\theta C_{E,Q_1} + \alpha + \beta}{\theta C_{E,Q_1} + \alpha}, \tag{4.15}$$

and the throughput in queue model II is given by

$$r_{th} = (1-\varepsilon) P_{ON} C_{E,Q_2} \frac{\theta}{e^{\theta} - 1} \frac{\theta C_{E,Q_2} + \alpha + \beta}{\theta C_{E,Q_2} + \alpha}, \tag{4.16}$$

where the effective capacities C_{E,Q_1} and C_{E,Q_2} are given in (4.8).

The throughput expressions in Theorem 9 can be obtained directly using the maximum average arrival rate given in (2.24).

4.1.2 Numerical Results

In this subsection, we further investigate the throughput of HARQ-CC with ON-OFF Markov arrivals in the presence of QoS constraints. Throughout this subsection, we assume Rayleigh fading channel with exponentially distributed fading power having a mean value of $\mathbb{E}\{z\} = 1$.

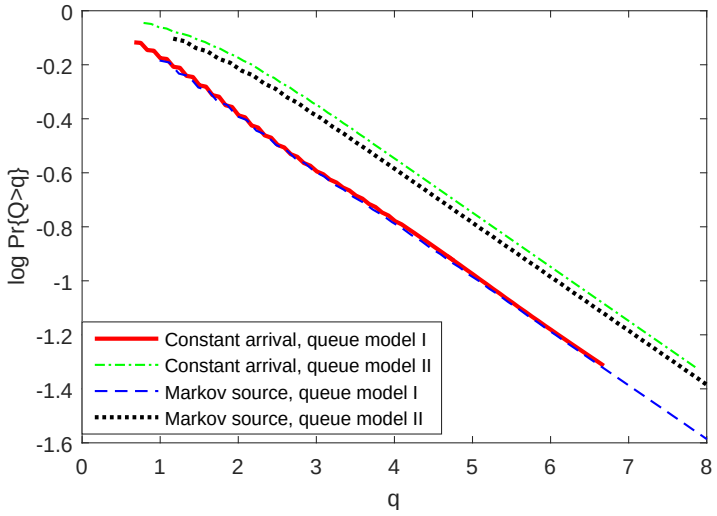

Figure 4.1: Logarithmic buffer overflow probability vs. buffer overflow threshold.

First, we present Monte Carlo simulation results to verify the characterizations in Proposition 2 and Theorem 7. In Fig. 4.1, we set the queuing constraint as $\theta = 0.2$, choose the outage probability as $\varepsilon = 0.1$, and we plot the logarithmic buffer overflow probabilities $\log \Pr\{Q \geq q\}$ as functions of the buffer overflow threshold q for both queue models with constant-rate arrivals and ON-OFF discrete-time Markov arrivals. For each curve, we repeat the simulations for 100 times, and in each time the simulation is conducted over 1×10^7 time blocks. For the constant-rate arrival model, the arrival rates are given by C_{E,Q_1} and C_{E,Q_2} described in (4.8), for queue models I and II, respectively. For the ON-OFF discrete-time Markov arrivals, we set $p_{11} = 0.4$, $p_{22} = 0.7$, and the arrival rates in the ON state are given by (2.13) for both queue models. From Fig. 4.1, we observe that the logarithmic buffer overflow probabilities decrease linearly even starting from relatively small q values, which agrees with the characterizations in (2.1) and (2.2)[1]. We estimate the slopes via linear regression. The estimated slopes of these four curves are -0.2005 (constant arrival for queue model I), -0.2006 (constant arrival for queue model II), -0.2005 (Markov source for queue model I), and -0.1995 (Markov source for queue model II), and the maximum slope

[1]Note that if (2.1) holds, we have $\log \Pr\{Q \geq q\} \approx -\theta q + \log \varsigma$

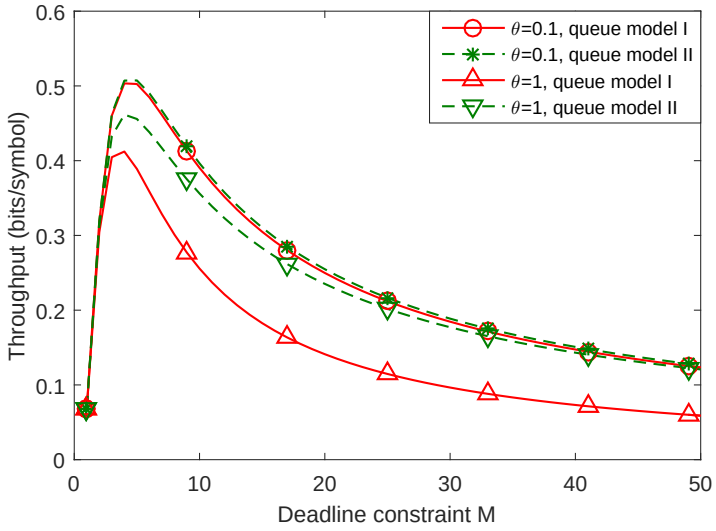

Figure 4.2: Throughput vs. deadline constraint M.

error of these four curves is 0.3%, which is very small, demonstrating the accurateness of the throughput characterizations.

In Fig. 4.2, we set the outage probability as $\varepsilon = 0.05$, and plot the throughput under constant-rate arrivals as a function of the deadline constraint M. We can observe that there exists an optimal M that maximizes the throughput. When M is small, the deadline constraint is strict and the transmitter has to reduce the fixed rate R to satisfy the target outage probability. As M increases, the transmission rate R increases, and the throughput is improved. After the throughput reaches its maximum value, further increase in M results in reduced throughput. For large M and fixed outage probability, the system wastes more time when outage happens, which is not favorable in the presence of the queuing constraint. Also, we find that as θ increases, the throughput becomes smaller in order to satisfy a stricter queuing constraint.

In Fig. 4.3, we set the deadline constraint as $M = 3$, and plot the throughput under constant-rate arrivals as a function of the outage probability ε. Similarly, we can observe that there exists an optimal ε that maximizes the throughput, and the explanation is similar to the influence of M in Fig. 4.2. When we compare the

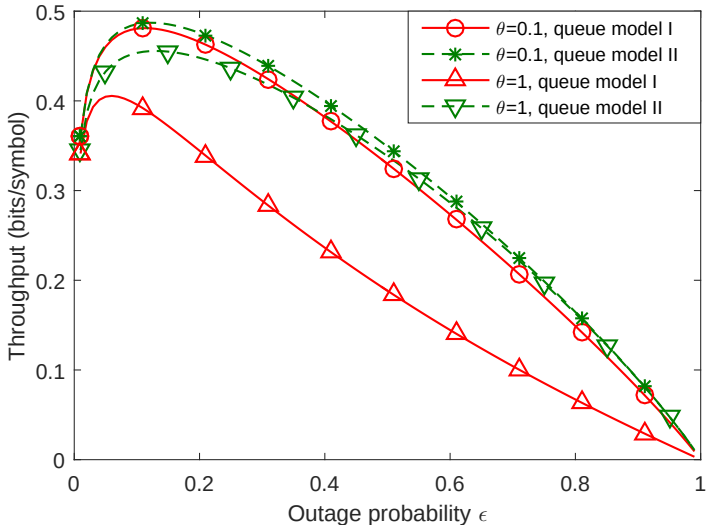

Figure 4.3: Throughput vs. outage probability ε.

throughputs of queue models I and II under constant-rate arrivals in Figs. 4.2 and 4.3, we observe that queue model II always has higher throughput than queue model I. When there is no queuing constraint, the throughput of queue model I is $r_{\text{th}} = r_{\text{avg}} = \frac{(1-\varepsilon)R}{\mathbb{E}\{T\}}$, and the throughput of queue model II is $r_{\text{th}} = (1-\varepsilon)r_{\text{avg}} = (1-\varepsilon)\frac{R}{\mathbb{E}\{T\}}$ [7] [15], where T is the duration of a transmission period. Therefore, these two queue models have the same throughput in the absence of queuing constraints. When the queuing constraint is imposed, queue model II has an advantage. Moreover, the throughput gap between these two queue models increases when we increase θ from 0.1 to 1.

In Figs. 4.4 and 4.5, we fix $M = 10$, $\varepsilon = 0.01$ and $\theta = 0.1$ to investigate the impact of source randomness on the throughput. Since source randomness has similar impact on queue models I and II, we only consider queue model I in the following discussion about the influence of source randomness. In Fig. 4.4, we plot the throughput as a function of P_{ON} for the ON-OFF discrete-time Markov source. We can observe that the throughput is an increasing function of P_{ON}. As P_{ON} decreases, the arrival rate in ON state r needs to increase with a certain rate in order to keep r_{avg} non-decreasing. However, with the same departure process, it is difficult to satisfy the queuing constraints and keep the throughput non-decreasing simultaneously.

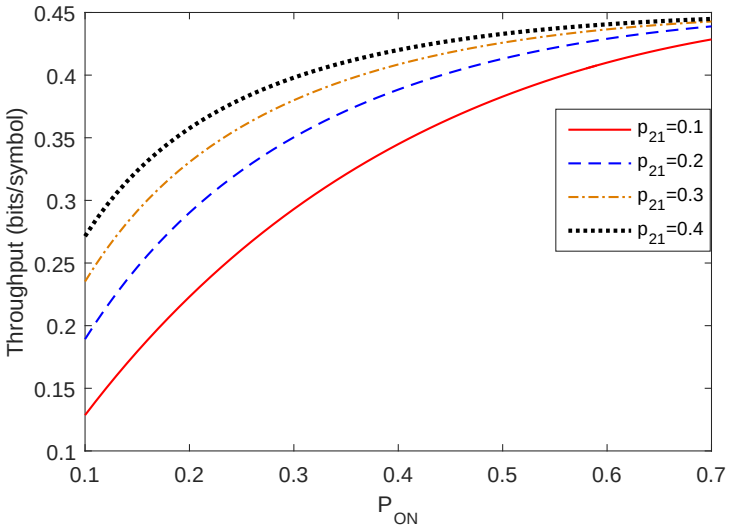

Figure 4.4: Throughput vs. P_{ON} for the ON-OFF discrete-time Markov source.

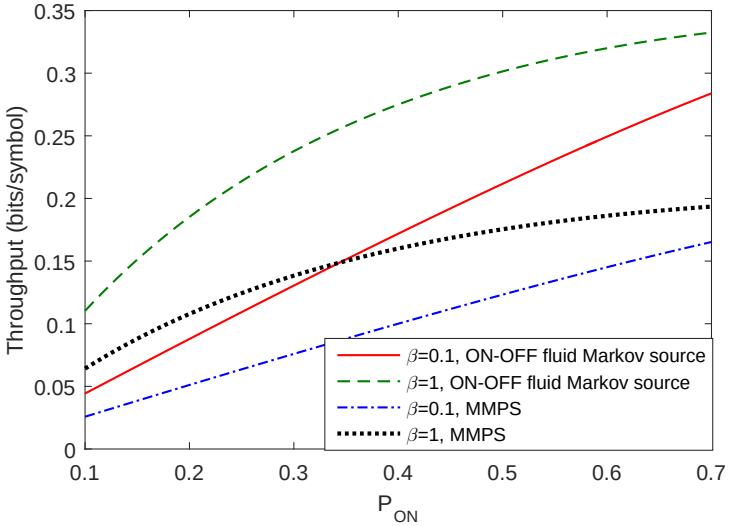

Figure 4.5: Throughput vs. P_{ON} for the ON-OFF Markov fluid source and MMPS.

Therefore, the throughput decreases as P_{ON} decreases. Similar explanation can be applied to the ON-OFF Markov fluid source and MMPS in Fig. 4.5. As $P_{ON} \to 1$, the ON-OFF discrete-time and fluid Markov sources become constant-rate arrival sources, which implies that constant-rate arrivals have higher throughput. We also observe in Fig. 4.4 that higher p_{21} values improve the throughput. For the same P_{ON}, higher p_{21} values make the source transitions from the ON state to the OFF state to occur more frequently, giving a better chance for the transmitter to shorten its queue length. Similarly, in Fig. 4.5, higher β values improve the throughput for both ON-OFF Markov fluid source and MMPS.

Finally, in Fig. 4.5, we find that the ON-OFF Markov fluid source has higher throughput than MMPS. Comparing the results in Theorems 8 and 9, we note that there is an additional factor $\frac{\theta}{e^{\theta}-1}$ in the throughput expression of MMPS. It is straightforward to show that $\lim_{\theta \to 0} \frac{\theta}{e^{\theta}-1} = 1$, and $\frac{\theta}{e^{\theta}-1}$ is an decreasing function of θ. These properties indicate that the throughput of MMPS improves as θ decreases, and it has the same throughput as the ON-OFF Markov fluid source when there is no queuing constraint.

4.2 Throughput of HARQ-IR with Finite Block-length Codes and QoS Constraints

4.2.1 System Model and Preliminaries

As depicted in Fig. 3.1, the same point-to-point wireless communication system is considered in this section. It is assumed that arriving data packets are initially stored in a buffer at the transmitter, which operates under queuing constraints, before being sent to the receiver. Also, we assume a block flat-fading channel in which the fading coefficients stay constant within one block, but change independently across blocks.

Each fading block is assumed to have a duration of l symbols. We use subscript i as the index of the fading block. Under these assumptions, the received signal in the i^{th} block can be written as

$$\mathbf{y}_i = h_i \mathbf{x}_i + \mathbf{n}_i \quad i = 1, 2, \ldots \tag{4.17}$$

Above, $\mathbf{x}_i$ and $\mathbf{y}_i$ are the transmitted and received signal vectors, respectively, and h_i denotes the channel fading coefficient. The average transmission energy per symbol of the transmitted signal $\mathbf{x}_i$ is given by $\mathcal{E} = \mathbb{E}\{\|\mathbf{x}_i\|^2\}/l$. Also, $\mathbf{n}_i$ represents the noise vector with i.i.d. circularly-symmetric, zero-mean Gaussian components, each with variance N_0. Therefore, we can denote the signal-to-noise ratio at the transmitter as $\text{SNR} = \frac{\mathcal{E}}{N_0}$.

4.2.1.1 HARQ-IR and Deadline Constraints

It is assumed that the system employs HARQ-IR scheme to guarantee the reliability of transmissions. The transmission rate is fixed at lR (bits/block) at the transmitter, where l is the number of symbols in each fading block, or equivalently the blocklength of each codeword, and R is the fixed rate in bits/symbol. Each packet is encoded into M codeword blocks (where M also represents the deadline constraint introduced below), and each block has a length of l symbols. In each time block, transmitter sends one codeword block to the receiver. If the receiver decodes the received packet correctly, it sends an ACK feedback to the transmitter through an error-free feedback link, and a new packet is sent in the next time block. If the receiver cannot decode the packet, a retransmission request is sent through the feedback link, and another codeword block of the same packet is sent in the next time block [5]. For simplicity, we assume an ideal ARQ protocol in our analysis, in which the transmitter gets the feedback immediately at the end of each time block without any delay.

Deadline constraint limits the maximum duration of a transmission period as M time blocks. We assume that an HARQ period lasts until either the receiver gets the packet without error or if the limit on the duration of the transmission period is reached, and then the transmitter starts with another packet in the next transmission period. An outage happens when the receiver does not receive the packet within one transmission period, or equivalently M decoding errors occur successively for a packet. The outdated packet is discarded in such a case. The duration of an HARQ period is denoted by the random variable T, where $1 \leq T \leq M$, and the outage probability (or equivalently deadline violation probability) is represented by ε.

In the HARQ-IR scheme, additional information is sent in each retransmission and the receiver combines all received code blocks in the same transmission period to decode the transmitted packets. Detailed introduction about HARQ-IR is provided in Section 3.1.

Under the constant-rate arrival assumption, the throughput (in bits/symbol) is given by $(1-\varepsilon)$ times the effective capacity (normalized by the blocklength l):

$$r_{\text{th}} = (1-\varepsilon)C_E(\theta, \text{SNR})/l = -\frac{1-\varepsilon}{l\theta}\Lambda_c(-\theta). \tag{4.18}$$

When the arrival rate is not constant, we need to formulate the LMGF of the arrival process as a function of the average arrival rate, and obtain the throughput by solving (2.4).

4.2.2 Throughput of HARQ-IR with Queuing Constraints and Finite Blocklength Codes

In this subsection, we study the throughput of HARQ-IR scheme with statistical queuing constraints, finite blocklength codes, and deadline limits. Initially, we consider constant-rate arrivals, and characterize the throughput by employing the effective

capacity formulation. Subsequently, we incorporate random arrival models by considering ON-OFF discrete-time Markov sources and determine the system throughput using the characterizations obtained in the constant-rate arrival model.

4.2.2.1 Outage Probability for HARQ-IR at Finite Blocklengths

As noted before, with HARQ-IR, the received information is accumulated at the receiver. At the end of the m^{th} trial in a transmission period, the receiver combines the m received codeword blocks to decode the packet, which is equivalent to decoding a codeword with m subblocks and each subblock has a length of l symbols from the perspective of achievable rate. According to the results in [85], the relationship between the fixed transmission rate and error probability is given by

$$R = \sum_{i=1}^{m} \log_2 \left(1 + \text{SNR} z_i\right) - \sqrt{\sum_{i=1}^{m} \frac{(\text{SNR} z_i + 2)\text{SNR} z_i}{l(\text{SNR} z_i + 1)^2}} Q^{-1}(\nu) \log_2 e + \frac{\log(ml)}{l} + \frac{o(1)}{l} \tag{4.19}$$

for the m^{th} trial, where l is the blocklength, $Q^{-1}(\cdot)$ represents the inverse Q-function, ν is the decoding error probability, and $z_i = |h_i|^2$ is the magnitude-square of the fading coefficient. From (4.19), we can express the decoding error probability for the m^{th} trial or attempt of packet transmission as

$$\nu_m = Q\left(\frac{\sum_{i=1}^{m} \log_2 \left(1 + \text{SNR} z_i\right) + \frac{\log(ml)}{l} - R}{\log_2 e \sqrt{\sum_{i=1}^{m} \frac{(2+\text{SNR} z_i)\text{SNR} z_i}{l(\text{SNR} z_i + 1)^2}}}\right) \tag{4.20}$$

for given channel fading $\mathbf{z} = (z_1, \cdots, z_m)$. Therefore, we can obtain the probability mass function (pmf) of T, which represents the duration of a transmission period, as

expressed in (4.21)

$$\Pr\{T=t\} = \begin{cases} 1-\mathbb{E}_{\mathbf{z}}\{\nu_1\} & \text{for } t=1 \\ \Pr\{T\le t\}-\Pr\{T\le t-1\} = \mathbb{E}_{\mathbf{z}}\{\nu_{t-1}\}-\mathbb{E}_{\mathbf{z}}\{\nu_t\} & \text{for } 2\le t\le M-1 \\ \mathbb{E}_{\mathbf{z}}\{\nu_{M-1}\} & \text{for } t=M \end{cases} \tag{4.21}$$

One assumption we have is that if the receiver cannot decode the packet correctly using all m received codeword blocks in the m^{th} trial, then it cannot decode the packet in the previous $m-1$ trials. This is due to the fact that it is more difficult to have correct decoding with less information, and the receiver uses less codeword blocks in previous trials. Therefore, the probability that the receiver decodes a packet within t trials is given by

$$\Pr\{T\le t\} = 1-\mathbb{E}_{\mathbf{z}}\{\nu_t\}. \tag{4.22}$$

In (4.21), for $2\le t\le M-1$, $T=t$ indicates that only the t^{th} trial has been successful, and the first $t-1$ trials of packet transmission has ended up with error. Having $T=M$ indicates that the first $M-1$ trials have failed, and the result of the last attempt does not have any influence on the pmf, because the deadline constraint forces the transmission to cease after M trials. Recall that an outage (or deadline violation) event happens if the receiver experiences M decoding errors successively in a transmission period. The outage probability can be expressed as

$$\varepsilon = \mathbb{E}_{\mathbf{z}}\{\nu_M\}. \tag{4.23}$$

In Fig. 4.6, we set the blocklength as $l=100$ and SNR as 0 dB, and plot the outage probabilities as functions of the fixed transmission rate R for different deadline

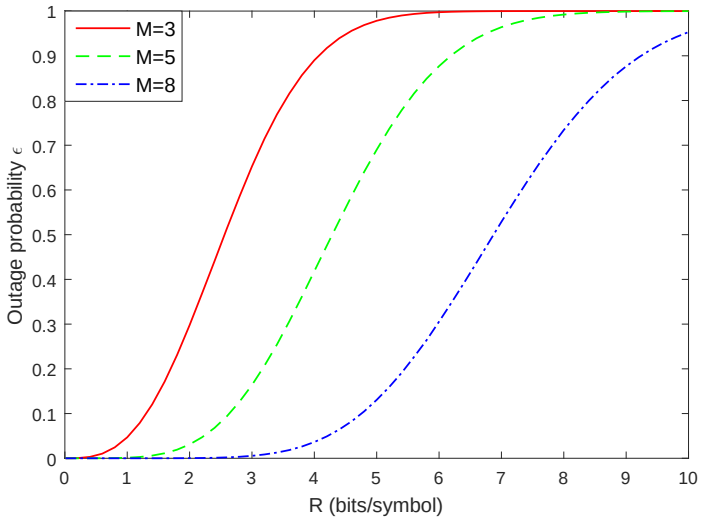

Figure 4.6: Outage probability vs. R

constraints M. As R increases, the decoding error probability ν increases, which leads to an increased outage probability. We also note that as M increases, the deadline constraints become looser, and we can expect lower outage probabilities. Finally, the fixed transmission rate R can be numerically determined using (4.23) and (4.20) once a target outage probability ε and deadline limit M are specified.

4.2.2.2 Throughput of HARQ-IR with Constant-rate Arrivals

Proposition 3 *The throughput of HARQ-IR scheme (in bits/symbol) with fixed transmission rate R (bits/symbol), deadline constraint M, QoS exponent θ, and constant-rate arrivals is given by*

$$r_{th} = (1-\varepsilon)C_E/l \tag{4.24}$$

where the effective capacity C_E (in bits/block) is given by

$$C_E = -\frac{1}{\theta}\log\left(\max\{|\lambda_1|, \cdots, |\lambda_M|\}\right). \tag{4.25}$$

Above, $\{\lambda_1, \cdots, \lambda_M\}$ *are the eigenvalues of the matrix* $\mathbf{A}$ *given in (4.26)*

$$\mathbf{A} = \begin{pmatrix} \Pr\{T=1\}e^{-\theta l R} & \Pr\{T=2\}e^{-\theta l R} & \cdots & \Pr\{T=M-1\}e^{-\theta l R} & \Pr\{T=M\}e^{-\theta l R} \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{pmatrix} \tag{4.26}$$

Proposition 3 can be shown by applying Theorem 1 in [15] to our model. Since the authors in [15] did not consider packet drop, we need to redefine the variable ν in [15] as the number of packets leaving the queue in a transmission period, which is always equal to 1 in our model due to the packet drop mechanism. In the Section 4.2.3, simulation results are provided to verify our effective capacity characterization.

4.2.2.3 Throughput of HARQ-IR with ON-OFF Discrete-Time Markov Source

Since the departure and arrival processes at the transmitter are independent, the effective capacity characterization in Proposition 3 is still valid.

Theorem 10 *For the ON-OFF discrete-time Markov source with fixed transmission rate* R *(bits/symbol), deadline constraint* M*, and QoS exponent* θ*, the throughput (in bits/symbol) is given by*

$$r_{th} = \frac{1-\varepsilon}{l}\frac{P_{ON}}{\theta}\log\left(\frac{e^{2\theta C_E} - p_{11}e^{\theta C_E}}{1-p_{11}-p_{22}+p_{22}e^{\theta C_E}}\right), \tag{4.27}$$

where the effective capacity C_E *is given in (4.25).*

Similar to the discussion in Section 4.1.1.2, the throughput is given by $(1-\varepsilon)r_{\text{avg}}$. Using the maximum average arrival rate given in (2.14), we readily obtain the through-

put expressions given in Theorem 10.

From (4.27), it is very easy to show that the throughput of HARQ-IR with ON-OFF discrete time Markov sources is an increasing function of the effective capacity C_E by checking the first order derivative of r_{th} with respect to C_E. Therefore, the transmission parameters, such as R and ε, that maximize the throughput for the constant-rate arrival models also maximize the throughput for the ON-OFF discrete time Markov arrival model.

The influence of the source burstiness was discussed in [86], in which it was shown that source burstiness degrades the energy efficiency under queuing constraints. Similar analysis can be applied to our scenario. The source is less bursty if it stays in the ON state for a longer period, resulting in smaller instantaneous arrival rates r for fixed average arrival rate r_{avg}. In other words, for different sources with the same r_{avg}, the one with less burstiness or equivalently smaller instantaneous arrival rate r is more favorable in terms of satisfying the queuing constraints.

4.2.3 Numerical Results

In this subsection, we further investigate the throughput of HARQ-IR with finite blocklength codes in the presence of deadline and QoS constraints. Throughout this subsection, we assume Rayleigh fading channel with exponentially distributed fading power having a mean value of $\mathbb{E}\{z\} = 1$, and SNR is set as 0 dB. We first verify our characterizations in Proposition 3 and Theorem 10 via Monte Carlo simulations. The logarithmic buffer overflow probabilities $\log \Pr\{Q \geq q\}$ are plotted as functions of the buffer overflow threshold q for both constant-rate arrivals and ON-OFF discrete-time Markov arrivals in Fig. 4.7. For both arrival models, we set the QoS exponent as $\theta = 0.01$, deadline constraint $M = 5$, fixed transmission rate $R = 3$ (bits/symbol), and blocklength $l = 100$. For each curve, we repeat the simulations 2000 times, and in each time the simulation is conducted over 1×10^5 time blocks. For the constant-rate

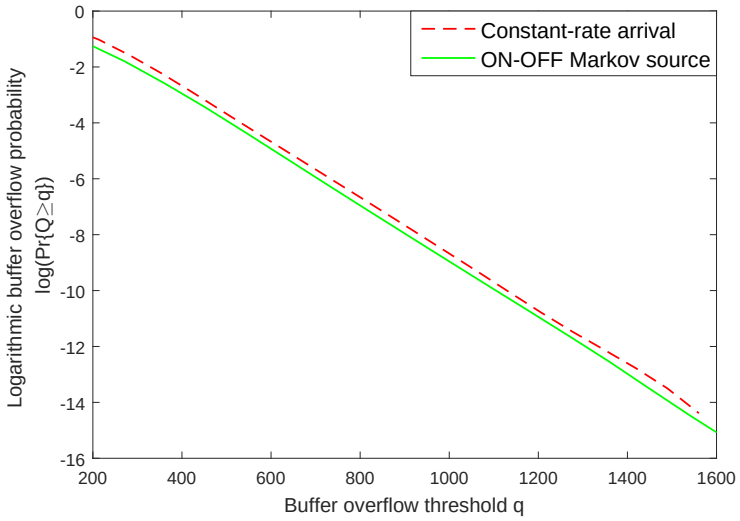

Figure 4.7: Logarithmic overflow probability vs. buffer overflow threshold.

arrival model, the arrival rate is given by the effective capacity in (4.25). For the ON-OFF discrete-time Markov arrivals, we set $p_{11} = 0.3$, $p_{22} = 0.7$, and the arrival rates in the ON state are given by lr (bits/block) in (2.13). It is observed in Fig. 4.7 that the logarithmic buffer overflow probabilities decrease almost linearly when q is sufficiently large, which agrees with the characterizations in (2.1) and (2.2)[2]. When $q > 1100$, we estimate the slopes of these two curves via linear regression, and the estimated slopes are -0.0099 and -0.0100 for the constant-rate and ON-OFF Markov arrival models, respectively. The slope errors are smaller than 1%, which is very small, demonstrating the accurateness of the throughput characterizations.

In Fig. 4.8, we plot the throughput as a function of the fixed transmission rate R for different deadline constraints with constant arrival sources, and QoS exponent $\theta = 0.1$. As shown in Fig. 4.8, there exists a unique optimal R value that maximizes the throughput. When R is small, the departure rate is also small, which limits the throughput. When R is too large, the outage probability ε is large, and most of the packets violate the deadline constraint and are discarded. Also, as M increases, the deadline constraints become looser, and we can achieve a higher maximum throughput

[2]Note that if (2.1) holds, we have $\log \Pr\{Q \geq q\} \approx -\theta q + \log \varsigma$

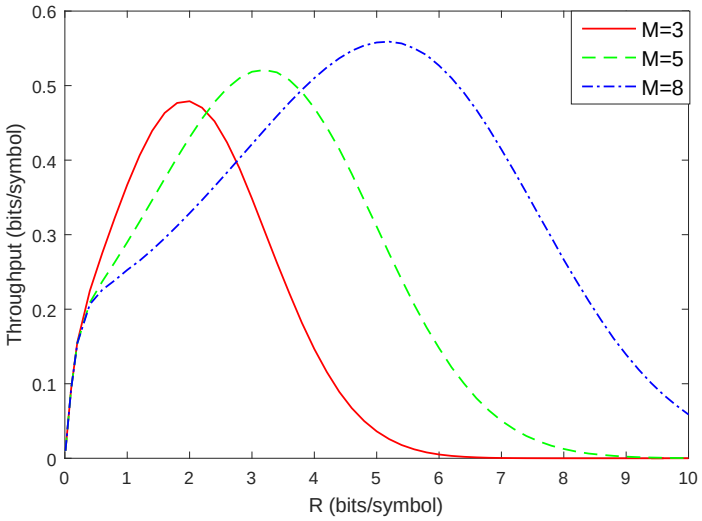

Figure 4.8: Throughput vs. fixed transmission rate R.

as seen in Fig. 4.8. Similar results were observed in [87] for small θ values without considering finite blocklength effects. In [87], it has been shown that the throughput of HARQ-IR is an increasing function of the fixed transmission rate R without deadline constraints. Therefore, we can have higher R values and low outage probabilities when M is large, which improves the maximum throughput.

Fig. 4.9 shows the influence of the finite blocklength l for constant arrival models with QoS exponent $\theta = 0.1$. In order to apply the approximation in (4.19), the blocklength l needs to be sufficiently large. In such a case, the throughput decreases as the block length l increases, because large l corresponds to slow fading[3], which is not favorable for delay sensitive systems with queueing constraints. In slow fading cases, strong attenuation would last for a longer time, leading to buffer overflows. Therefore, larger l value is expected to have a stronger impact on the throughput when the system has stricter queueing constraints. Indeed, this is observed in Fig. 4.9 where we see that the throughput curves associated with larger values of θ (indicating stricter queuing constraints) decrease faster with increasing l.

[3]This is by our block-fading assumption in which each fading block consists of l symbols.

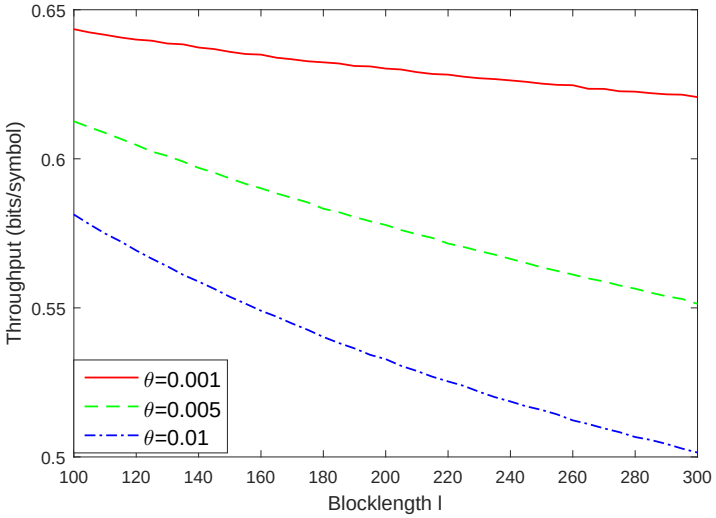

Figure 4.9: Throughput vs. blocklength l.

Chapter 5

Throughput of Cooperative Relay Networks under Statistical Queuing Constraints

In this chapter, we investigate the throughput of cooperative relay networks under statistical queuing constraints. Three types of cooperative relay networks are considered, namely two-hop relay channel, two-way relay channel and multi-source multi-destination relay network.

In Section 5.1, throughput of two-hop wireless relay channels is studied in the finite blocklength regime. Half-duplex relay operation, in which the source node initially sends information to the intermediate relay node and the relay node subsequently forwards the messages to the destination, is considered. It is assumed that all messages are stored in buffers before being sent through the channel, and both the source node and the relay operate under statistical queueing constraints. After characterizing the transmission rates in the finite blocklength regime, the system throughput is formulated via queueing analysis. Subsequently, several properties of the throughput function in terms of system parameters are identified, and an efficient algorithm is

proposed to maximize the throughput. Interplay between throughput, queueing constraints, relay location, time allocation, and code blocklength is investigated through numerical results.

In Section 5.2, throughput of two-way relaying under buffer constraints is studied. In the two-way relay system, source nodes initially send their messages to the relay in the multiple-access phase. Relay decodes and stores the messages from different sources in different buffers and subsequently broadcasts a superimposed signal. It is assumed that both source nodes and the relay operate in the presence of statistical queueing constraints. Under these assumptions, arrival rates that can be supported in this system are investigated through the LMGFs of the arrival and service processes. In particular, after identifying the service rates in the multiple-access and broadcast phases and addressing the stability conditions, characterizations of the maximum arrival rates are provided in terms of system resource allocation parameters, signal-to-noise ratios, and quality-of-service exponents. Impact of different parameters on the performance is investigated through numerical results.

In Section 5.3, the throughput of relay networks with multiple source-destination pairs under queueing constraints has been investigated for both variable-rate and fixed-rate schemes. When CSI is available at the transmitter side, transmitters can adapt their transmission rates according to the channel conditions, and achieve the instantaneous channel capacities. In this case, the departure rates at each node have been characterized for different system parameters, which control the power allocation, time allocation and decoding order. In the other case of no CSI at the transmitters, a simple ARQ protocol with fixed rate transmission is used to provide reliable communication. Under this ARQ assumption, the instantaneous departure rates at each node can be modeled as an ON-OFF process, and the probabilities of ON and OFF states are identified. With the characterization of the arrival and departure rates at each buffer, stability conditions are identified and effective capacity analysis

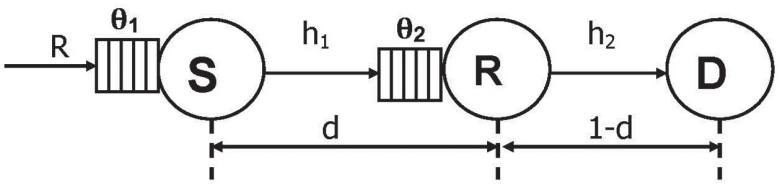

Figure 5.1: The two-hop relay system with buffer constraints.

is conducted for both cases to determine the system throughput under statistical queueing constraints. In addition, for the variable-rate scheme, the concavity of the sum rate is shown for certain parameters, helping to improve the efficiency of parameter optimization.

5.1 Throughput of Two-Hop Wireless Channels with Queuing Constraints and Finite Blocklength Codes

5.1.1 System Model and Preliminaries

5.1.1.1 System Model

The two-hop relay channel is shown in Fig. 5.1. In this model, source node **S** sends information to the receiver **D** with the help of the intermediate relay node **R**. We assume that there is no direct link between **S** and **D** (which, for instance, holds, if these nodes are sufficiently far apart in distance). All data packets are stored in buffers before being transmitted through wireless channels. Data arriving to the source node **S** is initially buffered before transmission to the relay. Similarly, the relay, upon receiving the signal from the source node and decoding the message, places the decoded data from the source in its own buffer before forwarding it to

the destination **D**. Both the source and the intermediate relay nodes operate under statistical queuing constraints. More specifically, buffer violation probabilities are constrained to decay exponentially for large buffer thresholds. Detailed discussion on the queuing constraints is provided in Chapter 2.

Since we consider half-duplex relay operation, reception and transmission at the relay occur in non-overlapping intervals. We introduce the parameter $\tau \in (0,1)$ as the fraction of time allocated to the initial phase, in which only the $\mathbf{S}-\mathbf{R}$ link is active. Then, the fraction of time allocated to the second phase is $1-\tau$, in which only the $\mathbf{R}-\mathbf{D}$ link is active. Next, we express the discrete-time input and output relationships in both phases. In the initial phase, the signal Y_r received at the relay can be expressed as

$$Y_r[i] = h_1[i]X[i] + n_r[i] \tag{5.1}$$

where X denotes the signal transmitted from the source node, h_1 represents the fading coefficient of the $\mathbf{S}-\mathbf{R}$ link, and n_r is the Gaussian noise at the relay. In the second phase, the received signal Y at receiver **D** is given by

$$Y[i] = h_2[i]X_r[i] + n_D[i] \tag{5.2}$$

where X_r denotes the signal sent from the relay node, h_2 represents the fading coefficient of the $\mathbf{R}-\mathbf{D}$ link, and n_D is the Gaussian noise at the receiver **D**. The inputs are subject to individual average energy constraints $\mathbb{E}\{|X|^2\} \leq \bar{P}_s/B$ and $\mathbb{E}\{|X_r|^2\} \leq \bar{P}_r/B$, where B is the bandwidth in the system. We assume that the fading coefficients $h_j, j = \{1,2\}$ are jointly stationary and ergodic discrete-time processes, and we denote the magnitude-square of the fading coefficients by $z_j[i] = |h_j[i]|^2$. Above, in the channel input-output relationships, the noise components are all zero-mean, circularly symmetric, complex Gaussian random variables

with variance $\mathbb{E}\{|n_k[i]|^2\} = N_k$ for $k = \{r, D\}$, and all noise samples are assumed to form an i.i.d. sequence. We denote the signal-to-noise ratios at the source and relay by $\text{SNR}_s = \frac{\bar{P}_s}{N_r B}$ and $\text{SNR}_r = \frac{\bar{P}_r}{N_D B}$, respectively.

5.1.1.2 Coding Rate in Finite Blocklength Regimes

In this section, we investigate the throughput achieved with finite blocklength coding. In the complex Gaussian noise channel with channel gain z, the coding rate (in bits per channel use) is approximated by [18] [19]

$$r = \log_2(1 + \text{SNR}z) - \sqrt{\frac{1}{m}\left(1 - \frac{1}{(\text{SNR}z + 1)^2}\right)} Q^{-1}(\epsilon) \log_2 e + \frac{\log m}{m} + \frac{O(1)}{m} \quad (5.3)$$

where m represents the coding blocklength, $\epsilon \in (0, 1)$ denotes the error probability, $Q^{-1}(\cdot)$ is the inverse Gaussian Q-function, and $O(1)$ denotes a constant term[1]. Hence, the approximation becomes more accurate as m increases. Also, the coding rate is an monotonic increasing function of the target error probability ϵ. Moreover, as the blocklength m grows without bound, the coding rate becomes $r = \log_2(1 + \text{SNR}z)$, which is the Shannon capacity.

For our half-duplex two-hop relay system, every time block is divided into two for the two transmission phases. Therefore, the coding blocklength of the source and relay are given by τm and $(1 - \tau)m$, respectively. From (5.3), we can characterize the coding rates of the source and relay nodes as

$$r_1 = \log_2(1 + \text{SNR}_s z_1) + \frac{\log(\tau m)}{\tau m} - \sqrt{\frac{1}{\tau m}\left(1 - \frac{1}{(\text{SNR}_s z_1 + 1)^2}\right)} Q^{-1}(\epsilon_1) \log_2 e \quad (5.4)$$

[1]Therefore, the term $\frac{O(1)}{m}$ vanishes fast with increasing blocklength m and is neglected in the remainder of the formulations and analysis

and

$$r_2 = \log_2\left(1+\mathrm{SNR}_r z_2\right) + \frac{\log\left((1-\tau)m\right)}{(1-\tau)m} - \sqrt{\frac{1}{(1-\tau)m}\left(1-\frac{1}{(\mathrm{SNR}_r z_2+1)^2}\right)}Q^{-1}(\epsilon_2)\log_2 e \tag{5.5}$$

respectively, where ϵ_1 and ϵ_2 are the target error probabilities of the $\mathbf{S}-\mathbf{R}$ and $\mathbf{R}-\mathbf{D}$ transmissions, respectively.

5.1.2 Throughput of Two-hop Relay Channels With Finite Blocklength Codes

In this subsection, we characterize the system throughput for the two-hop relay system operating under queuing constraints in the finite blocklength regime. In order to apply the effective capacity analysis, we have to guarantee that the stability conditions are satisfied, which require that the average arrival rate is smaller than the average departure rate at both the source and relay nodes. At the source node, the stability condition is guaranteed by (2.25) due to the fact that the constant arrival rate given by (2.25) is always smaller than the average transmission rate between the source and relay[2]. At the relay node, the stability condition is expressed as

$$(1-\epsilon_1)\tau\mathbb{E}\{r_1\} \leq (1-\epsilon_2)(1-\tau)\mathbb{E}\{r_2\} \tag{5.6}$$

where r_1 and r_2 are given by (5.4) and (5.5), respectively.

In [17], the throughput of the two-hop relay channel is studied for both half- and full-duplex scenarios without considering finite blocklength restrictions. Hence, instantaneous transmission rates were given by the Shannon capacities in [17]. Through

[2]It can be shown that the right-hand-side of (2.25) increases with decreasing $\tilde{\theta}$ and converges to the average departure rate as $\tilde{\theta}$ approaches zero [17].

a similar analysis, we can extend this result to the finite blocklength regime.

Theorem 11 *For the half-duplex two-hop relay system with finite code blocklength m, the maximum arrival rate (in bits per channel use) at the source node is given by*

$$\mathrm{R} = \begin{cases} \min\Big\{ -\frac{1}{m\theta_1}\log\mathbb{E}_{z_1}\{\epsilon_1+(1-\epsilon_1)e^{-\tau\theta_1 m r_1}\}, -\frac{1}{m\theta_2}\log\mathbb{E}\{\epsilon_2+(1-\epsilon_2)e^{-(1-\tau)\theta_2 m r_2}\}\Big\} & \theta_2 \le \theta_1 \\ \min\Big\{ -\frac{1}{m\theta_1}\big(\log\mathbb{E}\{\epsilon_2+(1-\epsilon_2)e^{-(1-\tau)\theta_2 m r_2}\}+\log\mathbb{E}\{\epsilon_1+(1-\epsilon_1)e^{\tau(\theta_2-\theta_1) m r_1}\}\big), & \\ \qquad -\frac{1}{m\theta_1}\log\mathbb{E}_{z_1}\{\epsilon_1+(1-\epsilon_1)e^{-\tau\theta_1 m r_1}\}\Big\} & \theta_2 > \theta_1 \end{cases} \tag{5.7}$$

when the stability condition (5.6) is satisfied.

Proof: See Appendix A.8.

5.1.3 Throughput Optimization for Two-hop Relay Systems

In this subsection, we investigate the throughput maximization problem for our two-hop relay system. We assume that both the source and relay nodes transmit at their maximum power level, the blocklength m is given, and the system maximizes its throughput by choosing the optimal τ, ϵ_1 and ϵ_2. This optimization problem can be formulated as

$$\begin{aligned} &\textbf{Maximize}_{\tau,\epsilon_1,\epsilon_2} \quad \mathrm{R} \\ &\textbf{Subject to} \ \ (1-\epsilon_1)\tau\mathbb{E}\{r_1\} \le (1-\epsilon_2)(1-\tau)\mathbb{E}\{r_2\} && (5.8) \\ &\qquad 0<\tau<1 && (5.9) \\ &\qquad 0<\epsilon_1<1 && (5.10) \\ &\qquad 0<\epsilon_2<1 && (5.11) \end{aligned}$$

Compared to the rate optimization in the absence of finite blocklength code restrictions, our optimization problem has more complicated rate expressions and higher dimensionality. Moreover, since the stability region described by (5.8) is not convex, and the objective function R given in (5.7) is not a convex function of the target error probabilities, this optimization problem is in general non-convex. In this setting, we initially establish several key properties of the throughput R and determine the optimal error probabilities ϵ_1 and ϵ_2 under certain conditions. Subsequently, we propose an algorithm that can efficiently solve the optimization problem.

Theorem 12 *For a given τ value, the error probability that maximizes $-\frac{1}{\theta_1}\Lambda_{\mathbf{S},\mathbf{R}}(-\theta_1)$ is given by the solution of*

$$1 = \mathbb{E}\Bigg\{ -\frac{(1-\epsilon_1)\theta_1}{\log 2}\sqrt{\tau m(1-\frac{1}{(1+SNR_s z_1)^2})} \times \dot{Q}^{-1}(\epsilon_1)e^{-\tau\theta_1 m r_1} + e^{-\tau\theta_1 m r_1}\Bigg\}, \tag{5.12}$$

and the error probability that maximizes $-\frac{1}{\theta_2}\Lambda_{\mathbf{R},\mathbf{D}}(-\theta_2)$ is given by the solution of

$$1 = \mathbb{E}\Bigg\{ -\frac{(1-\epsilon_2)\theta_2}{\log 2}\sqrt{(1-\tau)m(1-\frac{1}{(1+SNR_r z_2)^2})} \times \dot{Q}^{-1}(\epsilon_2)e^{-(1-\tau)\theta_2 m r_2} + e^{-(1-\tau)\theta_2 m r_2}\Bigg\}. \tag{5.13}$$

Proof 1 *It was shown in [20] that the unique optimal error probability that maximizes $-\frac{1}{\theta}\Lambda(-\theta)$ in a single-hop model is obtained by solving*

$$\frac{\partial}{\partial\epsilon}\Lambda(-\theta) = 0. \tag{5.14}$$

This property can be directly applied to our half-duplex two-hop system. After taking the derivatives of $\Lambda_{\mathbf{S},\mathbf{R}}(-\theta_1)$ and $\Lambda_{\mathbf{R},\mathbf{D}}(-\theta_2)$ with respect to ϵ_1 and ϵ_2 respectively, and plugging them into (5.14), we obtain (5.12) and (5.13).

Theorem 13 *For a given τ value, assume that ϵ_1^* and ϵ_2^* are the solutions of (5.12) and (5.13), respectively. Then, the throughput given in (5.7) is maximized at $(\epsilon_1^*, \epsilon_2^*)$ when $\theta_1 \geq \theta_2$.*

Proof 2 *When $\theta_1 \geq \theta_2$, we have*

$$\mathrm{R} = \min\Big\{ -\frac{1}{\theta_1}\Lambda_{\mathbf{S},\mathbf{R}}(-\theta_1), -\frac{1}{\theta_2}\Lambda_{\mathbf{R},\mathbf{D}}(-\theta_2)\Big\}. \tag{5.15}$$

Since ϵ_1^* and ϵ_2^* maximize $-\frac{1}{\theta_1}\Lambda_{\mathbf{S},\mathbf{R}}(-\theta_1)$ and $-\frac{1}{\theta_2}\Lambda_{\mathbf{R},\mathbf{D}}(-\theta_2)$, respectively, Theorem 13 follows immediately.

Theorem 14 *For a given τ value, assume that ϵ_2^* is the solution of (5.13), and $\widetilde{\epsilon}_1$ is the optimal probability that maximizes*

$$\hat{\mathrm{R}} = \min\Big\{ -\frac{1}{\theta_1}\Lambda_{\mathbf{S},\mathbf{R}}(-\theta_1), -\frac{1}{\theta_1}\Big(\Lambda_{\mathbf{R},\mathbf{D}}(-\theta_2)\Big|_{\epsilon_2=\epsilon_2^*} + \Lambda_{\mathbf{S},\mathbf{R}}(\theta_2-\theta_1)\Big)\Big\}. \tag{5.16}$$

Then, the throughput given in (5.7) is maximized at $(\widetilde{\epsilon}_1, \epsilon_2^*)$ when $\theta_1 < \theta_2$.

Proof 3 *When $\theta_1 < \theta_2$, we have*

$$\mathrm{R} = \min\Big\{ -\frac{1}{\theta_1}\Lambda_{\mathbf{S},\mathbf{R}}(-\theta_1), -\frac{1}{\theta_1}\Big(\Lambda_{\mathbf{R},\mathbf{D}}(-\theta_2) + \Lambda_{\mathbf{S},\mathbf{R}}(\theta_2-\theta_1)\Big)\Big\}. \tag{5.17}$$

It can be readily shown that the throughput is a non-decreasing function of $-\Lambda_{\mathbf{R},\mathbf{D}}(-\theta_2)$, and $-\Lambda_{\mathbf{R},\mathbf{D}}(-\theta_2)$ achieves its maximum value at $\epsilon_2 = \epsilon_2^*$. Therefore, the throughput achieves its maximum value when we choose $\epsilon_2 = \epsilon_2^*$ and the characterization in the theorem follows.

If $-\frac{1}{\theta_1}\Lambda_{\mathbf{S},\mathbf{R}}(-\theta_1)$ is always smaller than $-\frac{1}{\theta_1}\Big(\Lambda_{\mathbf{R},\mathbf{D}}(-\theta_2)\Big|_{\epsilon_2=\epsilon_2^*} + \Lambda_{\mathbf{S},\mathbf{R}}(\theta_2-\theta_1)\Big)$, then apparently $\widetilde{\epsilon}_1 = \epsilon_1^*$, where ϵ_1^* is the solutions of (5.12). Otherwise, we need to search for the $\widetilde{\epsilon}_1$ that maximizes $\hat{\mathrm{R}}$.

The results in Theorems 13 and 14 do not incorporate the stability condition. The following result gives a characterization when the optimal error probability pair determined by Theorem 13 does not satisfy the stability condition.

Theorem 15 *For a given τ value, if the error probability pair found using Theorem 13 does not satisfy the stability condition, in the case of $\theta_1 \geq \theta_2$, then the optimal error probability pair, satisfying the stability condition, lies on the boundary of the stability region, described by*

$$(1-\epsilon_1)\tau\mathbb{E}\{r_1\} = (1-\epsilon_2)(1-\tau)\mathbb{E}\{r_2\}. \tag{5.18}$$

Proof: See Appendix A.9.

Using the above characterizations, we can optimize the throughput efficiently. We search for the optimal τ value in the region $(0,1)$ with the following steps. For a given τ value, we first check whether the error probabilities given by Theorems 13 and 14 satisfy the stability condition. If satisfied, then we only have to perform a one-dimensional optimization over $\tau \in (0,1)$. If not, we determine the optimal error probability pair on the boundary of the stability region in the case of $\theta_1 \geq \theta_2$, or search in the entire bounded stability region in the case of $\theta_1 < \theta_2$.

5.1.4 Numerical Results

In this subsection, we provide our numerical results. We consider a simple scenario in which all three nodes are placed on a straight line. The distance between **S** and **D** has been normalized to 1, and $d \in (0,1)$ represents the distance between the source node and relay, which is shown in Fig. 5.1. We assume Rayleigh fading with path loss $\mathbb{E}\{z_1\} = d^{-4}$ and $\mathbb{E}\{z_2\} = (1-d)^{-4}$. Unless specified otherwise, we assume $\text{SNR}_s = \text{SNR}_r = 6\text{dB}$, and blocklength is $m = 150$.

In Figs. 5.2 and 5.3, we plot the maximum throughput R and the optimal time

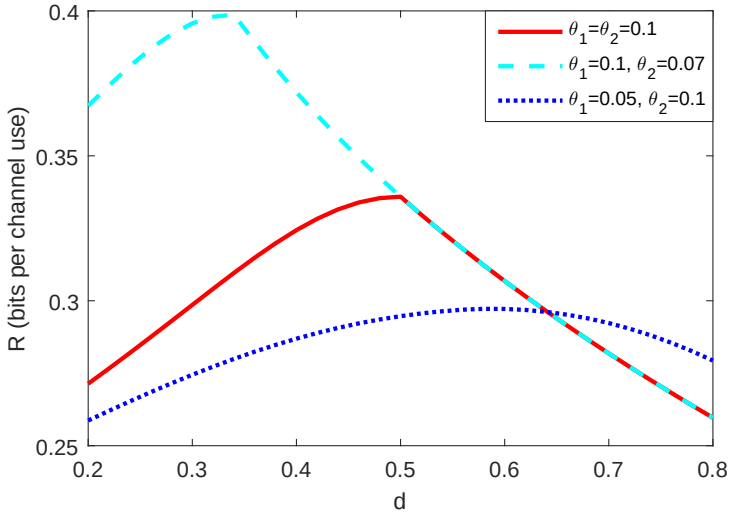

Figure 5.2: The maximum throughput vs. relay location parameter.

allocation parameter τ as functions of the distance parameter d for three different queuing constraint settings, respectively. When the source and relay nodes have the same QoS exponent value, the optimal location of the relay node is the midpoint between **S** and **D** to balance the channel conditions in the $\mathbf{S}-\mathbf{R}$ and $\mathbf{R}-\mathbf{D}$ links. When the source node has a stricter queuing constraint, the relay moves closer to **S** to enhance the departure rate of the source node. Similarly, when the relay node has a larger QoS exponent value, the optimal location of the relay is closer to **D** to improve the channel conditions in the $\mathbf{R}-\mathbf{D}$ link. From (5.7) we know that the link with stricter queuing constraint and smaller departure rate has more influence on the throughput, so the optimal location of the relay is chosen with the goal of improving the link experiencing stricter queuing constraints. Similar mechanisms can be observed in Fig. 5.3. Comparing two dashed lines in Fig. 5.3, we find that the system allocates more time to the link with a more stringent queuing constraint. When $\theta_1 = \theta_2$, the system allocates more time to the link with the worse channel condition. Note that these results are obtained with the optimal values of ϵ_1 and ϵ_2.

In Fig. 5.4, we place the relay at the midpoint and plot the maximum through-

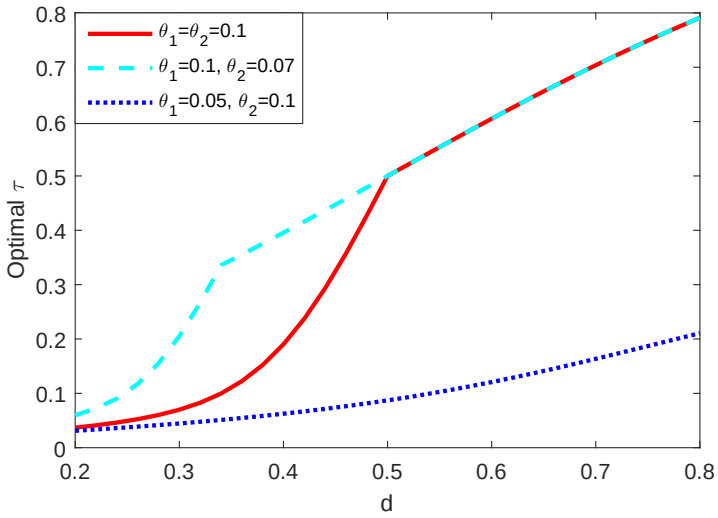

Figure 5.3: The optimal τ vs. relay location parameter.

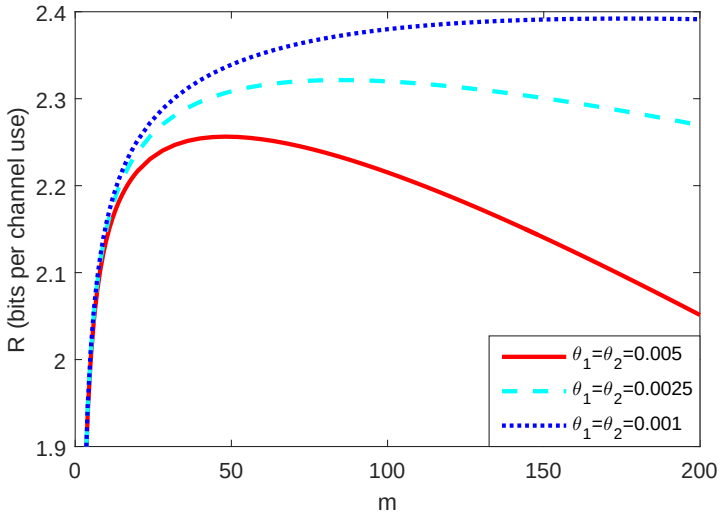

Figure 5.4: The maximum throughput vs. blocklength m.

put as a function of the blocklength m. When m is small, increasing m improves the performance because it increases the departure rates. When m grows beyond a threshold, the throughput starts decreasing, because large m corresponds to slow fading, which is not favorable for delay sensitive systems with queuing constraints. In slow fading cases, strong attenuation would last for a longer time, leading to buffer overflows. Therefore, large m value has a stronger influence on the throughput when

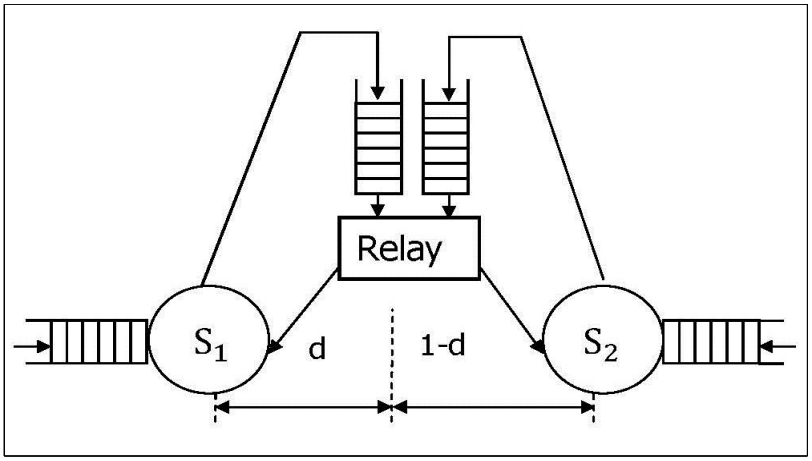

Figure 5.5: The two-way relay system with buffer constraints.

the system has stricter queuing constraints.

5.2 Throughput of Two-Way Relay Systems under Queueing Constraints

5.2.1 System Model

The two-way relay communication link is depicted in Fig. 5.5. In this model, sources $\mathbf{S}_1$ and $\mathbf{S}_2$ wish to exchange information with each other with the help of the intermediate relay node $\mathbf{R}$. We assume that there is no direct link between $\mathbf{S}_1$ and $\mathbf{S}_2$ (which, for instance, holds, if these nodes are sufficiently far apart in distance). Data arriving to sources $\mathbf{S}_1$ and $\mathbf{S}_2$ is initially buffered before transmission to the relay. Similarly, the relay, upon receiving the superimposed signals from the source nodes in the multiple-access phase and decoding the messages, places the decoded data from the sources in two different buffers before broadcasting the superimposed messages back to the source nodes. Both the source and the intermediate relay nodes operate under queuing limitations given in (2.1). The QoS exponents at the source nodes are denoted by θ_{s_j} for $j = 1, 2$. Similarly, the QoS exponents for the two buffers at the

relay are θ_{r,s_j} with $j = 1, 2$. In this section, we assume that the same asymptotic QoS constraints are imposed at the relay buffers, i.e., $\theta_{r,s_1} = \theta_{r,s_2} = \theta_r$. However, we also note that different QoS exponents at the relay buffers can easily be accommodated in the analysis as well.

Since we consider half-duplex relay operation, reception and transmission at the relay occur in non-overlapping intervals. Next, we express the discrete-time input and output relationships in both multiple-access and broadcast phases. In the multiple-access phase, the signal Y_r received at the relay can be expressed as

$$Y_r[i] = g_1[i]X_1[i] + g_2[i]X_2[i] + n_r[i] \tag{5.19}$$

where X_j for $j = 1, 2$ denotes the signal transmitted from source node $\mathbf{S}_j$, and g_j is the fading coefficient between the nodes $\mathbf{S}_j$ and $\mathbf{R}$. The decoded information from each source is stored in a separate buffer at the relay. In the broadcast phase, the signal Y_j received at source node $\mathbf{S}_j$ is given by

$$Y_j[i] = g_j[i]X_r[i] + n_j[i], j = 1, 2 \tag{5.20}$$

where X_r represents the signal sent from the relay node. The inputs are subject to individual average energy constraints $\mathbb{E}\{|X_j|^2\} \leq \bar{P}_j/B$ for $j = 1, 2$ and $\mathbb{E}\{|X_r|^2\} \leq \bar{P}_r/B$, where B is the bandwidth in the system. Assuming that the symbol rate is B complex symbols per second, we can easily see that the symbol energy constraint of $\bar{P}_k/B$ for $k = 1, 2, r$ implies that the nodes have a power constraint of $\bar{P}_k$. We assume that the fading coefficients $g_j, j = \{1, 2\}$ are jointly stationary and ergodic discrete-time processes, and we denote the magnitude-square of the fading coefficients by $z_j[i] = |g_j[i]|^2$. Above, in the channel input-output relationships, the noise component $n_k[i]$ are zero-mean, circularly symmetric, complex Gaussian random variables with variance $\mathbb{E}\{|n_k[i]|^2\} = N_k$ for $k = \{1, 2, r\}$. The additive Gaussian noise samples

$\{n_j[i]\}$ are assumed to form an i.i.d. sequence. We denote the signal-to-noise ratios as $\mathrm{SNR}_k = \frac{\bar{P}_k}{N_k B}$ where $k = 1, 2, r$.

Finally, we introduce two system parameters: We assume that the fraction of time allocated to the multiple-access phase, in which source nodes transmit to the relay, is $\tau \in (0, 1)$. Hence, broadcast phase occurs in the remaining fraction $(1 - \tau)$ of the time. In the broadcast phase, fraction of power allocated to data transmission to node $\mathbf{S}_1$ is denoted by $\rho \in (0, 1)$. Therefore, data intended for $\mathbf{S}_2$ is sent using $(1 - \rho)$ fraction of the relay power.

5.2.2 Throughput in Two-Way Relay Systems

5.2.2.1 Instantaneous Transmission Rates

We initially describe the instantaneous transmission or equivalently service rates at the source and relay nodes. Let us first consider the multiple-access phase in which $\mathbf{S}_1$ and $\mathbf{S}_2$ simultaneously transmit to the relay. Assume that the decoding order is fixed at the relay and is, for instance, given by $\{1, 2\}$, i.e., the information sent from user 1 is decoded first. Then, the maximum instantaneous achievable transmission rates at $\mathbf{S}_1$ and $\mathbf{S}_2$ are given, respectively, by

$$\begin{aligned} R_{s_1,r} &= B \log_2 \left(1 + \frac{\mathrm{SNR}_1 z_1}{1 + \mathrm{SNR}_2 z_2} \right), \\ R_{s_2,r} &= B \log_2 \left(1 + \mathrm{SNR}_2 z_2 \right). \end{aligned} \tag{5.21}$$

If, on the other hand, the decoding order is $\{2, 1\}$, we have

$$\begin{aligned} R_{s_1,r} &= B \log_2 \left(1 + \mathrm{SNR}_1 z_1 \right), \\ R_{s_2,r} &= B \log_2 \left(1 + \frac{\mathrm{SNR}_2 z_2}{1 + \mathrm{SNR}_1 z_1} \right). \end{aligned} \tag{5.22}$$

In the broadcast phase, we assume that the relay transmits the superimposition

of the signals intended for the source nodes, i.e., we have

$$X_r = X_{r,s_1} + X_{r,s_2} \tag{5.23}$$

with energies $\mathbb{E}\{|X_{r,s_1}|^2\} \leq \rho\bar{P}_r/B$ and $\mathbb{E}\{|X_r|^2\} \leq (1-\rho)\bar{P}_r/B$. Above, X_{r,s_j} is the signal intended for $\mathbf{S}_j$. We assume that source nodes know the channel gains and their own signals forwarded from the relay, i.e., $\mathbf{S}_1$ knows X_{r,s_2} and $\mathbf{S}_2$ knows X_{r,s_1}. Equipped with such knowledge, source nodes can eliminate the self-interference terms in the received signals in the broadcast phase and obtain

$$\breve{Y}_j[i] = g_j[i]X_{r,s_j}[i] + n_j[i], j = 1, 2. \tag{5.24}$$

With these assumptions, the instantaneous service rates for the two buffers at the relay become

$$\begin{aligned} R_{r,s_1} &= B\log_2\left(1 + \rho\mathrm{SNR}_r z_1\right) \\ R_{r,s_2} &= B\log_2\left(1 + (1-\rho)\mathrm{SNR}_r z_2\right). \end{aligned} \tag{5.25}$$

5.2.2.2 Stability Conditions

In this subsection, we discuss the conditions required to ensure stability in the relay buffers, which experience random arrivals and random departures. In particular, we investigate the conditions imposed on the parameters τ and ρ. As mentioned in Section 5.1.2, the stability condition is guaranteed by (2.25) at the source node.

Assume, without loss of generality, that the relay employs the decoding order $\{1, 2\}$. For the stability of the relay buffer storing data intended for $\mathbf{S}_2$, average arrival rate should be smaller than the average departure rate, i.e., we need to satisfy

$$\tau\mathbb{E}\left\{R_{s_1,r}\right\} \leq (1-\tau)\mathbb{E}\left\{R_{r,s_2}\right\}, \tag{5.26}$$

that is,

$$\tau\mathbb{E}\left\{B\log_2\left(1+\frac{\mathrm{SNR}_1 z_1}{1+\mathrm{SNR}_2 z_2}\right)\right\} \leq (1-\tau)\mathbb{E}\left\{B\log_2\left(1+(1-\rho)\mathrm{SNR}_r z_2\right)\right\}. \quad (5.27)$$

Similarly, for the stability of the relay buffer storing data intended for $\mathbf{S}_1$, we should have

$$\tau\mathbb{E}\left\{B\log_2\left(1+\mathrm{SNR}_2 z_2\right)\right\} \leq (1-\tau)\mathbb{E}\left\{B\log_2\left(1+\rho\mathrm{SNR}_r z_1\right)\right\}. \quad (5.28)$$

Hence, we need to identify (τ,ρ) pairs satisfying both (5.27) and (5.28). Note that as ρ increases, the right-hand-side (RHS) of (5.27) decreases, and hence τ must decrease to compensate the loss incurred by ρ. Indeed, when $\rho=1$, we should have $\tau=0$. On the other hand, when $\rho=0$, we have

$$\tau \leq \frac{\mathbb{E}\{\log_2(1+\mathrm{SNR}_r z_2)\}}{\mathbb{E}\left\{\log_2\left(1+\frac{\mathrm{SNR}_1 z_1}{1+\mathrm{SNR}_2 z_2}\right)\right\}+\mathbb{E}\{\log_2(1+\mathrm{SNR}_r z_2)\}}. \quad (5.29)$$

At the same time, the RHS of (5.28) increases with increasing ρ, so τ must increase to satisfy (5.28). When $\rho=0$, we need to set $\tau=0$. When $\rho=1$, we have

$$\tau \leq \frac{\mathbb{E}\{\log_2(1+\mathrm{SNR}_r z_1)\}}{\mathbb{E}\{\log_2\left(1+\mathrm{SNR}_2 z_2\right)\}+\mathbb{E}\{\log_2(1+\mathrm{SNR}_r z_1)\}}. \quad (5.30)$$

From the above observations, we conclude that if we plot the maximum value of τ as a function of ρ, the $\tau(\rho)$ curve, which satisfies (5.27), is a decreasing curve with end point at $\tau(1)=0$. The $\tau(\rho)$ curve, which satisfies (5.28), is an increasing curve, starting at $\tau(0)=0$. Therefore, there exists a crossing point of the two curves, where both (5.27) and (5.28) are satisfied. Let (ρ_1^*,τ_1^*) represent the intersection point. Now, the maximum value of τ, for which the queues are stable, is given by τ_1^*.

Finally, note that a similar discussion follows if the decoding order is $\{2,1\}$.

5.2.2.3 Throughput Region under Statistical QoS Constraints

Denote the region of (τ, ρ) pairs, for which the buffers are stable, by $\mathcal{W}_1$. It is clear from the descriptions in the previous subsection that $\mathcal{W}_1$ is well-defined and non-empty. The following definition characterizes the throughput region in two-way relay channels in the presence of statistical queuing constraints both at the source nodes and the relay.

In order to simplify the analysis, we henceforth consider block-fading channels in which the fading stays constant over a certain duration and then changes independently. Under block-fading assumption, asymptotic LMGF expressions of service processes simplify to

$$\Lambda_C(\theta) = \lim_{n\to\infty} \frac{\log \mathbb{E}\{e^{\theta \sum_{i=1}^{n} c[i]}\}}{n} = \log \mathbb{E}\{e^{\theta c[i]}\}. \tag{5.31}$$

Following the analysis given in Section 2.3, the maximum arrival rates for given $(\tau, \rho) \in \mathcal{W}_1$ can be expressed as

$$\mathrm{R}_1 = \begin{cases} \min\Big\{ -\frac{1}{\theta_{s_1}} \log \mathbb{E}_{z_1}\{e^{-\tau\theta_{s_1} R_{s_1,r}}\}, -\frac{1}{\theta_r} \log \mathbb{E}\{e^{-(1-\tau)\theta_r R_{r,s_2}}\}\Big\} & \theta_r \le \theta_{s_1} \\ \min\Big\{ -\frac{1}{\theta_{s_1}} \log \mathbb{E}_{z_1}\{e^{-\tau\theta_{s_1} R_{s_1,r}}\}, & \\ \qquad -\frac{1}{\theta_{s_1}}\big(\log \mathbb{E}\{e^{-(1-\tau)\theta_r R_{r,s_2}}\} + \log \mathbb{E}\{e^{\tau(\theta_r-\theta_{s_1}) R_{s_1,r}}\}\big)\Big\} & \theta_r > \theta_{s_1} \end{cases} \tag{5.32}$$

and

$$\mathrm{R}_2 = \begin{cases} \min\Big\{ -\frac{1}{\theta_{s_2}} \log \mathbb{E}_{z_2}\{e^{-\tau\theta_{s_2} R_{s_2,r}}\}, -\frac{1}{\theta_r} \log \mathbb{E}\{e^{-(1-\tau)\theta_r R_{r,s_1}}\}\Big\} & \theta_r \le \theta_{s_2} \\ \min\Big\{ -\frac{1}{\theta_{s_2}} \log \mathbb{E}_{z_2}\{e^{-\tau\theta_{s_2} R_{s_2,r}}\}, & \\ \qquad -\frac{1}{\theta_{s_2}}\big(\log \mathbb{E}\{e^{-(1-\tau)\theta_r R_{r,s_2}}\} + \log \mathbb{E}\{e^{\tau(\theta_r-\theta_2) R_{S_2,r}}\}\big)\Big\} & \theta_r > \theta_{s_2}. \end{cases} \tag{5.33}$$

The arrival rates in (5.32) and (5.33) can further be optimized over all $(\tau, \rho) \in \mathcal{W}_1$ as

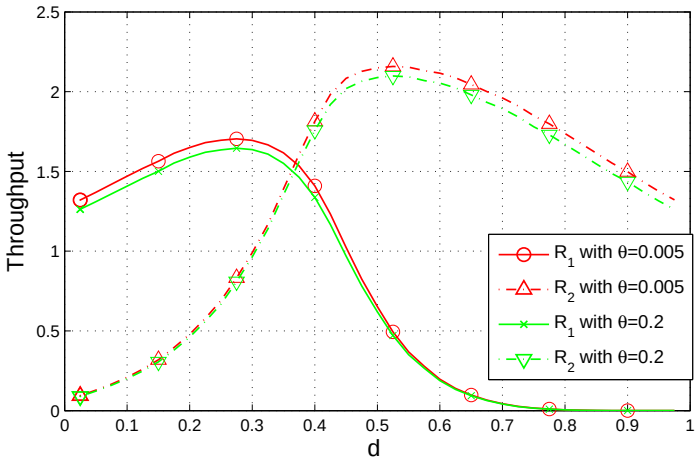

Figure 5.6: Maximum arrival rates R_1 and R_2 vs. d.

will be done in the numerical results below.

5.2.3 Numerical Results

In this subsection, we provide numerical results. We consider a simple scenario in which the sources and relay are located on a straight line. The distance between the two sources $\mathbf{S}_1$ and $\mathbf{S}_2$ has been normalized to 1, and $d \in (0,1)$ is the distance between $\mathbf{S}_1$ and the relay $\mathbf{R}$. Therefore, distance between $\mathbf{S}_2$ and $\mathbf{R}$ is $(1-d)$. We assume that the fading magnitude-squares z_1 and z_2 are independent exponential random variables with means $\mathbb{E}\{z_1\} = \frac{1}{d^\alpha}$ and $\mathbb{E}\{z_2\} = \frac{1}{(1-d)^\alpha}$. We set the path-loss exponent to $\alpha = 4$. Unless specified otherwise, we assume $\theta_{s_1} = \theta_{s_2} = \theta_{r,s_1} = \theta_{r,s_2} = \theta$ and the decoding order at the relay is $\{1,2\}$.

In Fig. 5.6, we plot the maximum arrival rates R_1 and R_2 at $\mathbf{S}_1$ and $\mathbf{S}_2$, respectively, as a function of the distance d for two different QoS exponents when $\mathrm{SNR}_1 = \mathrm{SNR}_2 = \mathrm{SNR}_r = 2$. Fig. 5.7 provides the corresponding optimal (τ^*, ρ^*) values as a function of d. We notice that if the relay is very close to one of the sources, the fraction of time allocated to the multiple-access phase, τ, diminishes due to the fact that downlink channel between the relay and the far-away source node becomes

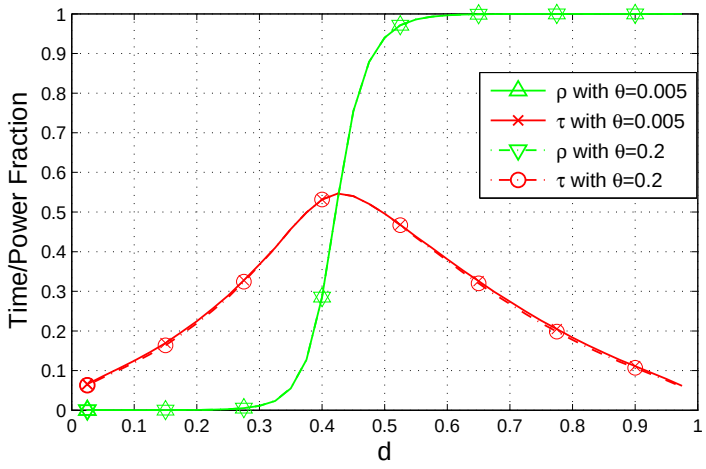

Figure 5.7: Optimal power fraction ρ^* and time fraction τ^* vs. d.

the bottleneck in the information exchange and more time needs to be allocated to relay broadcasting to avoid buffer overflows and/or instability. As a result, we see in Fig. 5.6 that both R_1 and R_2 start diminishing as d approaches 0 or 1. Another observation is that as d increases from 0.1 to 0.5, the relay approaches $\mathbf{S}_2$ and hence the channel between $\mathbf{S}_2$ and $\mathbf{R}$ improves, leading to larger values of R_2. Therefore, higher arrival rates can be supported at $\mathbf{S}_2$. Interestingly, we notice in Fig. 5.7 that ρ approaches to 1 in this case, meaning that the relay allocates more energy to forwarding data to $\mathbf{S}_1$, which is basically needed to support the higher arrival rates at $\mathbf{S}_2$. Finally, we note in Fig. 5.6 that arrival rates are smaller under more stringent queueing constraints, i.e., when θ is increased from 0.005 to 0.2. On the other hand, (τ^*, ρ^*) remain rather robust as seen in Fig. 5.7.

In the following numerical results, we set $d = 0.38$. In Fig. 5.8, we plot maximum arrival rate R_1 as a function SNR parameters. Expectedly, as SNR_1 and/or SNR_r increases, transmission/service rates from $\mathbf{S}_1$ and $\mathbf{R}$ increase and higher arrival rates at $\mathbf{S}_1$ can be supported. Fig. 5.9 plots R_1 now as a function of SNR_r and SNR_2, the signal-to-noise ratio of source $\mathbf{S}_2$'s transmissions. Note that since decoding order of $\{1, 2\}$ is considered, transmissions from $\mathbf{S}_1$ experience interference proportional to

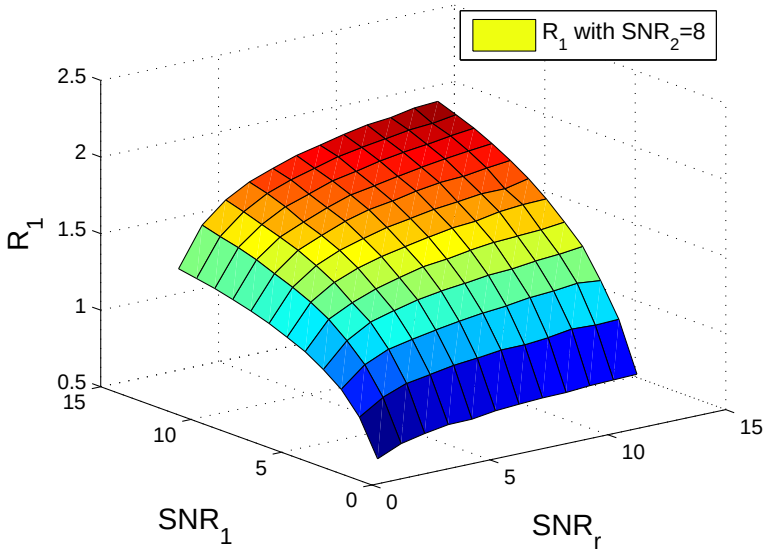

Figure 5.8: Maximum arrival rate R_1 vs. $(\mathrm{SNR}_1, \mathrm{SNR}_r)$ when $\mathrm{SNR}_2 = 8$, $\theta = 0.005$, and $d = 0.38$.

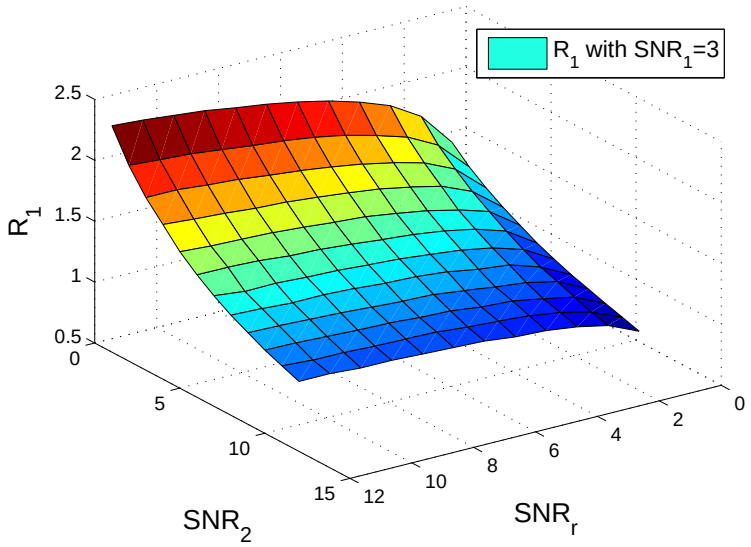

Figure 5.9: Maximum arrival rate R_1 vs. $(\mathrm{SNR}_2, \mathrm{SNR}_r)$ when $\mathrm{SNR}_1 = 3$ and $\theta = 0.005$.

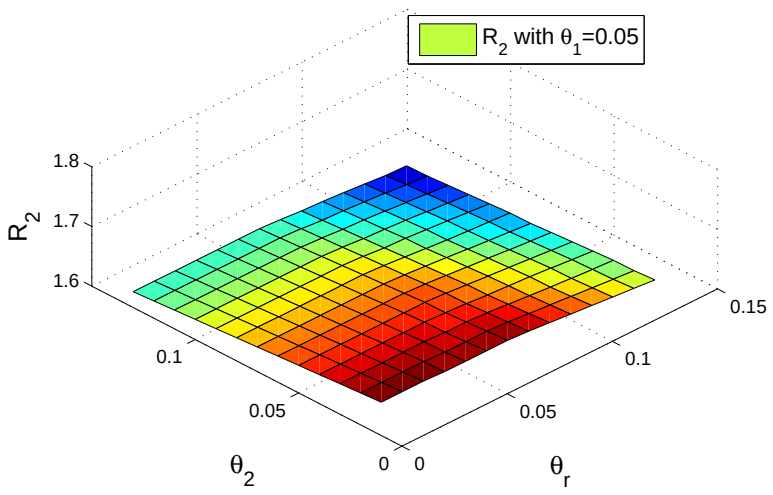

Figure 5.10: Maximum arrival rate R_2 vs. (θ_2, θ_r) when $\theta_1 = 0.05$ and $\text{SNR} = 2$.

SNR_2 as seen in the expression of $R_{s_1,r}$ in (5.21). Therefore, due to this coupling in the multiple-access phase, we see in Fig. 5.9 that R_1 diminishes with increasing SNR_2. As before, increasing SNR_r improves R_1.

In Fig. 5.10, we plot R_2 as a function of the QoS exponents. Note that the higher the QoS exponents, the more stringent the buffer constraints are. Therefore, as demonstrated in the figure, increasing QoS exponents results in reduced arrival rates. Sources basically admit lower-rate arrivals to satisfy more strict buffer constraints.

Until now, we have fixed the decoding order at $\{1, 2\}$ at the relay in the multiple-access phase. In general, varying the decoding order can enlarge the region of arrival rates. In Fig. 5.11, we plot the throughput region, i.e., region of arrival rates (R_1, R_2), achieved by time-sharing between decoding orders $\{1, 2\}$ and $\{2, 1\}$. The boundary of the region is determined by optimizing the resource allocation parameters (τ, ρ). This figure is obtained when $\text{SNR}_r = 4$, $\text{SNR}_1 = \text{SNR}_2 = 2$, $\theta = 0.005$, and $d = 0.5$, hence the relay is midway between the sources.

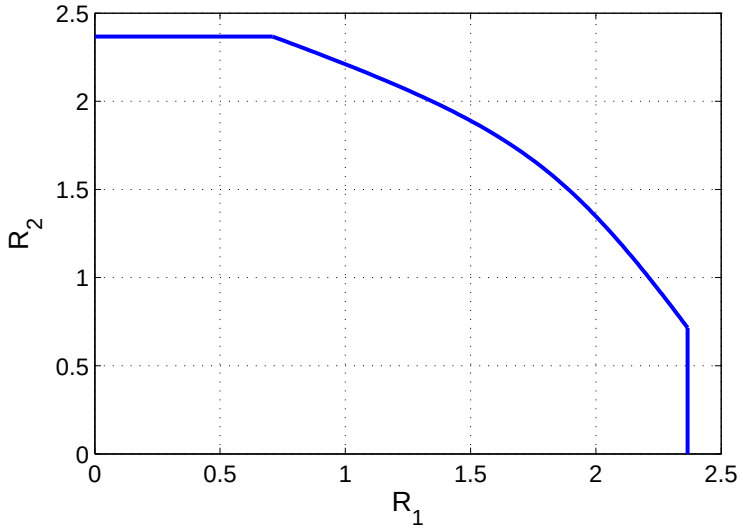

Figure 5.11: Throughput Region achieved with time-sharing between decoding orders $\{1, 2\}$ and $\{2, 1\}$. $\text{SNR}_r = 4, \text{SNR}_1 = \text{SNR}_2 = 2$, $\theta = 0.005$, and $d = 0.5$.

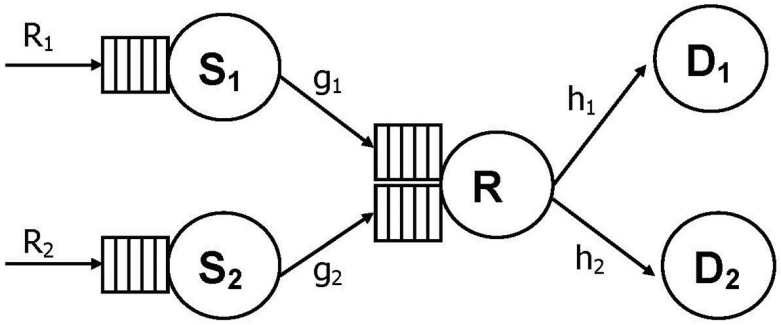

Figure 5.12: The relay network system with buffer constraints.

5.3 Throughput of Multi-Source Multi-Destination Relay Networks with Queuing Constraints

5.3.1 System Model

In this section, we consider a multi-source multi-destination relay network model with two pairs of sources and destinations, as depicted in Fig. 5.12. In this system, two sources $\mathbf{S}_1$ and $\mathbf{S}_2$ send information to their corresponding destinations $\mathbf{D}_1$ and $\mathbf{D}_2$

with the help of an intermediate relay node, and there is no direct link between the source nodes and their destinations. This assumption is accurate if the source and destination nodes are sufficiently far apart in distance. We assume that $\mathbf{D}_j$ only needs the packets coming from source $\mathbf{S}_j$, where $j = 1, 2$. Each source node has a buffer, keeping the packets to be transmitted to the relay node. The arrival rates at source nodes $\mathbf{S}_1$ and $\mathbf{S}_2$ are assumed to be constant, and are denoted as R_1 and R_2 respectively. At the relay node, there are two buffers[3], one for keeping the decoded information coming from source $\mathbf{S}_1$, and the other for the decoded data of $\mathbf{S}_2$.

In our setup, relay node performs decode-and-forward relaying and works in half-duplex mode, and hence it cannot transmit and receive at the same time. The entire transmission process can be divided into two phases, namely multiple-access phase and broadcast phase. In the multiple-access phase, both $\mathbf{S}_1$ and $\mathbf{S}_2$ transmit to the relay node simultaneously through a multiple-access channel. Relay node attempts to decode their messages by using certain decoding orders, and the decoded information bits are stored in their corresponding buffers at the relay. We assume that if fixed-rate transmissions are employed, transmission fails if the rate is greater than the instantaneous capacity of the link for a given decoding strategy at the relay[4].

The received discrete-time signal at the relay node can be expressed as

$$Y_r[i] = g_1[i]X_1[i] + g_2[i]X_2[i] + n_r[i], \tag{5.34}$$

where X_j for $j = 1, 2$ represents the transmitted signal from source node $\mathbf{S}_j$, g_j is the fading coefficient of the $\mathbf{S}_j - \mathbf{R}$ link, and n_r is the additive Gaussian noise at the relay.

[3]In practice, only one physical buffer is sufficient at the relay node to store the received packets from $\mathbf{S}_1$ and $\mathbf{S}_2$. In the analysis, we essentially decompose this physical buffer into two equivalent virtual buffers, in each of which data for only one destination is stored and first-in first-out policy is employed.

[4]It is assumed that errors are detected reliably at the receivers, and when the system employs ARQ protocol, ACK and retransmission request (RQ) packets are assumed to be received with no errors.

In the broadcast phase, relay node forwards information bits to their destinations through a broadcast channel. The received signal at $\mathbf{D}_j$ is

$$Y_j[i] = h_j[i]X_r[i] + n_j[i], \quad j = 1, 2 \tag{5.35}$$

where X_r stands for the transmitted signal from $\mathbf{R}$[5], n_j is the additive Gaussian noise at $\mathbf{D}_j$, and h_j represents the channel fading coefficient of the $\mathbf{R}-\mathbf{D}_j$ link. Magnitude-squares of the fading coefficients in both phases are denoted by $z_j[i] = |g_j[i]|^2$ and $\omega_j[i] = |h_j[i]|^2$, for $j = 1, 2$. In our analysis, we consider block fading and assume that fading coefficients stay constant in one time block, and change independently from block to block. While our analysis is general and applicable to any fading distribution with finite variances, we assume Rayleigh fading in all channels in our numerical analysis.

The transmitted signals are subject to energy constraints given by $\mathbb{E}\{|X_j|^2\} \leq \bar{P}_j/B$ for $j = 1, 2$ and $\mathbb{E}\{|X_r|^2\} \leq \bar{P}_r/B$, where B is the system bandwidth and $\bar{P}_k$ for $k = 1, 2, r$ is the transmit power constraint for the corresponding node. The additive noise terms $n_k[i]$ for $k = 1, 2, r$ are independent, zero-mean, circularly symmetric, complex Gaussian random variables with variances $\mathbb{E}\{|n_k[i]|^2\} = N_0$. Then, signal-to-noise ratios are defined as

$$\text{SNR}_k = \frac{\bar{P}_k}{N_0 B} \tag{5.36}$$

where $k = 1, 2, r$.

Finally, there are three important system parameters: τ, ρ and δ. $\tau \in (0, 1)$ denotes the fraction of time allocated to the multiple-access phase, and hence the fraction of time allocated to the broadcast phase is $1 - \tau$. $\rho \in (0, 1)$ represents the fraction of power allocated by the relay to the transmission of the message intended for $\mathbf{D}_1$, and therefore the fraction of power allocated to the transmission to $\mathbf{D}_2$ is

[5]The signal transmitted from the relay can be written as $X_r = X_{r1} + X_{r2}$, and hence is a combination of X_{r1} and X_{r2}, which are the signals intended for $\mathbf{D}_1$ and $\mathbf{D}_2$, respectively.

$1-\rho$. In the multiple-access phase, relay node decodes the received signal using different decoding orders, and the fraction of time allocated to decoding order $\{1,2\}$ and $\{2,1\}$ at the relay node are denoted by δ and $1-\delta$, respectively. This time sharing strategy between different decoding orders is used only for the case of variable-rate transmissions, performed when CSI is available at all transmitters. For fixed-rate transmission schemes, decoding order is part of the decoding strategy, which is fixed for each node.

In this section, system throughput is characterized by the pair of maximum constant arrival rates R_1 and R_2 that can be supported by the relay network with two pairs of source-destination nodes in the presence of statistical queuing constraints. Detailed discussion about the arrival rates for two-hop channels in the presence of statistical queuing constraints is given in Chapter 2.

Finally, we provide a list of notations together with their descriptions in Table 5.1.

5.3.2 Throughput of the Two-Source Two-Destination Relay Network With Variable Transmission Rates

In this subsection, we study the throughput of the two-source two-destination relay network with variable-rate transmissions. Under the assumption that CSI is available at each transmitter, transmitters adapt their transmission rate to the instantaneous channel conditions, and the departure rates at each buffer are given by the corresponding instantaneous channel capacities. To perform an effective capacity analysis at each node with a buffer, we have to first identify the instantaneous transmission rates as functions of the fading coefficients.

Table 5.1: Table of notations for Section 5.3

Notation	Definition
Y_j	Received signal at relay $\mathbf{R}$ (for $j = r$) or destination $\mathbf{D}_j$ (for $j = 1, 2$).
X_j	Transmitted signal from relay $\mathbf{R}$ (for $j = r$) or source $\mathbf{S}_j$ (for $j = 1, 2$).
g_j	Fading coefficient of the $\mathbf{S}_j - \mathbf{R}$ link.
z_j	Magnitude-square of the fading coefficient g_j.
h_j	Fading coefficient of the $\mathbf{R} - \mathbf{D}_j$ link.
ω_j	Magnitude-square of the fading coefficient h_j.
n_j	Additive Gaussian noise at the relay $\mathbf{R}$(for $j = r$) or destination $\mathbf{D}_j$ (for $j = 1, 2$) with variance N_0.
SNR_j	Signal-to-noise ratio of relay $\mathbf{R}$ (for $j = r$) or source $\mathbf{S}_j$ (for $j = 1, 2$).
θ_j	QoS exponent associated with the buffer constraint at relay $\mathbf{R}$ (for $j = r$) or source $\mathbf{S}_j$ (for $j = 1, 2$).
$\Lambda(\theta)$	LMGF of a departure or arrival process as a function of the QoS exponent θ.
τ	The fraction of time allocated to the multiple-access phase.
ρ	The fraction of power allocated by the relay to the transmission of the message intended for $\mathbf{D}_1$.
δ	The fraction of time allocated to decoding order $\{1, 2\}$ at relay $\mathbf{R}$.
R_j	The maximum constant arrival rate at source $\mathbf{S}_j$ that can be supported under queuing constraints.
$R_{A,B}$	The instantaneous channel capacity of link $\mathbf{A} - \mathbf{B}$.
$r_{A,B}$	The fixed transmission rate of link $\mathbf{A} - \mathbf{B}$ in the fixed rate scheme.

5.3.2.1 Instantaneous Transmission Rates in Multiple User Relay Networks

We initially describe the instantaneous transmission rates of four links. Let us first consider the multiple-access phase in which links $\mathbf{S}_1 - \mathbf{R}$ and $\mathbf{S}_2 - \mathbf{R}$ are active simultaneously. When the decoding order at the relay is given by $\{1, 2\}$, i.e., the information sent from node $\mathbf{S}_1$ is decoded first, and the information sent from node $\mathbf{S}_2$ is decoded after interference cancelation, then the maximum instantaneous achievable rates at $\mathbf{S}_1$ and $\mathbf{S}_2$ are given, respectively, by [32]

$$\begin{cases} R_{\mathbf{S}_1,\mathbf{R}\{1,2\}} & = B\log_2\left(1+\frac{\mathrm{SNR}_1 z_1}{1+\mathrm{SNR}_2 z_2}\right), \\ R_{\mathbf{S}_2,\mathbf{R}\{1,2\}} & = B\log_2\left(1+\mathrm{SNR}_2 z_2\right). \end{cases} \tag{5.37}$$

If the decoding order at the relay node is $\{2, 1\}$, then we have

$$\begin{cases} R_{\mathbf{S}_1,\mathbf{R}\{2,1\}} & = B\log_2\left(1+\mathrm{SNR}_1 z_1\right), \\ R_{\mathbf{S}_2,\mathbf{R}\{2,1\}} & = B\log_2\left(1+\frac{\mathrm{SNR}_2 z_2}{1+\mathrm{SNR}_1 z_1}\right). \end{cases} \tag{5.38}$$

If we perform time-sharing between two decoding orders with parameter δ, then the rates of links $\mathbf{S}_1 - \mathbf{R}$ and $\mathbf{S}_2 - \mathbf{R}$ are characterized by (5.37) in δ fraction of the time, and the rates are characterized by (5.38) rest of the time. Overall, the transmission rates between the source nodes and the relay node can be expressed as

$$R_{\mathbf{S}_j,\mathbf{R}} = \delta R_{\mathbf{S}_j,\mathbf{R}\{1,2\}} + (1-\delta) R_{\mathbf{S}_j,\mathbf{R}\{2,1\}}, \tag{5.39}$$

for $j = 1, 2$.

In the broadcast phase, relay node forwards packets to their corresponding destinations. In this phase, only links $\mathbf{R} - \mathbf{D}_1$ and $\mathbf{R} - \mathbf{D}_2$ are active. When the channel conditions are available at the relay node and destinations, the decoding order are

decided by the relationship between ω_1 and ω_2, and the instantaneous transmission rates are given by [88] [34]

$$\begin{cases} R_{\mathbf{R},\mathbf{D}_1} &= B\log_2\left(1+\frac{\rho \mathrm{SNR}_r \omega_1}{1+(1-\rho)\mathrm{SNR}_r\omega_1 \mathbb{1}\{\omega_1<\omega_2\}}\right), \\ R_{\mathbf{R},\mathbf{D}_2} &= B\log_2\left(1+\frac{(1-\rho)\mathrm{SNR}_r \omega_2}{1+\rho\mathrm{SNR}_r\omega_2 \mathbb{1}\{\omega_2<\omega_1\}}\right) \end{cases} \tag{5.40}$$

where $\mathbb{1}\{\bullet\}$ is indicator function.

5.3.2.2 Stability Conditions

With the expressions of the instantaneous rates for both the multiple-access channel and broadcast channel described above, we can characterize the stability region in the $\rho-\tau-\delta$ space. Stability at the source buffers is ensured by requiring the arrival rates to satisfy (2.25), which actually leads to compliance with the stricter condition that the tail distribution of the buffer length decays exponentially fast. The stability conditions at the relay node requires the average arrival rate to be less than or equal to the average departure rate at each buffer in the relay. Hence, the stability conditions can be formulated as

$$\begin{cases} \tau\left(\delta\mathbb{E}\{R_{\mathbf{S}_1,\mathbf{R}\{1,2\}}\}+(1-\delta)\mathbb{E}\{R_{\mathbf{S}_1,\mathbf{R}\{2,1\}}\}\right) \leq (1-\tau)\mathbb{E}\{R_{\mathbf{R},\mathbf{D}_1}\}, \\ \tau\left(\delta\mathbb{E}\{R_{\mathbf{S}_2,\mathbf{R}\{1,2\}}\}+(1-\delta)\mathbb{E}\{R_{\mathbf{S}_2,\mathbf{R}\{2,1\}}\}\right) \leq (1-\tau)\mathbb{E}\{R_{\mathbf{R},\mathbf{D}_2}\}. \end{cases} \tag{5.41}$$

Plugging (5.37), (5.38), and (5.40) into (5.41), we obtain

$$\begin{cases} (1-\tau)\mathbb{E}\left\{B\log_2\left(1+\frac{\rho \mathrm{SNR}_r\omega_1}{1+(1-\rho)\mathrm{SNR}_r\omega_1\mathbb{1}\{\omega_1<\omega_2\}}\right)\right\} \geq \\ \qquad \tau\left(\delta\mathbb{E}\{B\log_2\left(1+\frac{\mathrm{SNR}_1 z_1}{1+\mathrm{SNR}_2 z_2}\right)\}+(1-\delta)\mathbb{E}\{B\log_2\left(1+\mathrm{SNR}_1 z_1\right)\}\right), \\ (1-\tau)\mathbb{E}\left\{B\log_2\left(1+\frac{(1-\rho)\mathrm{SNR}_r\omega_2}{1+\rho\mathrm{SNR}_r\omega_2\mathbb{1}\{\omega_2<\omega_1\}}\right)\right\} \geq \\ \qquad \tau\left(\delta\mathbb{E}\{B\log_2\left(1+\mathrm{SNR}_2 z_2\right)\}+(1-\delta)\mathbb{E}\{B\log_2\left(1+\frac{\mathrm{SNR}_2 z_2}{1+\mathrm{SNR}_1 z_1}\right)\}\right). \end{cases} \tag{5.42}$$

All feasible (ρ,τ,δ)-tuples satisfying the inequalities in (5.42) form the stability region in the $\rho-\tau-\delta$ space. Hence, we formally define the the stability region Ξ in the $\rho-\tau-\delta$ space as

$$\Xi=\{(\rho,\tau,\delta)|\rho,\ \tau,\ \text{and } \delta \text{ that satisfy (5.42)}\}. \tag{5.43}$$

For a certain time sharing scheme at the relay node with fixed δ, since τ is the time fraction allocated to the multiple-access phase, lower τ value is more likely to satisfy the stability condition, and the two inequalities in (5.42) provide two upper bounds on τ as functions of ρ. The power allocation parameter ρ has a different influence on these two phases. With more power allocated to transmission to $\mathbf{D}_i$ in the broadcast phase, the corresponding buffer in the relay can support a higher τ value while satisfying the stability constraint.

5.3.2.3 Throughput Region under Statistical Queuing Constraints

As noted before, for a certain parameter setting, the system throughput is defined as the pair of constant arrival rates R_1 and R_2, which can be supported by two-hop links $\mathbf{S}_1-\mathbf{D}_1$ and $\mathbf{S}_2-\mathbf{D}_2$, respectively, under queuing constraints. Since stability is a prerequisite for effective capacity analysis, our system throughput is only defined with parameter values included in the stability region. For those parameter settings

outside the stability region, at least one of the queuing constraints cannot be satisfied, and the system throughput is set to zero. Using the results in the previous subsection, to comply with queuing constraints at all nodes, R_j for $j = 1, 2$ has to satisfy (2.25) and (2.26) simultaneously, which leads to the following characterization of the system throughput.

Theorem 16 *For any parameter setting $\{\tau, \rho, \delta\}$ that satisfies the stability conditions, the maximum constant arrival rate R_j, which can be supported at source node $\mathbf{S}_j$ for $j = 1, 2$ in the presence of all queuing constraints, is given by*

$$\mathrm{R}_j = \begin{cases} \min\left\{ -\frac{1}{\theta_j}\log(\mathbb{E}\{e^{-\theta_j \tau R_{\mathbf{S}_j,\mathbf{R}}}\}), -\frac{1}{\theta_r}\log(\mathbb{E}\{e^{-\theta_r(1-\tau)R_{\mathbf{R},\mathbf{D}_j}}\}) \right\} & \theta_r \leq \theta_j \\ \min\left\{ -\frac{1}{\theta_j}\log(\mathbb{E}\{e^{-\theta_j \tau R_{\mathbf{S}_j,\mathbf{R}}}\}), -\frac{1}{\theta_j}\left(\log(\mathbb{E}\{e^{-\theta_r(1-\tau)R_{\mathbf{R},\mathbf{D}_j}}\}) + \log(\mathbb{E}\{e^{(\theta_r-\theta_j)\tau R_{\mathbf{S}_j,\mathbf{R}}}\}) \right) \right\} & \theta_r > \theta_j, \end{cases} \tag{5.44}$$

Proof 4 *See Appendix A.10.*

Following this characterization, some properties of the system throughput are shown in the next subsection.

5.3.2.4 Properties of the System Throughput under Queuing Constraints

In the previous subsection, we have characterized the throughput of the two-source two-destination relay network. Based on (5.44), we next analyze the behavior of the throughput in the parameter space, and establish several convexity properties, which can lead to simplifications in parameter optimization.

Theorem 17 *In the stability region, for a given $\tau - \rho$ pair, the maximum arrival rates R_1, R_2 and the sum rate $\mathrm{R}_1 + \mathrm{R}_2$ are concave over the time sharing parameter δ between different decoding orders at the relay.*

Proof: See Appendix A.11.

Theorem 17 indicates that there exists a globally optimal time sharing parameter for the two possible decoding orders at the relay, which can be determined via convex optimization methods. Similarly, the system throughput functions are also concave functions of τ, which is the parameter for time allocation between the multiple-access and broadcast phases.

Theorem 18 *In the stability region, for given power allocation parameter ρ and time-sharing parameter δ, the maximum arrival rates R_1, R_2 and the sum rate $\mathrm{R}_1+\mathrm{R}_2$ are concave over the time allocation parameter τ.*

Proof: See Appendix A.12.

Using these results, we can maximize the system throughput over δ and τ under stability constraints by employing efficient convex optimization methods.

5.3.2.5 Throughput of Multi-Source Multi-Destination Networks

Our analysis in this subsection has primarily considered a two-source two-destination relay network. However, using similar techniques and approach, we can extend the analysis to multi-source multi-destination networks. For instance, let us consider a multiple-user model in which N sources send information to their corresponding destinations with the help of a relay node. The magnitude-squares of the fading coefficients of links $\mathbf{S}_j - \mathbf{R}$ and $\mathbf{R} - \mathbf{D}_j$ are represented by z_j and ω_j, respectively.

Compared with the two-user model, adding more users only increases the dimension of the parameter space while the analytical methods and results essentially remain the same. In this multi-user setting, system parameters $\boldsymbol{\rho} = (\rho_1, \rho_2, \cdots, \rho_N)$ and $\boldsymbol{\delta} = (\delta_1, \delta_2, \cdots, \delta_{N!})$ become vectors, while the time allocation parameter τ is still a scalar.

In the multiple-access phase, we denote the k^{th} decoding order at the relay as $\boldsymbol{\pi}_k = \{k_1, k_2, \cdots, k_N\}$, which is a permutation of $\{1, 2, \cdots, N\}$. With this decoding

order, the instantaneous rate of the $\mathbf{S}_{k_i} - \mathbf{R}$ link is characterized by

$$\mathbf{R}_{\mathbf{S}_{k_i},\mathbf{R},\boldsymbol{\pi}_k} = B \log_2 \left(1 + \frac{\mathrm{SNR}_{k_i} z_{k_i}}{1 + \sum_{j=i+1}^{N} \mathrm{SNR}_{k_j} z_{k_j}} \right). \tag{5.45}$$

Given a time sharing vector $\boldsymbol{\delta} = (\delta_1, \delta_2, \cdots, \delta_{N!})$, the rate of the $\mathbf{S}_j - \mathbf{R}$ link is given by

$$\mathbf{R}_{\mathbf{S}_j,\mathbf{R}} = \sum_{k=1}^{N!} \delta_k \mathbf{R}_{\mathbf{S}_j,\mathbf{R},\boldsymbol{\pi}_k}, \tag{5.46}$$

for $j = 1, 2, \cdots, N$. For the broadcast channel, the instantaneous rate is given by

$$R_{\mathbf{R},\mathbf{D}_j} = B \log_2 \left(1 + \frac{\rho_j \mathrm{SNR}_r \omega_j}{1 + \sum_{l=1, l\neq j}^{N} \rho_l \mathrm{SNR}_r \omega_j \mathbb{1}\{\omega_j < \omega_l\}} \right), \tag{5.47}$$

for $j = 1, 2, \cdots, N$. Similarly, the stability region in the parameter space is defined as

$$\begin{aligned} \Xi = \Bigg\{ & (\tau, \rho_1, \cdots, \rho_N, \delta_1, \cdots, \delta_{N!}) | \tau, \boldsymbol{\rho} \text{ and } \boldsymbol{\delta} \text{ that satisfy } \tau \mathbb{E}\{\mathbf{R}_{\mathbf{S}_j,\mathbf{R}}\} \leq (1-\tau)\mathbb{E}\{R_{\mathbf{R},\mathbf{D}_j}\}, \\ & \sum_{i=1}^{N} \rho_i = 1 \text{ and } \sum_{i=1}^{N!} \delta_i = 1, \text{for all } j = 1, 2, \cdots, N \Bigg\}. \end{aligned} \tag{5.48}$$

In this multiple-user setting, the dimension of the parameter space becomes much higher than that in the two-user model. For a set of parameters that guarantee the stability conditions, the throughput of the $\mathbf{S}_j - \mathbf{D}_j$ link under queuing constraints satisfies (2.25) and (2.26) simultaneously, and hence is given by (5.44), for $j = 1, 2, \cdots, N$, with the instantaneous rate expressions provided above.

5.3.2.6 Numerical Results

In this subsection, numerical results are provided to further analyze the throughput of the two-source two-destination relay network with variable transmission rates. Our numerical results are based on (5.44).

In order to verify our analysis, we have conducted Monte Carlo simulations in which we have generated arrivals to the buffer at constant rates determined by our theoretical characterization in (5.44) and also generated random (Rayleigh) fading coefficients to simulate the wireless channel and random transmission rates. We have tracked the buffer occupancy and overflows for different threshold levels. We plot the simulated logarithmic buffer overflow probabilities as functions of the overflow threshold $q_{\max}$ in Figs. 5.13 and 5.14. In each simulation, we generate 5×10^7 time blocks to estimate the buffer overflow probability, and repeat each simulation 1000 times to evaluate the averages. We set the queuing constraints as $\theta_1 = \theta_2 = \theta_r = 0.1$, and the constant arrival rates at nodes $\mathbf{S}_1$ and $\mathbf{S}_2$ are determined from (5.44). In both figures, $\mathbb{E}\{z_j\} = \mathbb{E}\{\omega_j\} = 1$, $\tau = \rho = \delta = 0.5$, $\text{SNR}_1 = \text{SNR}_2 = 10$dB. In Fig. 5.13, we set $\text{SNR}_r = 30$dB. Note that $\log \Pr\{Q \geq q_{\max}\} \approx \log \gamma - \theta q_{\max}$, the slope of the logarithmic overflow probability is expected to be proportional to $-\theta$. Although large $q_{\max}$ is required, our simulation results show that $\log \Pr\{Q \geq q_{\max}\}$ can be approximated as a linear function of $q_{\max}$ starting from relatively small $q_{\max}$. In Fig. 5.13, the slopes of the logarithmic overflow probabilities at buffers in $\mathbf{S}_1$ and $\mathbf{S}_2$ are -0.100 and -0.099, respectively. This implies that simulation results demonstrate perfect agreement with the analysis and the arrival rates given by (5.44) fit the queuing constraints at $\mathbf{S}_1$ and $\mathbf{S}_2$ exactly. We also observe that the logarithmic overflow probabilities of the two relay buffers decay faster with steeper slopes than our requirement of $\theta_r = 0.1$. In this specific example, due to relay having a relatively large transmit power, the system performance is mainly decided by the multiple-access phase, which is the bottleneck of the system. Although the relay can potentially

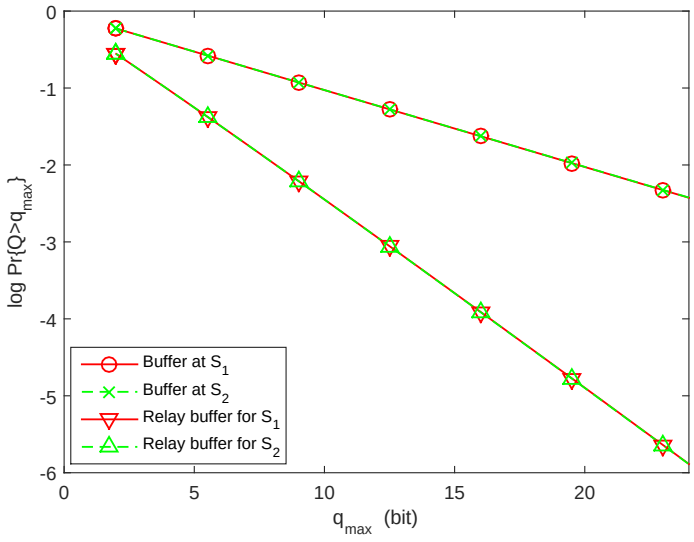

Figure 5.13: Logarithmic buffer overflow probability vs. buffer overflow threshold.

support higher R_1 and R_2, this is not allowed by the multiple-access phase. As we reduce the transmission power of the relay node, the system bottleneck shifts to the broadcast phase and the situation is reversed. In Fig. 5.14, we reduce SNR_r to 27.5dB. Now, the arrival rates given by (5.44) fit the queuing constraints at the relay exactly, and the corresponding slopes for the two relay buffers are -0.098 and -0.097, respectively. On the other hand, the decays of the overflow probabilities at the source nodes are faster, meaning that sources can potentially support higher arrival rates but this leads to the violation of the overflow constraints at the relay buffers and is therefore not allowed. Overall, these simulation results, while confirming the analysis, also interestingly unveil the critical interactions between the queues and buffer constraints.

For the rest numerical results in this subsection, we consider Rayleigh fading and we set $\text{SNR}_1 = \text{SNR}_2 = 3$ dB and $\text{SNR}_r = 6$ dB. Fig. 5.15 shows the influence of the position of the relay node for different θ values. We assume a symmetric model, in which $\theta_1 = \theta_2 = \theta_r$, and $Dist_{\mathbf{S}_1,\mathbf{R}} = Dist_{\mathbf{S}_2,\mathbf{R}}$ and $Dist_{\mathbf{R},\mathbf{D}_1} = Dist_{\mathbf{R},\mathbf{D}_2}$, where $Dist_{A,B}$ stands for the distance between A and B. The overall distance $D =$

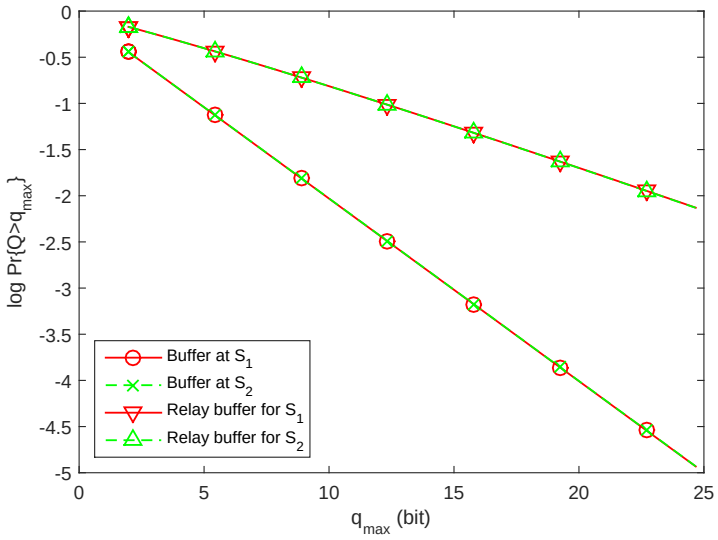

Figure 5.14: Logarithmic buffer overflow probability vs. buffer overflow threshold.

$Dist_{\mathbf{S}_1,\mathbf{R}} + Dist_{\mathbf{R},\mathbf{D}_1} = Dist_{\mathbf{S}_2,\mathbf{R}} + Dist_{\mathbf{R},\mathbf{D}_2} = 2$, and the position parameter $d = \frac{Dist_{\mathbf{S}_1,\mathbf{R}}}{D} = \frac{Dist_{\mathbf{S}_2,\mathbf{R}}}{D}$. Obviously, $d \in [0,1]$, and the smaller value of d indicates that relay is closer to the source. Path loss as a function of distance is incorporated into the statistics of fading powers, and hence, we have $\mathbb{E}\{z_j\} = (\frac{1}{D\,d})^4$ and $\mathbb{E}\{\omega_j\} = (\frac{1}{D(1-d)})^4$ for $j = 1, 2$. In the figure, we see that the maximum sum rate $\mathrm{R}_1 + \mathrm{R}_2$ is achieved when d is close to 0.5, which means that it is better to place the relay in the middle between the source and destination in this symmetric setting. When the relay is close to the source nodes, the channels between the relay and destinations deteriorate and the overall throughput is limited by the broadcast links. Similarly, the multiple-access links become the bottleneck when d is close to 1. Also, we observe that the system throughput decreases when θ increases due to tighter queuing constraints. This occurs because when θ is small, the effective capacity is closer to the Shannon capacity, and as θ increases, effective capacity diminishes and approaches the zero-outage capacity (which is, for instance, zero in Rayleigh fading).

In Fig. 5.16, we consider an asymmetric scenario in terms of QoS exponents, and again plot sum rate vs. relay location parameter d. We fix $\rho = \delta = 0.5$ and

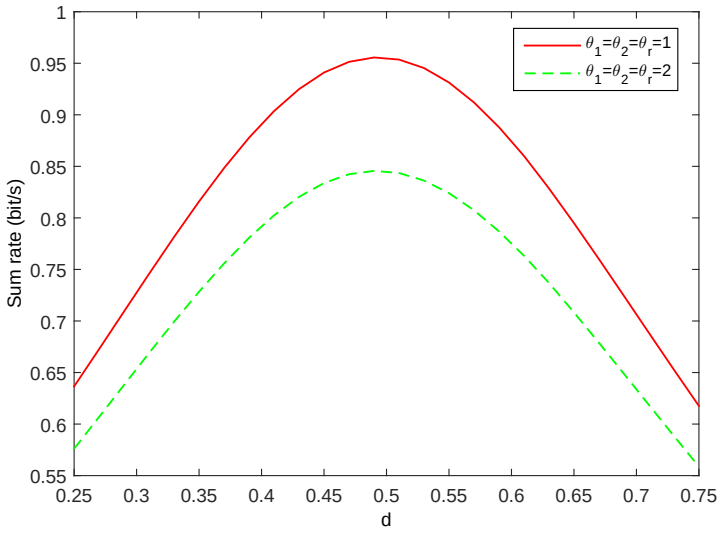

Figure 5.15: The sum rate vs. relay location parameter.

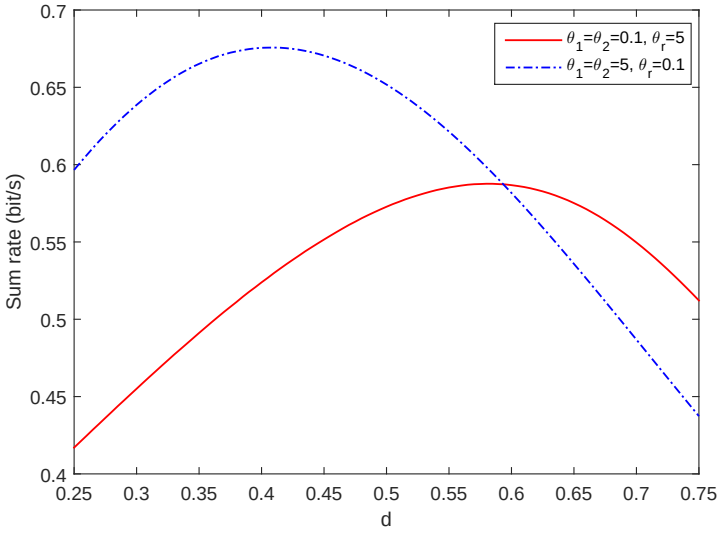

Figure 5.16: The sum rate vs. relay location parameter.

determine the optimal value of τ for each given d. When $\theta_r = 5$, $\theta_1 = \theta_2 = 0.1$, the maximum sum rate is achieved at $d = 0.58$. In this case, relay should be placed closer to the destinations to support more stringent queuing constraints at the relay. On the other hand, when $\theta_1 = \theta_2 = 5$, $\theta_r = 0.1$, the optimal position for the relay is at $d = 0.41$. Hence, the relay needs to be closer to the source nodes to support their stricter queuing constraints. These observations indicate the sensitivity of optimal relay placement to different QoS requirements.

Figs. 5.17 and 5.18 demonstrate the concavity[6] of the sum rate with respect to τ and δ, respectively, when the parameter values are in the stability region. In these two figures, $\theta_1 = \theta_2 = \theta_r = 1$, and $\mathbb{E}\{z_j\} = \mathbb{E}\{\omega_j\} = 1$. In Fig. 5.17, the sum rate curves first increase with τ, and then decrease very fast after reaching the maximum sum rate. As τ exceeds a threshold, the sum rates drop to 0, because stability conditions are violated beyond this threshold. In Fig. 5.18, the sum rate curves are concave with respect to the decoding parameter δ, and the optimal δ values which maximize the sum rate are all close to 0.5. In this case, relay allocates time to two decoding orders equally. However, note that these results are again for a symmetric scenario in which all QoS exponents are the same. In Fig. 5.19, we address a heterogeneous setting in terms of QoS exponents. For instance, when $\theta_1 = \theta_r = 1$ and $\theta_2 = 0.1$, the optimal value of δ is 1. Hence, sum rate is maximized when the decoding order at the relay is always fixed as $\{1, 2\}$, i.e., relay initially decodes data arriving from source $\mathbf{S}_1$ in the presence of interfering signal of $\mathbf{S}_2$. The underlying reason for this result is the following. Source $\mathbf{S}_1$ operates under stricter QoS constraints with respect to $\mathbf{S}_2$ and consequently can support smaller arrival rates and needs, in turn, smaller transmission rates which can be sustained even in the presence of interference. If the roles are switched (i.e., if we have $\theta_2 = \theta_r = 1$ and $\theta_1 = 0.1$), then the optimal value of δ is zero. If the QoS exponents are more comparable (e.g., $\theta_1 = 1$ and $\theta_2 = 0.5$ or

[6]These concavity results can simplify the search for the optimal parameter setting with the use of convex optimization tools.

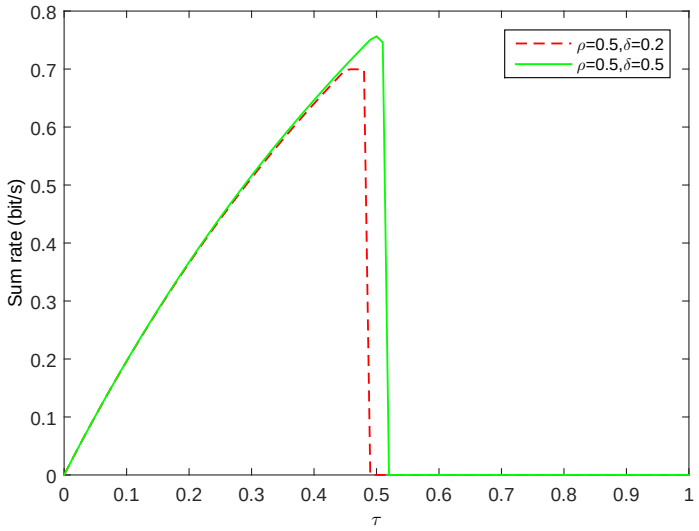

Figure 5.17: The sum rate vs. time allocation parameter τ.

$\theta_1 = 0.5$ and $\theta_2 = 1$), we notice that optimal values of δ start to slightly deviate from the two extremes of 0 and 1.

Fig. 5.20 shows the throughput regions of the two-source two-destination relay network under different queuing constraints. The boundary of the throughput region is obtained by searching over the three-dimensional parameter space. When R_1 achieves its maximum value, δ is close to 0, and ρ is slightly greater than 0.5, because decoding order $\{2, 1\}$ and more power in the $\mathbf{R} - \mathbf{D}_1$ link can help $\mathbf{S}_1 - \mathbf{D}_1$ link to support higher arrival rates. Similar results are also obtained for the maximum value of the arrival rate R_2.

5.3.3 Throughput of the Two-Source Two-Destination Relay Network With Fixed Transmission Rates

In practice, CSI may not be available at the transmitters. In such cases, the instantaneous departure rates from each buffer will be different. In this subsection, we investigate the system throughput when the transmitters do not have CSI and

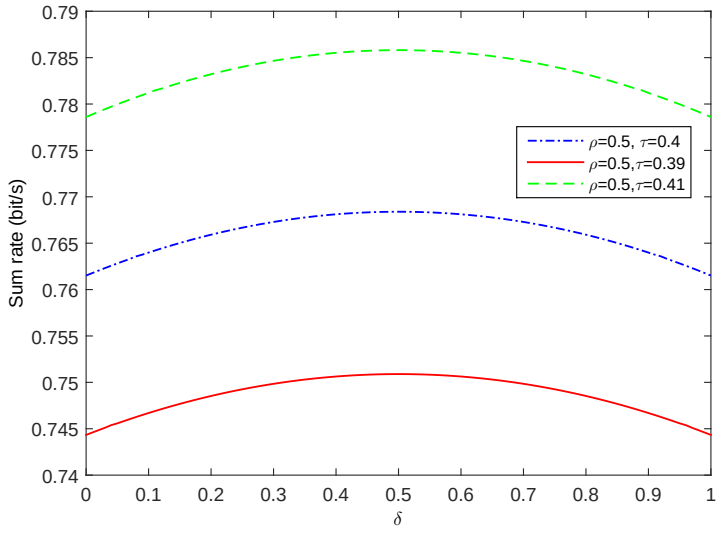

Figure 5.18: The sum rate vs. decoding parameter δ.

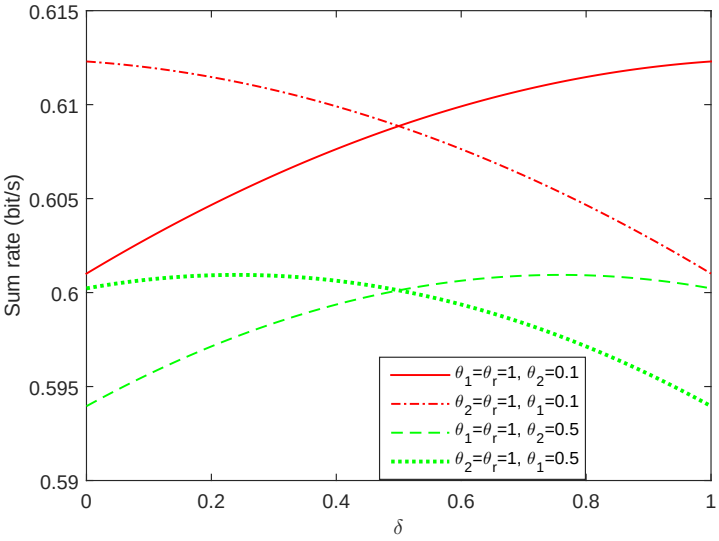

Figure 5.19: The sum rate vs. decoding parameter δ.

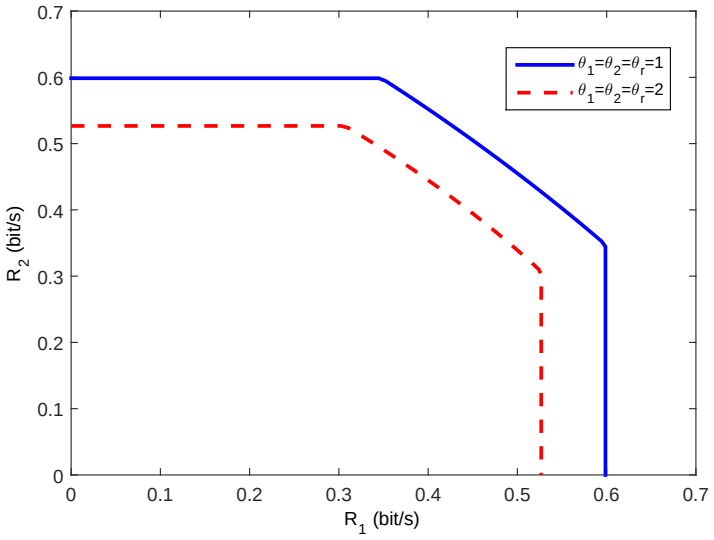

Figure 5.20: System throughput region R_1 vs. R_2.

transmit at fixed rates. We further assume that an ARQ protocol is employed and retransmissions are requested in case of communication failures. [89]

In the ARQ protocol, if the receiver decodes the packet, an ACK feedback is sent to the transmitter, otherwise the receiver asks for the retransmission of the same packet until the receiver gets the packet correctly. Here, the feedback signals are assumed to be transmitted without error and delay. In other words, the transmitter gets the error free feedback signal immediately after it completes the transmission of the corresponding packet. In this model, ARQ scheme guarantees the reliability, and the packets are kept in the buffer until the receiver decodes it correctly. With this ARQ assumption, the instantaneous departure rate at a buffer is equal to the fixed transmission rate if the receiver decodes the packet correctly, and it is 0 if the transmission fails.

In order to determine the asymptotic LMGFs $\Lambda_{\mathbf{S}_j,\mathbf{R}}$, $\Lambda_{\mathbf{R}}$ and $\Lambda_{\mathbf{R},\mathbf{D}_j}$, we have to first identify the success and failure probabilities of these fixed-rate transmissions.

5.3.3.1 State Probabilities in the Multiple Access Phase

As noted before, source node $\mathbf{S}_j$ transmits in the multiple-access phase with fixed rate $r_{\mathbf{S}_j,\mathbf{R}}$ for $j = 1, 2$. In the broadcast phase, relay node transmits to destination $\mathbf{D}_j$ with fixed rate $r_{\mathbf{R},\mathbf{D}_j}$, for $j = 1, 2$. Since all transmitters are using the ARQ protocol, all links can be regarded to be in either ON or OFF state at a given time. The link is in the ON state if the fixed transmission rate is less than the instantaneous channel capacity, and the receiver can decode the packet correctly. Otherwise, failure occurs and the link is in the OFF state in which the transmission rate is effectively zero.

In the multiple-access phase, the channel capacity is related to the decoding strategy of the relay, which is described as follows:

1. Relay tries to decode the first packet while treating the interference as noise. Without loss of generality, we assume that the relay always starts with the packets sent by $\mathbf{S}_1$.

   (a) If the receiver decodes the packet correctly, then it moves to the interference cancelation step (i.e., Step 2 below).

   (b) If the receiver cannot decode the packet from $\mathbf{S}_1$, it tries to decode the packet from $\mathbf{S}_2$.

   (c) If the receiver decodes it correctly, then it moves to the interference cancelation step. Otherwise, it asks retransmission from both transmitters, and decoding process ends.

2. The receiver performs interference cancelation by subtracting the decoded message from the received signal.

3. The receiver attempts to decode the remaining packet after interference cancelation. If it cannot decode the packet, retransmission is required from the corresponding transmitter.

Later in our analysis, we show that it does not make any difference if the relay starts with the packets sent by $\mathbf{S}_2$. In the multiple-access phase, according to the states of the links $\mathbf{S}_1 - \mathbf{R}$ and $\mathbf{S}_2 - \mathbf{R}$, we identify four possible cases:

Case 1: In this case, the relay node cannot decode any of the received messages. Relay node attempts to decode the message from $\mathbf{S}_1$ first, while treating the signal from $\mathbf{S}_2$ as noise. Following unsuccessful decoding, relay tries to decode the message from $\mathbf{S}_2$ while treating the interference as noise, and cannot succeed either. Hence, we in this scenario have

$$\begin{cases} r_{\mathbf{S}_1,\mathbf{R}} > & \tau B \log_2\left(1+\frac{\mathrm{SNR}_1 z_1}{1+\mathrm{SNR}_2 z_2}\right) \\ r_{\mathbf{S}_2,\mathbf{R}} > & \tau B \log_2\left(1+\frac{\mathrm{SNR}_2 z_2}{1+\mathrm{SNR}_1 z_1}\right) \end{cases}. \tag{5.49}$$

(5.49) can be transformed into the following bounds on fading magnitude-squares z_1 and z_2:

$$\begin{cases} z_1 > & \frac{1}{\mathrm{SNR}_1}\left(\mathrm{SNR}_2 z_2 / \left(2^{\frac{r_{\mathbf{S}_2,\mathbf{R}}}{\tau B}}-1\right)-1\right) \\ z_1 < & \frac{1}{\mathrm{SNR}_1}\left(2^{\frac{r_{\mathbf{S}_1,\mathbf{R}}}{\tau B}}-1\right)\left(1+\mathrm{SNR}_2 z_2\right) \\ z_2 > & 0 \\ z_2 < & -\left(2^{\frac{r_{\mathbf{S}_2,\mathbf{R}}}{\tau B}}-1\right)2^{\frac{r_{\mathbf{S}_1,\mathbf{R}}}{\tau B}} \Big/ \left\{\mathrm{SNR}_2\left[\left(2^{\frac{r_{\mathbf{S}_1,\mathbf{R}}}{\tau B}}-1\right)\left(2^{\frac{r_{\mathbf{S}_2,\mathbf{R}}}{\tau B}}-1\right)-1\right]\right\}, \\ & \qquad \text{if} \quad \left(2^{\frac{r_{\mathbf{S}_1,\mathbf{R}}}{\tau B}}-1\right)\left(2^{\frac{r_{\mathbf{S}_2,\mathbf{R}}}{\tau B}}-1\right) < 1. \end{cases} \tag{5.50}$$

(5.50) defines a region on the first quadrant of (z_1, z_2) plane, which we denote by Ψ_1. Therefore, the probability of Case 1 is given by

$$P_{M,1} = \iint_{\Psi_1} p_{z_1,z_2}(z_1, z_2) dz_1 dz_2, \tag{5.51}$$

where $p_{z_1,z_2}(z_1, z_2)$ is the joint probability density function (pdf) of z_1 and z_2. For instance, if we consider independent Rayleigh fading, then joint pdf is given by

$$p_{z_1,z_2}(z_1, z_2) = \frac{1}{\overline{z_1}\,\overline{z_2}} \exp\left(-\frac{z_1}{\overline{z_1}} - \frac{z_2}{\overline{z_2}}\right), \tag{5.52}$$

where $\overline{z_j}$ represents the expected value of z_j for $j = 1, 2$.

In this case, since the relay can decode none of them, switching the decoding order will not make a difference.

Case 2: In this case, the relay can decode the message from $\mathbf{S}_1$ in the presence of interference from $\mathbf{S}_2$, but the message from $\mathbf{S}_2$ cannot be decoded successfully even after interference cancelation. This scenario can be expressed by the following two inequalities:

$$\begin{cases} r_{\mathbf{S}_1,\mathbf{R}} \leq & \tau B \log_2\left(1 + \frac{\mathrm{SNR}_1 z_1}{1+\mathrm{SNR}_2 z_2}\right) \\ r_{\mathbf{S}_2,\mathbf{R}} > & \tau B \log_2\left(1 + \mathrm{SNR}_2 z_2\right) \end{cases}, \tag{5.53}$$

which can further be expressed as

$$\begin{cases} z_1 \geq & \left(2^{\frac{r_{\mathbf{S}_1,\mathbf{R}}}{\tau B}} - 1\right)(1 + \mathrm{SNR}_2 z_2)/\mathrm{SNR}_1 \\ z_2 < & \left(2^{\frac{r_{\mathbf{S}_2,\mathbf{R}}}{\tau B}} - 1\right)/\mathrm{SNR}_2 \end{cases}. \tag{5.54}$$

(5.54) defines the region Ψ_2 on the first quadrant of (z_1, z_2) plane, and the probability of Case 2 is given by

$$P_{M,2} = \iint\limits_{\Psi_2} p_{z_1,z_2}(z_1, z_2) dz_1 dz_2. \tag{5.55}$$

Notice that since the relay cannot decode the message from $\mathbf{S}_2$ even after interference cancelation, changing the decoding order would not help.

Case 3: This is the symmetric version of Case 2 with the roles of $\mathbf{S}_1$ and $\mathbf{S}_2$ interchanged. Hence, the relay can decode the message from $\mathbf{S}_2$ with interference, but not the message from $\mathbf{S}_1$. The probability of this case is given by

$$P_{M,3} = \iint\limits_{\Psi_3} p_{z_1,z_2}(z_1, z_2) dz_1 dz_2, \tag{5.56}$$

where Ψ_3 is the region in the first quadrant of (z_1, z_2) plane described by

$$\begin{cases} z_2 \geq & \left(2^{\frac{r_{\mathbf{S}_2,\mathbf{R}}}{\tau B}} - 1\right) (1 + \mathrm{SNR}_1 z_1)/\mathrm{SNR}_2 \\ z_1 < & \left(2^{\frac{r_{\mathbf{S}_1,\mathbf{R}}}{\tau B}} - 1\right) /\mathrm{SNR}_1. \end{cases} \tag{5.57}$$

Case 4: In this case, the relay can decode both messages from two source nodes. Although the description of this case is more involved, we can fortunately express the probability of this case as

$$P_{M,4} = 1 - \sum_{i=1}^{3} P_{M,i}. \tag{5.58}$$

Note now that the ON state probability of the $\mathbf{S}_j - \mathbf{R}$ link is given by

$$P_j = P_{M,j+1} + P_{M,4} \text{ for } j = 1, 2. \tag{5.59}$$

5.3.3.2 State Probabilities in Broadcast Phase

In the broadcast phase, the decoding strategy of the destination node $\mathbf{D}_j$ for $j = 1, 2$ is described as follows:

1. $\mathbf{D}_j$ attempts to decode its own packet first while treating the interference as noise.

   (a) If the receiver decodes correctly, then the decoding process ends.

   (b) If the receiver cannot decode its own packet first, it tries to decode the packet intended for the other destination first.

   (c) If the receiver decodes the other packet correctly, then it moves to the interference cancelation step. Otherwise, it asks for a retransmission from the relay node, and decoding process ceases.

2. The receiver performs interference cancelation by subtracting the decoded message from the received signal.

3. The receiver tries to decode its own packet after interference cancelation. If it still cannot decode the packet, retransmission is required from the relay.

There are two possibilities for link $\mathbf{R} - \mathbf{D}_1$ being in the ON state. $\mathbf{D}_1$ may decode its message while treating interference as noise, or it may decode the message for $\mathbf{D}_2$ first, and then decode its own message after interference cancelation. These are described by the following conditions:

$$r_{\mathbf{R},\mathbf{D}_1} \leq (1-\tau)B\log_2\left(1+\frac{\mathrm{SNR}_r\rho\omega_1}{1+\mathrm{SNR}_r(1-\rho)\omega_1}\right) \tag{5.60}$$

or

$$\begin{cases} r_{\mathbf{R},\mathbf{D}_1} & > (1-\tau)B\log_2\left(1+\frac{\mathrm{SNR}_r\rho\omega_1}{1+\mathrm{SNR}_r(1-\rho)\omega_1}\right) \\ r_{\mathbf{R},\mathbf{D}_1} & \leq (1-\tau)B\log_2(1+\mathrm{SNR}_r\rho\omega_1) \\ r_{\mathbf{R},\mathbf{D}_2} & \leq (1-\tau)B\log_2\left(1+\frac{\mathrm{SNR}_r(1-\rho)\omega_1}{1+\mathrm{SNR}_r\rho\omega_1}\right) \end{cases} \tag{5.61}$$

where $\omega_1 = |h_1|^2$. We first define

$$a_1 = \left(2^{\frac{r_{\mathbf{R},\mathbf{D}_1}}{(1-\tau)B}}-1\right) \Big/ \left\{\mathrm{SNR}_r\left[1-(1-\rho)2^{\frac{r_{\mathbf{R},\mathbf{D}_1}}{(1-\tau)B}}\right]\right\} \tag{5.62}$$

$$a_2 = \left(2^{\frac{r_{\mathbf{R},\mathbf{D}_1}}{(1-\tau)B}}-1\right)/(\mathrm{SNR}_r\rho) \tag{5.63}$$

$$a_3 = \left(2^{\frac{r_{\mathbf{R},\mathbf{D}_2}}{(1-\tau)B}}-1\right) \Big/ \left\{\mathrm{SNR}_r\left[1-\rho 2^{\frac{r_{\mathbf{R},\mathbf{D}_2}}{(1-\tau)B}}\right]\right\}. \tag{5.64}$$

Using the conditions in (5.60) and (5.61), we can express the ON probability of the $\mathbf{R}-\mathbf{D}_1$ link as

$$P_3 = \begin{cases} 0, & a_1 < 0 \text{ and } a_3 < 0 \\ \int_{a_1}^{\infty} p_{\omega_1}(\omega_1)d\omega_1, & (a_1 > 0 \text{ and } a_3 < 0) \text{ or } (a_3 > a_1 > 0) \\ \int_{\max\{a_2,a_3\}}^{\infty} p_{\omega_1}(\omega_1)d\omega_1, & \text{otherwise,} \end{cases} \tag{5.65}$$

where, for instance, in Rayleigh fading, $p_{\omega_1}(\omega_1) = \frac{1}{\overline{\omega_1}}\exp(-\omega_1/\overline{\omega_1})$ is the pdf of ω_1, and $\overline{\omega_1}$ is the expected value of ω_1. A similar analysis can be applied to obtain the ON state probability of the link $\mathbf{R}-\mathbf{D}_2$ as

$$P_4 = \begin{cases} 0, & b_1 < 0 \text{ and } b_3 < 0 \\ \int_{b_1}^{\infty} p_{\omega_2}(\omega_2) d\omega_2, & (b_1 > 0 \text{ and } b_3 < 0) \text{ or } (b_3 > b_1 > 0) \\ \int_{\max\{b_2, b_3\}}^{\infty} p_{\omega_2}(\omega_2) d\omega_2, & \text{otherwise,} \end{cases} \tag{5.66}$$

where, for instance, if again Rayleigh fading is considered, $p_{\omega_2}(\omega_2) = \frac{1}{\overline{\omega_2}} \exp(-\omega_2/\overline{\omega_2})$ is the pdf of ω_2, $\overline{\omega_2}$ is the expected value of ω_2, and parameters b_j for $j = 1, 2, 3$ are defined as

$$b_1 = \left(2^{\frac{r_{\mathbf{R},\mathbf{D}_2}}{(1-\tau)B}} - 1\right) \Big/ \left\{ \mathrm{SNR}_r \left[1 - \rho 2^{\frac{r_{\mathbf{R},\mathbf{D}_2}}{(1-\tau)B}} \right] \right\} \tag{5.67}$$

$$b_2 = \left(2^{\frac{r_{\mathbf{R},\mathbf{D}_2}}{(1-\tau)B}} - 1\right) / (\mathrm{SNR}_r (1-\rho)) \tag{5.68}$$

$$b_3 = \left(2^{\frac{r_{\mathbf{R},\mathbf{D}_1}}{(1-\tau)B}} - 1\right) \Big/ \left\{ \mathrm{SNR}_r \left[1 - (1-\rho) 2^{\frac{r_{\mathbf{R},\mathbf{D}_1}}{(1-\tau)B}} \right] \right\}. \tag{5.69}$$

From the view of ARQ, outage happens when the receiver cannot decode the received signal, thus $1 - P_j$ can be regarded as outage probabilities of their corresponding links.

5.3.3.3 Stability Conditions

Similar to the variable-rate case, stability at the source buffers is ensured by requiring the arrival rates to satisfy (2.25). The stability at the relay buffer requires the average arrival rate to be smaller than the average departure rate. This can be ensured by choosing the parameters (ρ, τ) accordingly. Now, our parameter space is just a two dimensional plane, and we can describe the feasible set of (ρ, τ) for stability as

$$\mathbf{\Xi} = \{(\rho, \tau) | r_{\mathbf{S}_1,\mathbf{R}} P_1 \leq r_{\mathbf{R},\mathbf{D}_1} P_3 \text{ and } r_{\mathbf{S}_2,\mathbf{R}} P_2 \leq r_{\mathbf{R},\mathbf{D}_2} P_4\}. \tag{5.70}$$

It can be easily seen that both average arrival rates $r_{\mathbf{S}_1,\mathbf{R}}P_1$ and $r_{\mathbf{S}_2,\mathbf{R}}P_2$ are monotonic increasing functions of τ, because allocating more time to the multiple-access phase is beneficial to links $\mathbf{S}_1-\mathbf{R}$ and $\mathbf{S}_2-\mathbf{R}$. For the same reason, average departure rates $r_{\mathbf{R},\mathbf{D}_1}P_3$ and $r_{\mathbf{R},\mathbf{D}_2}P_4$ are decreasing functions of τ. Therefore, for given ρ, conditions in (5.70) provide two upper bound curves on τ. Then, feasible set for stability is the region under these two upper bounds on the (ρ,τ) plane.

5.3.3.4 Throughput Region under Statistical Queuing Constraints

Similar to the variable-rate case, the system throughput is only defined for the feasible parameter setting, which guarantees the stability. For those parameter values outside the stability region, the system throughput is set to 0. For a given feasible (ρ,τ) pair and given fixed transmission rates, we next formulate the maximum constant arrival rates R_1 and R_2 at the source nodes under statistical queuing constraints parameterized by QoS exponents θ_1,θ_2 and θ_r. For the described ON-OFF link model with independent fading coefficients, the asymptotic LMGFs can be simplified as

$$\Lambda_{\mathbf{S}_j,\mathbf{R}}(\theta) = \log\left(e^{\theta r_{\mathbf{S}_j,\mathbf{R}}}P_j + e^{\theta 0}(1-P_j)\right) \tag{5.71}$$

$$= \log\left(e^{\theta r_{\mathbf{S}_j,\mathbf{R}}}P_j + 1 - P_j\right) \tag{5.72}$$

$$\Lambda_{\mathbf{R},\mathbf{D}_j}(\theta) = \log\left(e^{\theta r_{\mathbf{R},\mathbf{D}_j}}P_{j+2} + e^{\theta 0}(1-P_{j+2})\right) \tag{5.73}$$

$$= \log\left(e^{\theta r_{\mathbf{R},\mathbf{D}_j}}P_{j+2} + 1 - P_{j+2}\right) \quad j=1,2, \tag{5.74}$$

noting that the transmission rates are either equal to the fixed rates $r_{\mathbf{S}_j,\mathbf{R}}$ from the sources and $r_{\mathbf{R},\mathbf{D}_j}$ from the relay if transmissions are successful and the corresponding links are in the ON state with probabilities P_j and P_{j+2} for $j=1,2$, and are zero in case of failures. Recall that in order to satisfy the queuing constraints at both the sources and the relay, the arrival rates at the two source nodes should satisfy (2.25) and (2.26) simultaneously. Then, using (5.72) and (5.74) and considering (2.25) and

(2.26), we can characterize the maximum constant arrival rates as

$$\mathrm{R}_1 = \begin{cases} \min\Big\{ -\frac{1}{\theta_1}\log\left(e^{-\theta_1 r_{\mathbf{S}_1,\mathbf{R}}}P_1 + 1 - P_1\right), \\ \qquad -\frac{1}{\theta_r}\log\left(e^{-\theta_r r_{\mathbf{R},\mathbf{D}_1}}P_3 + 1 - P_3\right)\Big\} & \theta_r \le \theta_1 \\ \min\Big\{ -\frac{1}{\theta_1}\log\left(e^{-\theta_1 r_{\mathbf{S}_1,\mathbf{R}}}P_1 + 1 - P_1\right), \\ \qquad -\frac{1}{\theta_1}\big(\log\left(e^{-\theta_r r_{\mathbf{R},\mathbf{D}_1}}P_3 + 1 - P_3\right) \\ \qquad\quad + \log\left(e^{(\theta_r-\theta_1) r_{\mathbf{S}_1,\mathbf{R}}}P_1 + 1 - P_1\right)\big)\Big\} & \theta_r > \theta_1 \end{cases} \tag{5.75}$$

$$\mathrm{R}_2 = \begin{cases} \min\Big\{ -\frac{1}{\theta_2}\log\left(e^{-\theta_2 r_{\mathbf{S}_2,\mathbf{R}}}P_2 + 1 - P_2\right), \\ \qquad -\frac{1}{\theta_r}\log\left(e^{-\theta_r r_{\mathbf{R},\mathbf{D}_2}}P_4 + 1 - P_4\right)\Big\} & \theta_r \le \theta_2 \\ \min\Big\{ -\frac{1}{\theta_2}\log\left(e^{-\theta_2 r_{\mathbf{S}_2,\mathbf{R}}}P_2 + 1 - P_2\right), \\ \qquad -\frac{1}{\theta_2}\big(\log\left(e^{-\theta_r r_{\mathbf{R},\mathbf{D}_2}}P_4 + 1 - P_4\right) \\ \qquad\quad + \log\left(e^{(\theta_r-\theta_2) r_{\mathbf{S}_2,\mathbf{R}}}P_2 + 1 - P_2\right)\big)\Big\} & \theta_r > \theta_2. \end{cases} \tag{5.76}$$

Searching over the stability region $\boldsymbol{\Xi}$, the arrival rates R_1, R_2 and their sum rate can be further optimized over ρ and τ, which will be numerically evaluated in the next subsection.

5.3.3.5 Numerical Results

In this subsection, numerical results for the two-source two-destination relay network with fixed transmission rates are provided. First, we verify our analysis through Monte Carlo simulations. In each simulation, we generate 2×10^7 time blocks to estimate the buffer overflow probability, and repeat each simulation 500 times to evaluate the averages. We set the queuing constraints as $\theta_1 = \theta_2 = \theta_r = 0.1$, and the constant arrival rates at nodes $\mathbf{S}_1$ and $\mathbf{S}_2$ are chosen according to (5.75) and (5.76), respectively. We further assume that $r_{\mathbf{S}_1,\mathbf{R}} = r_{\mathbf{S}_2,\mathbf{R}} = r_{\mathbf{R},\mathbf{D}_1} = r_{\mathbf{R},\mathbf{D}_2} = 0.3$ bit/s, $\overline{z_1} = \overline{z_2} = \overline{\omega_1} = \overline{\omega_2} = 2$, $\mathrm{SNR}_1 = 6.02$ dB, $\mathrm{SNR}_2 = 4.77$ dB, $\mathrm{SNR}_r = 7.78$ dB, $\tau = 0.39$,

$\rho = 0.7$. We plot the logarithmic buffer overflow probabilities as functions of the overflow thresholds in Fig. 5.21. In this specific example, the system throughput is mainly decided by the multiple-access channel, and the overflow probabilities at the buffers of the two source nodes almost exactly meet the queuing constraints. The simulated slopes of the logarithmic overflow probabilities at $\mathbf{S}_1$ and $\mathbf{S}_2$ are -0.099 and -0.101, respectively. The overflow probabilities at the relay buffers diminish with steeper slopes than required and hence satisfy even stricter queuing constraints. Among the two relay buffers, we note that the overflow probability in the buffer keeping the data from $\mathbf{S}_1$ decays much more faster. This is because we set $\rho = 0.7$, meaning that the $\mathbf{R} - \mathbf{D}_1$ link gets more power than the $\mathbf{R} - \mathbf{D}_2$ link.

Fig. 5.22 demonstrates the case in which the performance bottleneck is in the broadcast phase. In Fig. 5.22, we again set the queuing constraints as $\theta_1 = \theta_2 = \theta_r = 0.1$. Other system parameters are given as $r_{\mathbf{S}_1,\mathbf{R}} = r_{\mathbf{S}_2,\mathbf{R}} = 0.3000$ bit/s, $r_{\mathbf{R},\mathbf{D}_1} = r_{\mathbf{R},\mathbf{D}_2} = 0.7673$bit/s, $\overline{z_1} = \overline{z_2} = \overline{\omega_1} = \overline{\omega_2} = 2$, $\text{SNR}_1 = 4.77$dB, $\text{SNR}_2 = 4.77$dB, $\text{SNR}_r = 10$dB, $\tau = 0.45$, $\rho = 0.65$. In this case, the overflow probabilities at the buffers of the two source nodes decrease faster than the imposed queuing constraints, and the overflow probabilities at the two relay buffers almost exactly meet the QoS requirements. Specifically, the simulated slopes of the logarithmic overflow probabilities at the two relay buffers are -0.100 and -0.098, respectively.

The rest numerical results are obtained for the following parameter values: $r_{\mathbf{S}_1,\mathbf{R}} = r_{\mathbf{S}_2,\mathbf{R}} = r_{\mathbf{R},\mathbf{D}_1} = r_{\mathbf{R},\mathbf{D}_2} = 0.3$ bit/s, $\overline{z_1} = \overline{z_2} = \overline{\omega_1} = \overline{\omega_2} = 2$, $\text{SNR}_1 = 6.02$ dB, $\text{SNR}_2 = 4.77$dB, $\text{SNR}_r = 7.78$dB, $\theta_1 = \theta_2 = 1$ and $\theta_r = 3$. Fig. 5.23 shows the influence of the power allocation parameter ρ on the successful transmission probabilities P_3 and P_4 in the broadcast phase. We observe that as ρ increases from 0 to 0.5 and hence a larger fraction of the power is allocated to the transmission of the message to $\mathbf{D}_1$, P_3 grows dramatically while P_4 diminishes by a relatively small amount. This indicates that the sum arrival rate increases initially with increasing ρ. We also notice that both

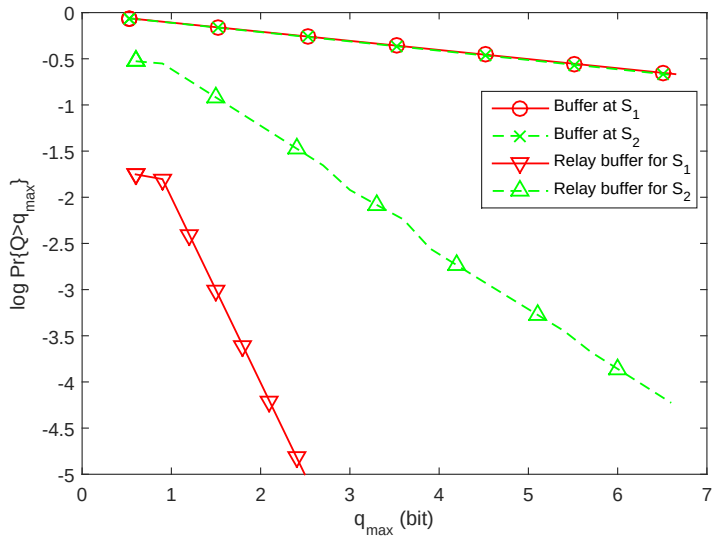

Figure 5.21: Logarithmic buffer overflow probability vs. buffer overflow threshold.

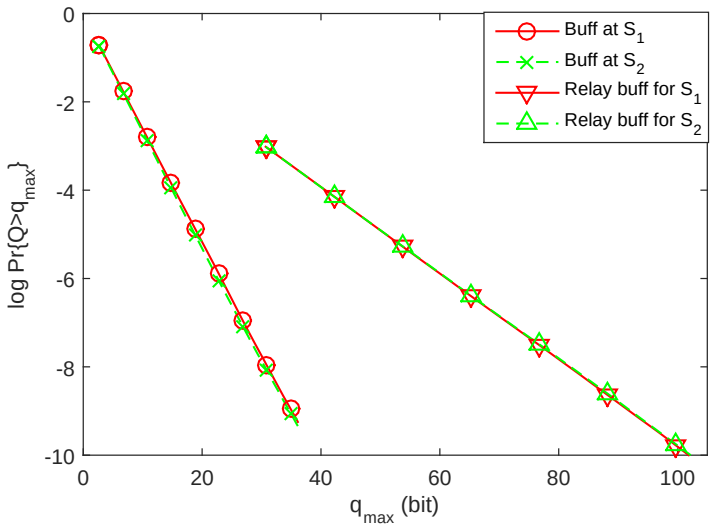

Figure 5.22: Logarithmic buffer overflow probability vs. buffer overflow threshold.

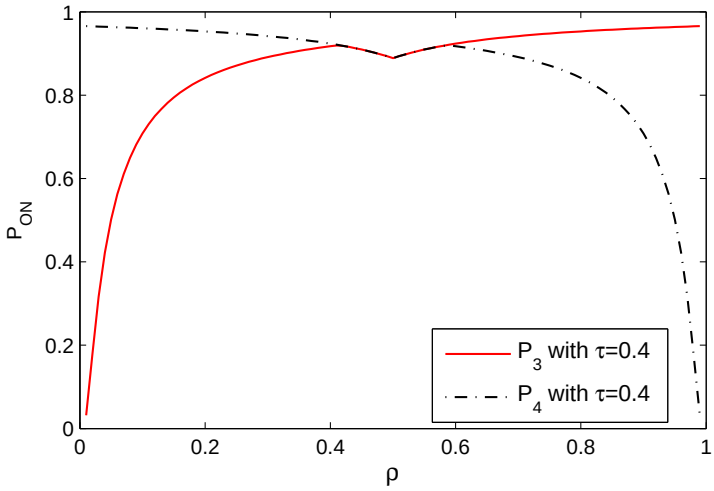

Figure 5.23: ON state probabilities in the broadcast phase vs. ρ with $\tau = 0.4$.

P_3 and P_4 decrease slightly at around $\rho = 0.5$ due to the increased interference caused by the joint transmission of messages at similar power levels in the broadcast phase. Similarly, the boundary of region of feasible (ρ, τ) pairs for stability at the relay buffer shown in Fig. 5.24 has a local minimum τ value at ρ close to 0.5. Additionally, we see in the figure that the boundary, which essentially bounds τ from above, can be regarded as the intersection of two upper bounds on τ as discussed in Section 5.3.3.3.

Fig. 5.25 shows the arrival rate R_1 as a function of (ρ, τ). Outside the feasible region, rate is set to zero. We note that as τ increases, R_1 initially increases and then decreases within the feasible region. From (5.75), we know that R_1 is characterized as the point-wise minimum of two functions, one being an increasing function of τ, while the other being a decreasing function. Therefore, there exists an optimal τ that maximizes R_1 for given ρ. We also observe that the maximum of R_1 over all (ρ, τ) is achieved when $\tau = 0.39$ and $\rho = 0.7$. Note that with this relatively large ρ value, the $\mathbf{S}_1 - \mathbf{R} - \mathbf{D}_1$ link can in general support higher arrival rates because the relay allocates more power for the transmission of the message coming from $\mathbf{S}_1$. Similar numerical results can be obtained for R_2. Expectedly, R_2 has higher values when $\rho < 0.5$.

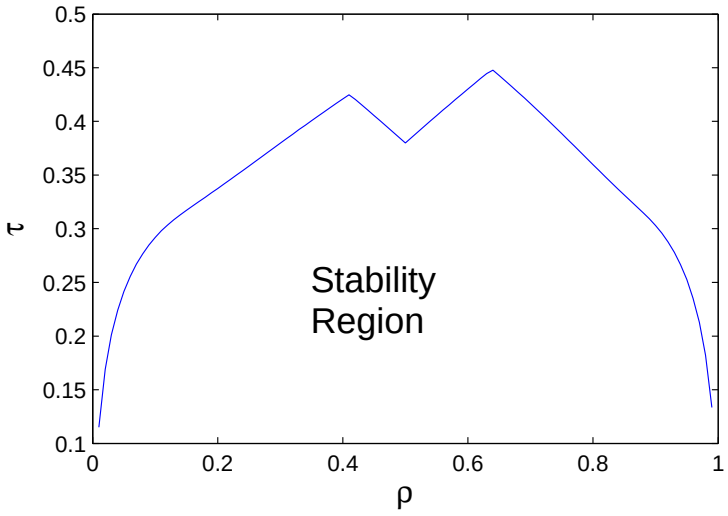

Figure 5.24: Region of feasible (ρ, τ) pairs for stability at the relay buffer.

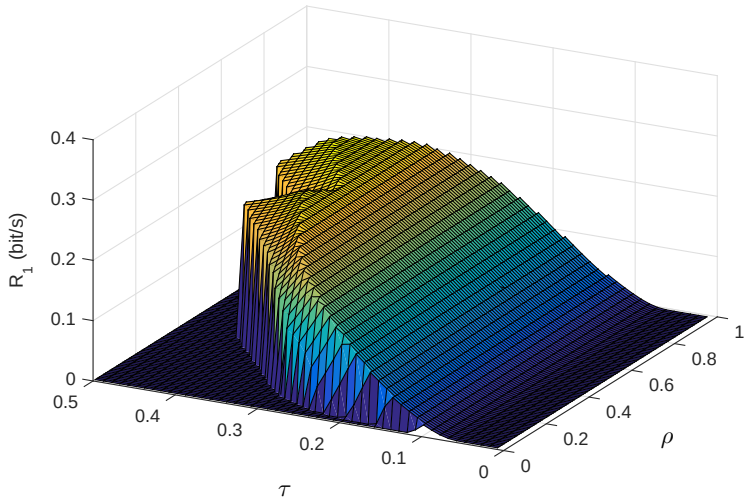

Figure 5.25: The arrival rate R_1 as a function of (ρ, τ).

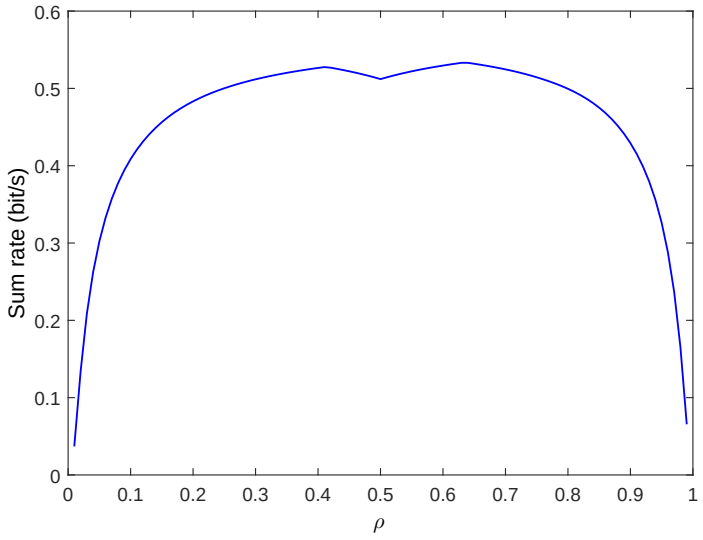

Figure 5.26: Maximum sum arrival rate vs. ρ.

Finally, we consider the maximum sum arrival rate. Fig. 5.26 plots the maximum sum arrival rate $\max\{R_1+R_2\}$ as a function of ρ. As ρ approaches 0, the performance of the $\mathbf{S}_1-\mathbf{R}-\mathbf{D}_1$ link is limited by the low transmission power of the relay, leading to the adoption of a very small τ value as seen in Fig. 5.24. Small value of τ lowers the throughput of the $\mathbf{S}_2-\mathbf{R}-\mathbf{D}_2$ link as well. Hence R_2 is also small. Similar concerns arise as ρ approaches 1. Hence, allocating the power almost exclusively for the transmission of one message is not an efficient strategy in terms of maximizing the sum rate. Indeed, the sum arrival rate is maximized when $\rho = 0.64$. However, it is interesting to note that equal allocation (i.e., having $\rho = 0.5$) is not the optimal strategy either, because the sum rate has a local minimum point around $\rho = 0.5$ again as a reflection of increased interference.

Chapter 6

Throughput and Mode Selection in Two-way MIMO Systems under Queuing Constraints

In this chapter, the throughput of and mode selection between half-duplex and full-duplex modes are studied in two-way MIMO systems operating under statistical queuing constraints. In particular, the effective capacity of these systems is determined in order to identify the throughput under constraints on the buffer overflow probability. In the low SNR regime, the optimal input covariance matrices that achieve the minimum energy per bit of the system are investigated. Full-duplex mode is found to have better performance at low SNRs and short distances, while half-duplex mode outperforms full-duplex operation at high SNR levels and long distances.

6.1 System Model

We consider a two-way MIMO system shown in Fig. 6.1, in which two users **A** and **B** want to exchange information with each other. Both of them are equipped with multiple antennas. N_A and N_B are the numbers of antennas at users **A** and **B**,

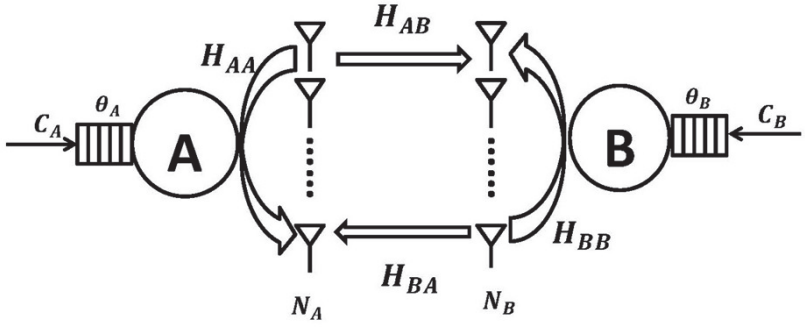

Figure 6.1: System model for two-way MIMO channel

respectively. We assume a discrete-time system model with block flat-fading. In each time block, the fading coefficients stay fixed, and change independently across blocks.

Also, we assume that there is a buffer at each node to store the arriving packets, and these two users have to satisfy the statistical queuing constraints described in Chapter 2. Packets will be cleared from the buffer only after the corresponding receiver successfully decodes them. Perfect CSI is assumed to be available at both nodes, so that transmitters can adapt their transmission rates to the channel conditions. In this section, we investigate the system throughput in both half-duplex and full-duplex modes.

6.1.1 Half-Duplex Mode

In the half-duplex mode, we consider both time division multiplexing (TDM) and frequency division multiplexing (FDM). In both cases, there is no self interference, and the input-output relationships are given by

$$\bar{y}_A = \mathbf{H}_{BA}\mathbf{x}_B + \bar{n}_A \tag{6.1}$$

$$\bar{y}_B = \mathbf{H}_{AB}\mathbf{x}_A + \bar{n}_B, \tag{6.2}$$

where $\mathbf{x}_j$, $\bar{y}_j$ and $\bar{n}_j$ are the received signal vector, transmitted signal vector and Gaussian noise vector of node j, for $j = \mathbf{A}, \mathbf{B}$, and $\mathbf{H}_{AB}$ and $\mathbf{H}_{BA}$ are the channel matrices for the links $\mathbf{A}-\mathbf{B}$ and $\mathbf{B}-\mathbf{A}$, respectively, whose components are the channel fading coefficients between the corresponding transmitting and receiving antenna pairs. The average energy of the transmitted signal is $\mathbb{E}\{||\mathbf{x}_j||^2\} = \frac{P_j}{B_j}$, where B_j is the bandwidth allocated to the j^{th} link. The noise vectors are assumed to be zero-mean Gaussian random vectors with covariance matrices given by $\mathbb{E}\{\bar{n}_j \bar{n}_j^\dagger\} = \sigma_n^2 \mathbf{I}$. At this point, we can define the system SNR for link $\mathbf{A} - \mathbf{B}$ as $\text{SNR}_A = \frac{\mathbb{E}\{||\mathbf{x}_A||^2\}}{\mathbb{E}\{||\bar{n}_B||^2\}} = \frac{P_A}{N_B B_A \sigma_n^2}$. Similarly, we have $\text{SNR}_B = \frac{P_B}{N_A B_B \sigma_n^2}$. In the special case of Rayleigh fading, the components of the channel fading matrices are assumed to follow zero-mean circularly symmetric Gaussian distribution with variance σ_h^2, which we denote by $\mathcal{CN}(0, \sigma_h^2)$. Note that $\bar{n}_B$ and $\mathbf{x}_B$ are $N_B \times 1$ dimensional vectors, $\bar{n}_A$ and $\mathbf{x}_A$ are $N_A \times 1$ dimensional vectors, $\mathbf{H}_{AB}$ is an $N_B \times N_A$ dimensional matrix, and $\mathbf{H}_{BA}$ is an $N_A \times N_B$ dimensional matrix.

In the half-duplex TDM mode, two users cannot transmit and receive at the same time. In this case, we denote the fraction of time allocated to link $\mathbf{A} - \mathbf{B}$ as τ, where $\tau \in [0, 1]$. Therefore, the fraction of time allocated to link $\mathbf{B} - \mathbf{A}$ is $1 - \tau$. In the half-duple FDM mode, the system divides the frequency band into two subbands. We denote the fraction of bandwidth allocated to link $\mathbf{A} - \mathbf{B}$ as ρ. Hence, the bandwidth allocated to the transmission of node A is $B_A = \rho B$, where $B = B_A + B_B$ is the total bandwidth of the system, and $B_B = (1 - \rho)B$ is the bandwidth allocated to the transmission of node $\mathbf{B}$. Since the numbers of antennas are fixed at $\mathbf{A}$ and $\mathbf{B}$ for all transmission modes, including the full-duplex mode, each antenna should transmit and receive at the same time in half-duplex FDM mode and full-duplex mode to achieve the best performance. For half-duplex FDM mode, self-interference is negligible since transmitting and receiving are being performed in different frequency bands. Self-interference cancellation in full-duplex mode will be discussed in the next subsection.

6.1.2 Full-Duplex Mode

In the full-duplex mode, two nodes **A** and **B** transmit and receive at the same time using a common frequency band, and there is self-interference which originates from the transmitting antennas at the same node. The discrete-time input-output relationships in the full-duplex mode can be expressed as

$$\bar{y}_A = \mathbf{H}_{BA}\mathbf{x}_B + \gamma_A \mathbf{H}_{AA}\mathbf{x}_A + \bar{n}_A \tag{6.3}$$

$$\bar{y}_B = \mathbf{H}_{AB}\mathbf{x}_A + \gamma_B \mathbf{H}_{BB}\mathbf{x}_B + \bar{n}_B \tag{6.4}$$

where γ_j and $\mathbf{H}_{jj}$ represent the self-interference cancelation parameter and self-interference channel matrix, respectively for node j, where $j = \mathbf{A}, \mathbf{B}$. The rest of the notation has the same description and SNR_j is defined in the same way as in half-duplex mode. The self-interference components, $\gamma_j \mathbf{H}_{jj}\mathbf{x}_j$, do not necessarily follow a Gaussian distribution. However, since Gaussian distribution leads to lower bounds on the channel capacities, we consider these components as zero-mean circularly-symmetrical complex Gaussian distributed, leading to a worst-case analysis.

To have a fair comparison of the full-duplex and half-duplex modes, we assume that the same number of antennas are employed at both nodes for transmission and reception in both modes. This requires each antenna to transmit and receive simultaneously in full-duplex mode. Feasibility of this was demonstrated by a practical design that has been proposed in [90], guaranteeing over 40 dB channel isolation between the transmitter and receiver circuity. With this assumption, we have two types of self-interference. The first type occurs between two antennas at the same node. For this type, interference signals are received from the transmitting antennas, and the channel fading coefficients are described by the off-diagonal components of $\mathbf{H}_{AA}$ and $\mathbf{H}_{BB}$, which are assumed to have zero-mean and variance σ_{s1}^2. At the receiver side, self-interference cancellation is described and quantified with the parameter γ_j

at node j. From (6.3) and (6.4), we note that $\gamma = 1$ represents no self-interference cancelation scheme, while $\gamma = 0$ represents perfect cancelation. The second type of self-interference occurs on a single antenna, and the main component of the interference comes from its own transmitting circuit. By applying the techniques in [90], this self-interference can be controlled by introducing isolation between the transmitting and receiving circuity sharing the same antenna. If we denote the cancelation parameter of this type of self-interference by β_j at node j, then the diagonal components of $\mathbf{H}_{jj}$ can be assumed to have zero-mean and variance $(\sigma_{s2}\beta_j/\gamma_j)^2$ to satisfy (6.3) and (6.4).

6.2 Throughput for Two-way MIMO systems

In this section, we formulate the system throughput under queuing constraints for both half-duplex and full-duplex modes. Apparently, the covariance matrices of the transmitted signals have significant influence on the system throughput. We define the normalized input covariance matrix of $\mathbf{x}_j$ as

$$\mathbf{K}_{\mathbf{x}_j} = \frac{\mathbb{E}\{\mathbf{x}_j\mathbf{x}_j^\dagger\}}{P_j/B_j}, \qquad \text{for} \quad j = \mathbf{A},\ \mathbf{B}. \tag{6.5}$$

Throughout this section, we mainly discuss the effective capacity for given input covariance matrices. The maximum throughput can be obtained through optimization over $\mathbf{K}_{\mathbf{x}_A}$ and $\mathbf{K}_{\mathbf{x}_B}$ pairs.

6.2.1 System Throughput for Half-Duplex TDM Mode

The transmission in the half-duplex TDM two-way MIMO systems can be divided into two phases. In the first phase, node $\mathbf{A}$ sends information to node $\mathbf{B}$, and node $\mathbf{B}$ keeps silent. This phase occupies τ fraction of the time. In this phase, the instantaneous

rate of user **A** can be expressed as

$$r_A = B\log_2 \det\left[\mathbf{I} + \frac{P_A}{B\sigma_n^2}\mathbf{H}_{AB}\mathbf{K}_{\mathbf{x}_A}\mathbf{H}_{AB}^{\dagger}\right] \tag{6.6}$$

$$= B\log_2 \det\left[\mathbf{I} + N_B \mathrm{SNR}_A \mathbf{H}_{AB}\mathbf{K}_{\mathbf{x}_A}\mathbf{H}_{AB}^{\dagger}\right], \tag{6.7}$$

for given $\mathbf{K}_{\mathbf{x}_A}$. In the next phase, users **A** and **B** exchange their roles, and only node **B** transmits. This phase occupies $1-\tau$ fraction of the time, and the instantaneous rate of user **B** is given by

$$r_B = B\log_2 \det\left[\mathbf{I} + \frac{P_B}{B\sigma_n^2}\mathbf{H}_{BA}\mathbf{K}_{\mathbf{x}_B}\mathbf{H}_{BA}^{\dagger}\right] \tag{6.8}$$

$$= B\log_2 \det\left[\mathbf{I} + N_A \mathrm{SNR}_B \mathbf{H}_{BA}\mathbf{K}_{\mathbf{x}_B}\mathbf{H}_{BA}^{\dagger}\right], \tag{6.9}$$

for given $\mathbf{K}_{\mathbf{x}_B}$. By plugging (6.7) and (6.9) into (2.7), the effective capacities of links $\mathbf{A}-\mathbf{B}$ and $\mathbf{B}-\mathbf{A}$ are given by

$$C_A = -\frac{1}{\theta_A}\log_e\left(\mathbb{E}\left\{e^{-\tau\theta_A r_A}\right\}\right), \quad \text{and} \tag{6.10}$$

$$C_B = -\frac{1}{\theta_B}\log_e\left(\mathbb{E}\left\{e^{-(1-\tau)\theta_B r_B}\right\}\right), \tag{6.11}$$

respectively, given the covariance matrices $\mathbf{K}_{\mathbf{x}_A}$ and $\mathbf{K}_{\mathbf{x}_B}$.

Theorem 19 *For given input covariance matrices, the sum throughput $C_A + C_B$ is a concave function of the time-fraction parameter τ.*

Proof 5 *By taking the second derivative of C_A with respect to τ, we get*

$$\frac{\partial^2 C_A}{\partial \tau^2} = -\frac{\theta_A}{\left(\mathbb{E}\left\{e^{-\tau\theta_A r_A}\right\}\right)^2}\left\{\mathbb{E}\left\{e^{-\tau\theta_A r_A}\right\}\mathbb{E}\left\{r_A^2 e^{-\tau\theta_A r_A}\right\} - \left(\mathbb{E}\left\{r_A e^{-\tau\theta_A r_A}\right\}\right)^2\right\}. \tag{6.12}$$

Applying Cauchy-Schwarz inequality, we have

$$\mathbb{E}\left\{e^{-\tau\theta_A r_A}\right\}\mathbb{E}\left\{r_A^2 e^{-\tau\theta_A r_A}\right\} \geq \left(\mathbb{E}\left\{r_A e^{-\tau\theta_A r_A}\right\}\right)^2. \tag{6.13}$$

From (6.13) which implies that the second derivative in (6.12) is non-positive, we determine that the effective capacity of the $\mathbf{A}-\mathbf{B}$ *link is a concave function of* τ*. Similarly, we can prove that* C_B *is also a concave function of* τ*, leading to the desired result that the sum throughput* $C_A + C_B$ *is concave.*

With this characterization, the optimal τ value which maximizes the sum throughput can be obtained via convex optimization algorithms.

6.2.2 System Throughput for Half-Duplex FDM Mode

In the half-duplex FDM mode, both users can transmit and receive simultaneously, using different frequency bands. The instantaneous transmission rates of users **A** and **B** are given by

$$r_A = \rho B \log_2 \det\left[\mathbf{I} + \frac{P_A}{\rho B\sigma_n^2}\mathbf{H}_{AB}\mathbf{K}_{\mathbf{x}_A}\mathbf{H}_{AB}^\dagger\right] \tag{6.14}$$

$$= \rho B \log_2 \det\left[\mathbf{I} + N_B \mathrm{SNR}_A \mathbf{H}_{AB}\mathbf{K}_{\mathbf{x}_A}\mathbf{H}_{AB}^\dagger\right], \tag{6.15}$$

and

$$r_B = (1-\rho) B \log_2 \det\left[\mathbf{I} + \frac{P_B}{(1-\rho)B\sigma_n^2}\mathbf{H}_{BA}\mathbf{K}_{\mathbf{x}_B}\mathbf{H}_{BA}^\dagger\right]$$

$$= (1-\rho) B \log_2 \det\left[\mathbf{I} + N_A \mathrm{SNR}_B \mathbf{H}_{BA}\mathbf{K}_{\mathbf{x}_B}\mathbf{H}_{BA}^\dagger\right], \tag{6.16}$$

respectively.

Comparing (6.15) and (6.16) with (6.7) and (6.9), we have the following observation.

Remark 1 *For a half-duplex FDM two-way MIMO system with system parameter ρ, there exists a half-duplex TDM system with parameter $\tau = \rho$, which can achieve the same system throughput with the same average power consumption.*

We use subscripts TDM and FDM to distinguish the corresponding quantities. For user **A**, we set $\mathrm{SNR}_{A,\mathrm{TDM}} = \mathrm{SNR}_{A,\mathrm{FDM}}$. Since we have $\tau = \rho$, by comparing (6.7) and (6.15), apparently we should have $r_{A,\mathrm{TDM}} = r_{A,\mathrm{FDM}}$, which also implies that we have $C_{A,\mathrm{TDM}} = C_{A,\mathrm{FDM}}$. From $\mathrm{SNR}_{A,\mathrm{TDM}} = \mathrm{SNR}_{A,\mathrm{FDM}}$, we can get $P_{A,\mathrm{TDM}} = P_{A,\mathrm{FDM}}/\rho$. Then, the average power of TDM system can be expressed as

$$\overline{P}_{A,\mathrm{TDM}} = \tau P_{A,\mathrm{TDM}} = \frac{\tau}{\rho} P_{A,\mathrm{FDM}} = P_{A,\mathrm{FDM}} = \overline{P}_{A,\mathrm{FDM}}.$$

Now, we have shown that for user **A**, we can achieve the same throughput using the same average power in both TDM and FDM systems, when we set $\mathrm{SNR}_{A,\mathrm{TDM}} = \mathrm{SNR}_{A,\mathrm{FDM}}$ and $\tau = \rho$. Similarly, we can show the same result for user **B**, when we set $\mathrm{SNR}_{B,\mathrm{TDM}} = \mathrm{SNR}_{B,\mathrm{FDM}}$ and $\tau = \rho$.

Therefore, TDM and FDM systems have the same performance, and we can only address one of them in the half-duplex mode. Hence, we subsequently consider the TDM system to represent the performance of the half-duplex mode.

6.2.3 System Throughput for Full-Duplex Mode

In full-duplex mode, since the two users transmit and receive simultaneously in the same frequency band, we have additional self-interference terms as seen in (6.3) and (6.4). As mentioned in Section 6.1.2, there are two types of self-interference, experienced due to transmissions from the other antennas at the same node and from the transmitting circuitry of the same antenna, respectively. These interference terms are characterized by the off-diagonal and diagonal elements of the self-interference channel matrices, respectively.

We define the overall interference covariance matrix $\mathbf{K}_{zj}$ as

$$\mathbf{K}_{zj} = \left(\frac{\gamma_j}{\sigma_n}\right)^2 \mathbf{H}_{jj}\mathbb{E}\{\mathbf{x}_j\mathbf{x}_j^\dagger\}\mathbf{H}_{jj}^\dagger + \frac{1}{\sigma_n^2}\mathbb{E}\{\bar{n}_j\bar{n}_j^\dagger\} \tag{6.17}$$

$$= N_j\gamma_j^2 \mathrm{SNR}_j \mathbf{H}_{jj}\mathbf{K}_{\mathbf{x}_j}\mathbf{H}_{jj}^\dagger + \mathbf{I}, \tag{6.18}$$

for $j = \mathbf{A}, \mathbf{B}$. Then, the instantaneous transmission rates for $\mathbf{A}$ and $\mathbf{B}$ can be written, respectively, as

$$r_A = B\log_2 \det\left[\mathbf{I} + N_B \mathrm{SNR}_A \mathbf{H}_{AB}\mathbf{K}_{\mathbf{x}_A}\mathbf{H}_{AB}^\dagger \mathbf{K}_{zB}^{-1}\right], \tag{6.19}$$

and

$$r_B = B\log_2 \det\left[\mathbf{I} + N_A \mathrm{SNR}_B \mathbf{H}_{BA}\mathbf{K}_{\mathbf{x}_B}\mathbf{H}_{BA}^\dagger \mathbf{K}_{zA}^{-1}\right], \tag{6.20}$$

where $\mathbf{K}_{zA}$ and $\mathbf{K}_{zB}$ are given by (6.18). Inserting (6.19) and (6.20) into (2.7), the effective capacities of the $\mathbf{A}-\mathbf{B}$ and $\mathbf{B}-\mathbf{A}$ links are given by

$$C_A = -\frac{1}{\theta_A}\log_e\left(\mathbb{E}\left\{e^{-\theta_A r_A}\right\}\right), \quad \text{and} \tag{6.21}$$

$$C_B = -\frac{1}{\theta_B}\log_e\left(\mathbb{E}\left\{e^{-\theta_B r_B}\right\}\right), \tag{6.22}$$

respectively.

6.3 Mode Selection

In this subsection, we investigate the mode selection protocol in two-way MIMO systems with the goal of maximizing the sum throughput $C_{sum} = C_A + C_B$. At the beginning of the transmission, two users evaluate the sum throughput of half-duplex and full-duplex modes, and choose the one with the higher sum throughput.

In Section 6.2, we have mentioned that the throughput depends on the input signal covariance matrices, $\mathbf{K}_{\mathbf{x}_A}$ and $\mathbf{K}_{\mathbf{x}_B}$. In general, it is not easy to determine the optimal covariance matrices that maximize the sum throughput, but we can identify them in the low-SNR regime.

6.3.1 Mode Selection in the Low-SNR Regime

In the low-SNR regime, minimum energy per bit is a widely used performance metric [84], which characterizes the minimum energy required to send one bit of information reliably. In our system setting, the minimum energy per bit is achieved when SNR approaches zero, and is given by

$$\frac{E_b}{N_0}_{\min} = \lim_{\mathrm{SNR}\to 0} \frac{\mathrm{SNR}}{C_E(\mathrm{SNR})} = \frac{1}{\dot{C}_E(0)}, \tag{6.23}$$

where $\dot{C}_E(0)$ denotes the first derivative of the the effective capacity function $C_E(\mathrm{SNR})$ with respect to SNR at $\mathrm{SNR} = 0$. Now, our purpose is to find the optimal covariance matrices that achieves the smallest $\frac{E_b}{N_0}_{\min}$ for both half-duplex and full-duplex modes.

Remark 2 *In the low-SNR regime, for half-duplex two-way MIMO systems, the optimal input covariance matrices that achieve the smallest* $\frac{E_b}{N_0}_{\min}$ *are given by*

$$\mathbf{K}_{\mathbf{x}_A} = \mathbf{u}_A\mathbf{u}_A^\dagger \quad \textit{and} \quad \mathbf{K}_{x,B} = \mathbf{u}_B\mathbf{u}_B^\dagger \tag{6.24}$$

where $\mathbf{u}_A$ *is the eigenvector corresponding to the maximum eigenvalue of* $\mathbf{H}_{AB}^\dagger\mathbf{H}_{AB}$*, and* $\mathbf{u}_B$ *is the eigenvector corresponding to the maximum eigenvalue of* $\mathbf{H}_{BA}^\dagger\mathbf{H}_{BA}$*.*

It has been proved in [39] that the optimal input covariance matrix that minimize the minimum energy per bit for a point to point MIMO channel is

$$\mathbf{K}_x = \mathbf{u}\mathbf{u}^\dagger, \tag{6.25}$$

where $\mathbf{u}$ is the eigenvector corresponding to the maximum eigenvalue of $\mathbf{H}^\dagger\mathbf{H}$, and $\mathbf{H}$ is the channel fading matrix. For our half-duplex two-way MIMO systems, transmissions over $\mathbf{A}-\mathbf{B}$ and $\mathbf{B}-\mathbf{A}$ links are separated, and minimizing the overall bit energy is equivalent to minimizing the individual bit energies of users $\mathbf{A}$ and $\mathbf{B}$. Applying (6.25) to links $\mathbf{A}-\mathbf{B}$ and $\mathbf{B}-\mathbf{A}$, we immediately have the observation in Remark 2.

In full-duplex mode, transmission over these two links are now interacting due to self-interference. In this case, we can identify the optimal solution, which minimizes the bit energies separately for users $\mathbf{A}$ and $\mathbf{B}$, via an iterative procedure. First, we have the following characterization.

Theorem 20 *In the low-SNR regime for full-duplex two-way MIMO systems, for given $\mathbf{K}_{\mathbf{x}_B}$, the optimal input covariance matrix $\mathbf{K}_{\mathbf{x}_A}$ that achieves the smallest $\frac{E_b}{N_0}_{\min}$ of user $\mathbf{A}$ is*

$$\mathbf{K}_{\mathbf{x}_A} = \mathbf{\Psi}_A\mathbf{\Psi}_A^\dagger, \tag{6.26}$$

where $\mathbf{\Psi}_A$ is the eigenvector corresponding to the maximum eigenvalue of $\mathbf{H}_{AB}^\dagger\mathbf{K}_{zB}^{-1}\mathbf{H}_{AB}$. Similarly for user $\mathbf{B}$, for given $\mathbf{K}_{\mathbf{x}_A}$, the optimal $\mathbf{K}_{\mathbf{x}_B}$ is

$$\mathbf{K}_{\mathbf{x}_B} = \mathbf{\Psi}_B\mathbf{\Psi}_B^\dagger, \tag{6.27}$$

where $\mathbf{\Psi}_B$ is the eigenvector corresponding to the maximum eigenvalue of $\mathbf{H}_{BA}^\dagger\mathbf{K}_{zA}^{-1}\mathbf{H}_{BA}$.

Proof: See Appendix A.13.

Using Theorem 20, we can determine the optimal covariance matrices through an iterative process. Initially, we set $\mathbf{K}_{\mathbf{x}_A} = \frac{1}{N_A}\mathbf{I}$ and $\mathbf{K}_{\mathbf{x}_B} = \frac{1}{N_B}\mathbf{I}$. In each iteration, we update $\mathbf{K}_{\mathbf{x}_A}$ and $\mathbf{K}_{\mathbf{x}_B}$ using (6.26) and (6.27), until convergence or the number of iterations reaches a limit.

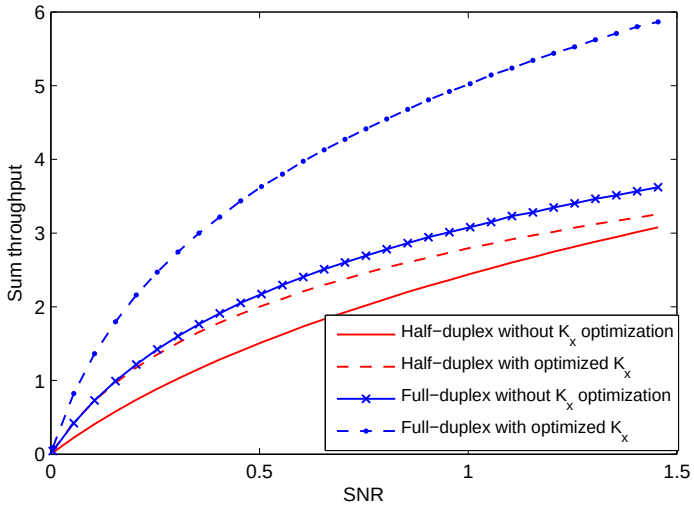

Figure 6.2: Sum throughput vs. SNR

Fig. 6.2 shows the low SNR performance for half-duplex and full-duplex modes with different input covariance matrices. For all numerical results in the paper, we assume Rayleigh fading. All fading coefficients between the corresponding transmitting and receiving antenna pairs follow zero-mean circularly symmetrical complex Gaussian distribution. In this figure, we set $\text{SNR}_A = \text{SNR}_B = \text{SNR}$, $N_A = N_B = 3$, $\theta_A = \theta_B = 1$, $\gamma_A = \gamma_B = 0.1$, $\sigma_n = 0.33$, $\sigma_h = 0.7$, $\sigma_{s1} = 10$, $\sigma_{s2}\beta_A = \sigma_{s2}\beta_B = 0.05$. For the curves with no covariance matrix optimization, we set $\mathbf{K}_{\mathbf{x}_A} = \frac{1}{N_A}\mathbf{I}$ and $\mathbf{K}_{\mathbf{x}_B} = \frac{1}{N_B}\mathbf{I}$. Fig. 6.2 shows that full-duplex mode has better performance at low SNR values, if self-interference is under control, because it allows two users to utilize the channel simultaneously. For both half-duplex and full-duplex modes, the sum throughput with optimized covariance matrices are greater. This is due to the facts that the direction with the best channel gain is selected and the influence of self-interference is reduced. Since the covariance matrix optimization is done for low SNR levels, when we have relatively higher SNR values, the advantage of the optimization is diminished in the half-duplex case.

Additionally, in our numerical analysis, we have observed that our iterative algorithm converges 99.47% of the time, and the average number of iterations needed to

achieve convergence is 13.77, when we set the criterion as $\mathrm{tr}(e_A e_A^\dagger)+\mathrm{tr}(e_B e_B^\dagger) \leq 10^{-10}$, where $e_j = \mathbf{K}_{\mathbf{x}_j}^i - \mathbf{K}_{\mathbf{x}_j}^{i-1}$ and $\mathbf{K}_{\mathbf{x}_j}^i$ is the covariance matrix after the i^{th} iteration, for $j = \mathbf{A}, \mathbf{B}$.

6.3.2 Mode Selection in the High-SNR Regime

Apparently, the sum throughput of the half-duplex mode grows without bound when both SNR_A and SNR_B are increased. However, the situation is different for the full-duplex mode.

Proposition 4 *In the high-SNR regime, when the power ratio P_A/P_B is kept fixed, the sum throughput approaches a constant, which only depends on the power ratio.*

Proof 6 *If we keep $\frac{P_A}{P_B} = \eta$ constant, plugging (6.18) into (6.19), the instantaneous rate for the $\mathbf{A}-\mathbf{B}$ link as $SNR_A \to \infty$ becomes*

$$\begin{aligned}\lim_{SNR_A\to\infty} r_A &= \lim_{SNR_A\to\infty} B\log_2\det\left[\mathbf{I}+N_B SNR_A \mathbf{H}_{AB}\mathbf{K}_{\mathbf{x}_A}\mathbf{H}_{AB}^\dagger \mathbf{K}_{zB}^{-1}\right] \\ &= \lim_{SNR_A\to\infty} B\log_2\det\left[\mathbf{I}+\frac{\eta}{\gamma_B^2}\mathbf{H}_{AB}\mathbf{K}_{\mathbf{x}_A}\mathbf{H}_{AB}^\dagger(\mathbf{H}_{BB}\mathbf{K}_{\mathbf{x}_B}\mathbf{H}_{BB}^\dagger)^{-1}\right] \\ &= B\log_2\det\left[\mathbf{I}+\frac{\eta}{\gamma_B^2}\mathbf{H}_{AB}\mathbf{K}_{\mathbf{x}_A}\mathbf{H}_{AB}^\dagger(\mathbf{H}_{BB}\mathbf{K}_{\mathbf{x}_B}\mathbf{H}_{BB}^\dagger)^{-1}\right],\end{aligned}$$

which approaches a constant for large SNR_A and SNR_B values. A similar result can be found for r_B under the high SNR assumption. Then the sum throughput $C_A + C_B$ approaches a constant that only depends on η.

Proposition 4 implies at high SNRs that half-duplex mode has better performance than the full-duplex mode. When the transmission power increases, self-interference also increases proportionally, and becomes difficult to control. Or equivalently, when the noise power vanishes, the system will primarily be self-interference limited. In this case, the advantage of half-duplex operation, which avoids self-interference, becomes

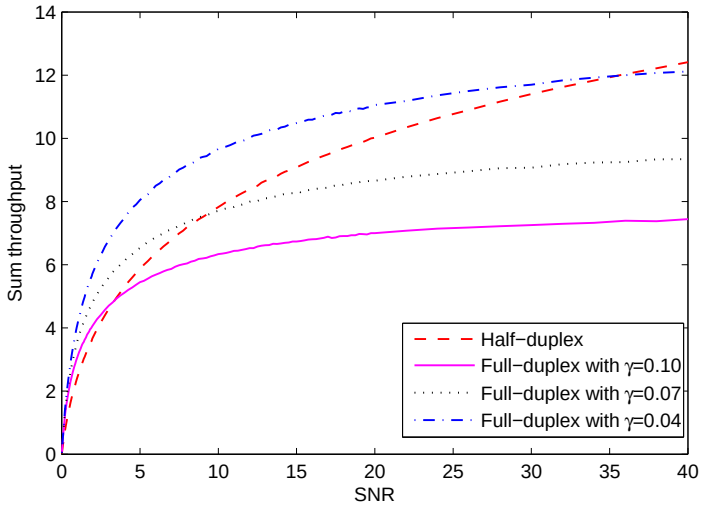

Figure 6.3: Sum throughput vs. SNR

significantly beneficial. Fig. 6.3 shows the sum throughput of the full-duplex mode with different self-interference cancelation parameters and of the half-duplex mode over a relative wide SNR range. We set $\gamma_A = \gamma_B = \gamma$, $\mathbf{K}_{\mathbf{x}_A} = \frac{1}{N_A}\mathbf{I}$, $\mathbf{K}_{\mathbf{x}_B} = \frac{1}{N_B}\mathbf{I}$, and all other parameter settings are the same as in Fig. 6.2. Although full-duplex mode has better performance at low SNRs, half-duplex operation starts outperforming when the SNR level is sufficiently high, and the gap increases as SNR increases. When γ decreases, we have better self-interference cancelation in the full-duplex mode, and the performance improves. If we have perfect interference cancelation, full-duplex performance should always be better than that of the half-duplex mode.

6.3.3 Mode Selection at Different Transmission Distances

In addition to the SNR level, the transmission distance also has an impact on mode selection. The distance between the two users affects the distribution of $\mathbf{H}_{AB}$ and $\mathbf{H}_{BA}$. Assuming that the transmission distance is d, we set $\sigma_h^2 = \frac{1}{d^4}$. In Fig. 6.4, we plot the sum throughput of half-duplex and full-duplex modes as a function of d for different queuing constraints. We set $\theta_A = \theta_B = \theta$, $\gamma_A = \gamma_B = 0.8$, $\text{SNR}_A = \text{SNR}_B = 7$, and other parameter settings are the same as in Fig 6.2. When the distance is

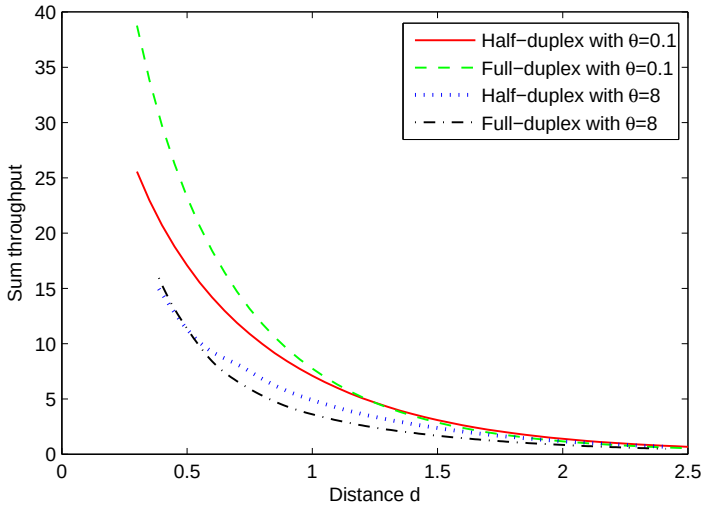

Figure 6.4: Sum throughput vs. transmission distance

small, full-duplex mode has better performance. As the distance increases, channel conditions become worse, and the received signal power diminishes, but the self-interference is still unchanged. As a result, the SINR in full-duplex mode decreases very fast, and the performance drops below that of half-duplex mode. Therefore, self-interference control is a critical problem in the full-duplex mode in long-distance transmissions. Also, Fig. 6.4 shows that increasing the values of the QoS exponent θ lowers the system throughput.

Chapter 7

Mode Selection and Resource Allocation Algorithms for D2D Cellular Networks

In this chapter, we study the mode selection and resource allocation algorithms for D2D cellular networks. D2D communication underlaid with cellular networks is a new paradigm, proposed to enhance the performance of cellular networks. By allowing a pair of D2D users to communicate directly and share the same spectral resources with the cellular users, D2D communication can achieve higher spectral efficiency, improve the energy efficiency, and lower the traffic delay.

In Section 7.1, transmission mode selection and resource allocation in a TDM cellular network with one cellular user, one base station, and a pair of D2D users is investigated under rate and queueing constraints. In particular, four possible modes are considered, namely the cellular mode, dedicated mode, uplink reuse mode, and downlink reuse mode. Using tools from stochastic network calculus, the system throughput under statistical queueing constraints is formulated, efficient resource allocation algorithms for all possible modes are proposed, and the influence of the

positions of each node and the queueing constraints is analyzed via numerical results. Scenarios and conditions for different modes to be optimal in the sense of maximizing the sum-throughput are identified.

In Section 7.2, we propose a novel channel matching algorithm for joint mode selection and channel allocation with the goal of maximizing the system throughput under statistical queueing constraints. Seven possible modes are considered, namely the D2D cellular mode, D2D dedicated mode, uplink dedicated mode, downlink dedicated mode, uplink reuse mode, downlink reuse mode, and D2D reuse mode. The throughput is characterized by determining the effective capacity. We formulate the channel allocation problem as a maximum-weight matching problem, which can be solved by employing the Hungarian algorithm. Via simulation results, we verify the performance improvements achieved by our proposed matching algorithm.

In Section 7.3, we propose a novel joint mode selection and channel resource allocation algorithm via the vertex coloring approach. We decompose the problem into three subproblems and design algorithms for each of them. In the first step, we divide the users into groups using a vertex coloring algorithm. In the second step, we solve the power optimization problem using the interior-point method for each group and conduct mode selection between the cellular mode and D2D mode for D2D users, and we assign channel resources to these groups in the final step. Numerical results show that our algorithm achieves higher sum rate and serves more users with relatively small time consumption compared with other algorithms. Also, the influence of system parameters and the tradeoff between sum rate and the number of served users are studied through simulation results.

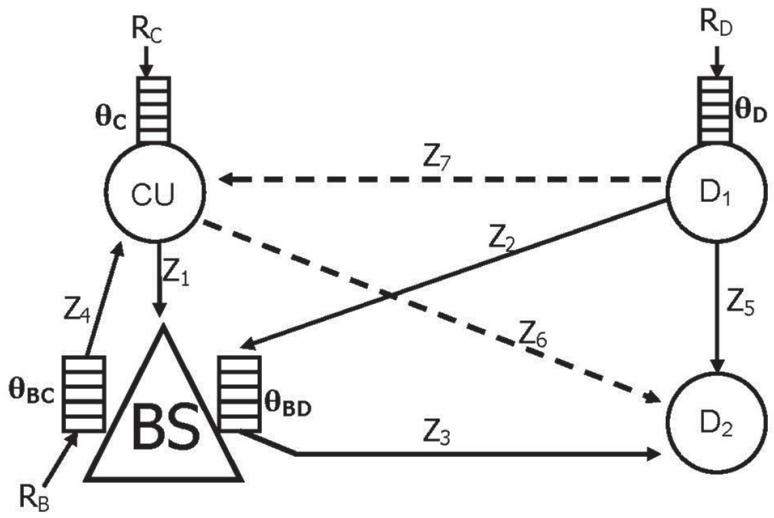

Figure 7.1: System model with queuing constraints (Dashed lines represent interference only links.)

7.1 Mode Selection of Device-to-Device Communication in Cellular Networks under Statistical Queuing Constraints

7.1.1 System Model

As mentioned above, we study mode selection and resource allocation in a cellular network with D2D users with the goal of maximizing the overall sum rate of the network under queuing constraints. For simplicity, we consider a typical model shown in Fig. 7.1, in which there is only one cellular user (**CU**), one base station (**BS**), and one pair of D2D users denoted by $\mathbf{D}_1$ and $\mathbf{D}_2$, respectively. We assume the transmission between D2D users is one-way, and $\mathbf{D}_1$ is the transmitter and $\mathbf{D}_2$ is the receiver. Therefore, there are overall three transmitters in this network, namely **BS**, **CU** and $\mathbf{D}_1$, and their maximum transmission powers are denoted by P_b, P_c and P_{dmax}. There are three main communication links corresponding to the three transmitters. In the uplink, the cellular user **CU** sends information to the base

station; in the downlink, the base station sends information to the cellular user; and in the D2D link, $\mathbf{D}_1$ transmits to $\mathbf{D}_2$. Before the transmitters send their packets to the corresponding receivers, the packets are stored in buffers. All transmitters are operating under statistical queuing constraints imposed as limitations on buffer overflow probabilities.

We consider a block-fading model. The fading coefficients and their magnitude-squares in different communication and interference links are represented by h_i and $z_i = |h_i|^2$, respectively, as depicted in Fig. 7.1 for $i = 1, 2, \cdots, 7$. The overall bandwidth of this system is B, and the time in each transmission period is divided into several phases and allocated to different links. The fractions of time allocated to each link are given by the elements of the time allocation vector $\boldsymbol{\tau}$. For different modes, the dimension of $\boldsymbol{\tau}$ is different. In this section, we consider four modes for this network, namely the cellular mode, the dedicated mode, the uplink reuse mode, and the downlink reuse mode.

Cellular Mode: In this mode, the transmission is divided into 4 phases. In the first phase, only the uplink is active, in which **CU** sends information to **BS**; in the second phase, only the downlink is active, in which **BS** sends information to **CU**; in the third phase, $\mathbf{D}_1$ sends information to the base station (only $\mathbf{D}_1 - \mathbf{BS}$ link is active), and the base station decodes and forwards the information to $\mathbf{D}_2$ in the last phase (only $\mathbf{BS} - \mathbf{D}_2$ link is active). In the cellular mode, the base station works as a relay node between $\mathbf{D}_1$ and $\mathbf{D}_2$, and hence these two D2D users essentially communicate just like the cellular users. We denote the fractions of time allocated to the cellular uplink and downlink as τ_1 and τ_2, respectively, and we denote the overall fraction of time allocated to the two D2D links as τ_3. The time allocation between the two D2D links is discussed in Section 7.1.2.1. Now, we have $\|\boldsymbol{\tau}\|_1 = \sum_{j=1}^{3} \tau_j = 1$. For simplicity, we assume that there are two separate buffers at the base station. One buffer is for the packets that will be sent to **CU**, and the other one stores the packets

that has arrived from $\mathbf{D}_1$ and will be sent to $\mathbf{D}_2$. These two buffers operate under different queuing constraints. Since there is only one pair of transmitter and receiver being active in each phase, there is no interference in the cellular mode. In each time block, the received signal at each receiver has the form

$$y = h_i x + n_i, \tag{7.1}$$

where x is the transmitted signal, h_i is the corresponding channel fading coefficient, and n_i is the additive noise component. At different receiver nodes, the noise terms are assumed to be independent, zero-mean, circularly symmetric, complex Gaussian random variables with variances $\mathbb{E}\{|n_i|^2\} = \sigma_0^2$.

Dedicated Mode: In this mode, the transmission is divided into three phases, which are the uplink phase, downlink phase and D2D phase, and the fractions of time allocated to each phase are given by $\boldsymbol{\tau} = (\tau_1, \tau_2, \tau_3)$. Compared to the cellular mode, the only difference is that $\mathbf{D}_1$ sends information directly to $\mathbf{D}_2$ without the participation of the base station in the third phase. Similar to the cellular mode, there is no interference, and the received signals have the form in (7.1).

Reuse Modes: In the uplink and downlink reuse modes, the D2D link reuses the channel resource with uplink and downlink, respectively. Therefore, the transmission is divided into two phases in these two modes. In the uplink reuse mode, the first phase is allocated to the downlink, and the second phase is allocated to the uplink and D2D link. Since $\mathbf{D}_1$ and **CU** transmit in the second phase simultaneously, $\mathbf{D}_2$ and **BS** experience interference. In the uplink channel, the received signals follow the form

$$y = h_i x + h_{\text{inter}} x_{\text{inter}} + n_i, \tag{7.2}$$

where x is the desired signal, h_i is the fading coefficient of the channel between this

receiver and its corresponding transmitter, x_{inter} is the interference signal, h_{inter} is the fading coefficient of the interfering link, and n_i is the Gaussian noise. Since there is no interference in the downlink phase, the received signal at **CU** node follows the form in (7.1). In the downlink reuse mode, the first phase is allocated to the uplink, and the second phase is allocated to the downlink and D2D link. Similarly, the received signals at $\mathbf{D}_2$ and **CU** follow (7.2) in the downlink reuse mode, and the received signal at the base station in the uplink channel follows (7.1).

In order to guarantee certain performance levels, we impose throughput/rate constraints for all users in all 4 modes. The minimum throughput required in the uplink, downlink and D2D link are denoted by $\mathbf{R}_{Cmin}$, $\mathbf{R}_{Bmin}$ and $\mathbf{R}_{Dmin}$, respectively. If one of the rate constraints cannot be satisfied, then the corresponding mode is not used.

7.1.2 Throughput of Cellular Network with D2D Users

In the previous subsection, we introduce the formulation of the system throughput under queuing constraints. In order to determine the effective capacities, we have to characterize the instantaneous transmission rates of each communication link. In this subsection, we formulate the overall system throughput of the cellular network in each mode, under given resource allocation strategies, and efficient resource allocation algorithms are provided in Section 7.1.3.

7.1.2.1 Throughput in the Cellular Mode

In the cellular mode, $\mathbf{CU}-\mathbf{BS}$, $\mathbf{BS}-\mathbf{CU}$, $\mathbf{D}_1-\mathbf{BS}$ and $\mathbf{BS}-\mathbf{D}_2$ links are active. The fractions of time allocated to $\mathbf{CU}-\mathbf{BS}$, $\mathbf{BS}-\mathbf{CU}$ and $\mathbf{D}_1-\mathbf{BS}-\mathbf{D}_2$ links are given by τ_1, τ_2 and τ_3, respectively. In this mode, all transmitters transmit with their maximum power since there is no interference. The instantaneous transmission rate

of the uplink channel is

$$r_{C,B} = B \log_2 \left(1 + \frac{P_c}{B\sigma_0^2} z_1\right). \tag{7.3}$$

Plugging (7.3) into (2.7), we obtain the effective capacity of the $\mathbf{CU} - \mathbf{BS}$ link as

$$\mathbf{R}_C = -\frac{1}{\theta_C} \log \mathbb{E}\left\{e^{-\tau_1 \theta_C r_{C,B}}\right\}. \tag{7.4}$$

For the downlink channel, the instantaneous rate is

$$r_{B,C} = B \log_2 \left(1 + \frac{P_b}{B\sigma_0^2} z_4\right), \tag{7.5}$$

and the corresponding effective capacity is

$$\mathbf{R}_B = -\frac{1}{\theta_{BC}} \log \mathbb{E}\left\{e^{-\tau_2 \theta_{BC} r_{B,C}}\right\}. \tag{7.6}$$

In the cellular mode, the D2D link is a two-hop channel with two queues in tandem. More specifically, $\mathbf{D}_1$ first sends information to **BS** in the third phase, and then **BS** forwards the information to $\mathbf{D}_2$ in the last phase. The instantaneous transmission rates of $\mathbf{D}_1 - \mathbf{BS}$ and $\mathbf{BS} - \mathbf{D}_2$ links are given, respectively, by

$$r_{D,B} = B \log_2 \left(1 + \frac{P_{dmax}}{B\sigma_0^2} z_2\right), \tag{7.7}$$

$$r_{B,D} = B \log_2 \left(1 + \frac{P_b}{B\sigma_0^2} z_3\right). \tag{7.8}$$

Define $\hat{\tau}$ as the fraction of time allocated to the third phase. Then the fraction of time allocated to the last phase is $\tau_3 - \hat{\tau}$. In this two-hop case, the arrival rate at node $\mathbf{D}_1$ has to satisfy the QoS constraints at $\mathbf{D}_1$ and **BS**, simultaneously. The effective capacity analysis of this half-duplex two-hop channel is much more involved,

and a detailed analysis is given in [17, Section III-B]. Through a similar process, we express the effective capacity of the $\mathbf{D}_1 - \mathbf{BS} - \mathbf{D}_2$ link as

$$\mathbf{R}_D = -\frac{1}{\theta_D} \log \mathbb{E}\{e^{-\hat{\tau}\theta_D r_{D,B}}\}, \tag{7.9}$$

where $\hat{\tau} = \min\{\tau_0, \tau^*\}$, and τ_0 is the solution to

$$\tau_0 \mathbb{E}\{r_{D,B}\} = (\tau_3 - \tau_0)\mathbb{E}\{r_{B,D}\}, \tag{7.10}$$

and τ^* is the solution to

$$-\frac{1}{\theta_D} \log \mathbb{E}\{e^{-\tau^*\theta_D r_{D,B}}\} = -\frac{1}{\theta_{BD}} \log \mathbb{E}\{e^{-(\tau_3-\tau^*)\theta_{BD} r_{B,D}}\} \tag{7.11}$$

when $\theta_D \geq \theta_{BD}$, or

$$-\frac{1}{\theta_D} \log \mathbb{E}\{e^{-\tau^*\theta_D r_{D,B}}\} = -\frac{1}{\theta_D} \left(\log \mathbb{E}\{e^{-(\tau_3-\tau^*)\theta_{BD} r_{B,D}}\} + \log \mathbb{E}\{e^{\tau^*(\theta_{BD}-\theta_D) r_{D,B}}\} \right) \tag{7.12}$$

when $\theta_D < \theta_{BD}$.

7.1.2.2 Throughput in the Dedicated Mode

In the dedicated mode, $\mathbf{CU} - \mathbf{BS}$, $\mathbf{BS} - \mathbf{CU}$, and $\mathbf{D}_1 - \mathbf{D}_2$ links are active, and the fraction of time allocated to these links are given by $\boldsymbol{\tau} = (\tau_1, \tau_2, \tau_3)$. In this mode, $\mathbf{D}_1$ transmits to $\mathbf{D}_2$ directly without the help of the base station. Since there is no interference, all transmitters transmit using their maximum power. Similar to the analysis above, the instantaneous transmission rates of the uplink, downlink are also given by (7.3) and (7.5), and their effective capacities still follow (7.4) and (7.6). For

the direct D2D link, the instantaneous transmission rate is

$$r_{D1,D2} = B \log_2 \left(1 + \frac{P_{dmax}}{B\sigma_0^2} z_5\right). \tag{7.13}$$

Plugging (7.13) into (2.7), we can get its corresponding effective capacity expressed as

$$\mathbf{R}_D = -\frac{1}{\theta_D} \log \mathbb{E}\left\{e^{-\tau_3 \theta_D r_{D1,D2}}\right\}. \tag{7.14}$$

7.1.2.3 Throughput in Reuse Modes

In the two reuse modes, only one cellular link is active in the first phase, and the other cellular link share the channel resource with the D2D direct link in the second phase. We denote the fraction of time allocated to the first phase as τ_1. Then the fraction of time left for the second phase is $1-\tau_1$. In reuse modes, we assume that the base station and cellular user transmit with their maximum power, due to cellular links being assumed to have higher priority over the D2D link. Assume that the transmission power of $\mathbf{D}_1$ is given by P_d. Then the instantaneous transmission rates for the downlink, uplink and D2D link are given by

$$r_{B,C} = B \log_2 \left(1 + \frac{P_b}{B\sigma_0^2} z_4\right) \tag{7.15}$$

$$r_{C,B} = B \log_2 \left(1 + \frac{P_c/B\sigma_0^2}{1 + (P_d/B\sigma_0^2) z_2} z_1\right) \tag{7.16}$$

$$r_{D1,D2} = B \log_2 \left(1 + \frac{P_d/B\sigma_0^2}{1 + (P_c/B\sigma_0^2) z_6} z_5\right), \tag{7.17}$$

in the uplink reuse mode, and their corresponding effective capacities are given, respectively, by

$$\mathbf{R}_B = -\frac{1}{\theta_{BC}} \log \mathbb{E}\left\{e^{-\tau_1 \theta_{BC} r_{B,C}}\right\} \tag{7.18}$$

$$\mathbf{R}_C = -\frac{1}{\theta_C} \log \mathbb{E}\left\{e^{-(1-\tau_1)\theta_C r_{C,B}}\right\} \tag{7.19}$$

$$\mathbf{R}_D = -\frac{1}{\theta_D} \log \mathbb{E}\left\{e^{-(1-\tau_1)\theta_D r_{D1,D2}}\right\}. \tag{7.20}$$

Similar characterizations can be obtained for the downlink reuse mode, and the optimization over P_d in reuse modes is discussed in the next subsection.

7.1.3 Resource Allocation

In the previous subsection, we have formulated the throughput under statistical queuing constraints for all 4 modes. In this subsection, we investigate efficient resource allocation strategies with the goal of maximizing the sum throughput. In the cellular and dedicated modes, only the time allocation vector $\boldsymbol{\tau}$ is optimized, and in the two reuse modes, both τ_1 and P_d are optimized.

7.1.3.1 Cellular Mode

In cellular mode, the throughput maximization problem is formulated as

$$\text{Maximize}_{\boldsymbol{\tau}} \qquad \mathbf{R}_{sum} = \mathbf{R}_C + \mathbf{R}_B + \mathbf{R}_D \tag{7.21}$$

$$\text{Subject to} \qquad \|\boldsymbol{\tau}\|_1 = 1 \tag{7.22}$$

$$\mathbf{R}_j \geq \mathbf{R}_{j\,min}, \quad \text{for } j = C, B, D \tag{7.23}$$

where $\mathbf{R}_{Cmin}$, $\mathbf{R}_{Dmin}$ and $\mathbf{R}_{Dmin}$ are the minimum rates required for the uplink, downlink and D2D link, respectively. This optimization problem can be solved through auction game approach. Auction game has been studied in the game theory literature,

Table 7.1: Algorithm 7.1

Resource allocation for the cellular mode

1. **Initialization**: Initialize τ_1 and τ_2 by solving $\mathbf{R}_C = \mathbf{R}_{Cmin}$ and $\mathbf{R}_B = \mathbf{R}_{Bmin}$. $\hat{\tau}$ is given by the solution of $-\frac{1}{\theta_D}\log\mathbb{E}\{e^{-\hat{\tau}\theta_D r_{D,B}}\} = \mathbf{R}_{Dmin}$. With this $\hat{\tau}$ value, solve τ_3 from (7.10), and denote this solution by $\tau_{3,1}$. If $\theta_D \geq \theta_{BD}$, solve τ_3 from (7.11), otherwise solve τ_3 from (7.12), and denote this solution by $\tau_{3,2}$. Set $\tau_3 = \max\{\tau_{3,1}, \tau_{3,2}\}$.

   Check if $\|\boldsymbol{\tau}\|_1 > 1$, then end the allocation process.

2. **Auction**: Divide the remaining time resource equally into N parts, and set $\Delta\tau = (1 - \tau_1 - \tau_2 - \tau_3)/N$. Set the bids of the uplink, downlink and D2D link as $\text{Bid}_C = \mathbf{R}_C(\tau_1 + \Delta\tau) - \mathbf{R}_C(\tau_1)$, $\text{Bid}_B = \mathbf{R}_B(\tau_2 + \Delta\tau) - \mathbf{R}_B(\tau_2)$ and $\text{Bid}_D = \mathbf{R}_D(\tau_3 + \Delta\tau) - \mathbf{R}_D(\tau_3)$, respectively.

   For $i = 1 : N$

   (a) Select the link with the highest bid as the winner of the i^{th} round, and increase its time allocation parameter by $\Delta\tau$.

   (b) Update the winner's bid with its new time allocation parameter.

   end

and it has been used to design algorithms for resource allocation in D2D models. For example, in [47], spectrum allocation problem was solved through an iterative combinatorial auction approach. In our setting, we build our auction game as follows: The time resource is divided into small time slots (of the same duration), which are regarded as resource units. In each round, three bidders, which are uplink, downlink and D2D link, compete for a single resource unit. Bidders bid according to the throughput increment they gain with an additional time slot, and the resource unit will be allocated to the link giving the highest bid. The detailed algorithm is provided in Table 7.1.

The initialization step guarantees the rate constraints. If the rate constraints cannot be satisfied, we have $\|\boldsymbol{\tau}\|_1 > 1$. It is immediately seen that this iterative algorithm is completed within finite time, and the system only needs to evaluate

effective capacity $N+3$ times, where N is the number of resource units. Therefore, the complexity of this algorithm does not depend on the number of bidders.

The performance of this auction algorithm is evaluated via numerical results. We generate the location of the nodes randomly, and compare the optimal throughput values obtained from Algorithm 7.1 and exhaustive search. The normalized error is defined as $\varepsilon = \frac{|\mathbf{R}_{\text{sum,auction}} - \mathbf{R}_{\text{sum,search}}|}{\mathbf{R}_{\text{sum,search}}}$. Over 10^5 rounds of testing, the average normalized error is 2.01×10^{-5}. In each time, N is selected as the smallest integer that makes the step $\Delta\tau \leq 5 \times 10^{-3}$.

7.1.3.2 Dedicated Mode

In the dedicated mode, the optimization problem has the same formulation as in the cellular mode. Different from the cellular mode case, the throughput maximization problem can be solved by convex optimization algorithms.

Theorem 21 *In the dedicated mode, the sum rate* $\mathbf{R}_{sum} = \mathbf{R}_C + \mathbf{R}_B + \mathbf{R}_D$ *is concave with respect to the time allocation vector* $\boldsymbol{\tau}$.

Proof: See Appendix A.14

Also, the constraints define a convex region in the $\tau_1 - \tau_2 - \tau_3$ space. Therefore, the throughput maximization problem in the dedicated mode is a concave maximization problem, which can be solved efficiently by convex optimization algorithms.

7.1.3.3 Reuse Modes

In the two reuse modes, we need to optimize the throughput over τ_1 and the D2D transmission power P_d, and the optimization problem is formulated as

$$\begin{aligned} \text{Maximize}_{\tau_1, P_d} \quad & \mathbf{R}_{sum} = \mathbf{R}_C + \mathbf{R}_B + \mathbf{R}_D && (7.24) \\ \text{Subject to} \quad & 0 \le P_d \le P_{dmax} && (7.25) \\ & \mathbf{R}_j \ge \mathbf{R}_{j\,min}, \quad \text{for } j = C, B, D && (7.26) \end{aligned}$$

Due to the interference between the D2D and cellular users, the objective function is not concave. To solve this problem, we search for the optimal τ_1 from 0 to 1. For a given value of τ_1, we determine the optimal P_d through the successive convex approximation (SCA) algorithm. If the rate constraints cannot be all satisfied, the sum rate is set to 0.

The idea of SCA was proposed in [91]. In this approach, the nonconvex optimization problem is transformed into a series of convex optimization problems. More specifically in our model, we replace the sum throughput by a series of concave functions of P_d, which are denoted as $\{U_l\}$ for $l = 1, 2, \cdots$. In the uplink reuse mode, since $\mathbf{R}_B$ does not depend on P_d, we only need to consider the maximization of $\mathbf{R} = \mathbf{R}_C + \mathbf{R}_D$, and we construct $\{U_l\}$ as

$$U_l = \mathbf{R}_D(P_d) + \mathbf{R}_C(P_d^{l-1}) + (P_d - P_d^{l-1}) \left. \frac{\partial \mathbf{R}_C}{\partial P_d} \right|_{P_d = P_d^{l-1}}$$

where P_d^{l-1} is the optimal P_d value in the $(l-1)^{\text{th}}$ iteration, and the derivative of $\mathbf{R}_C$ with respect to P_d is given by

$$\frac{\partial \mathbf{R}_C}{\partial P_d} = -\frac{1-\tau_1}{\mathbb{E}\{e^{-(1-\tau_1)\theta_C r_{C,B}}\}} \mathbb{E}\left\{ e^{-(1-\tau_1)\theta_C r_{C,B}} \frac{B P_c z_1 z_2}{(P_d z_2 + B\sigma_0^2)(P_d z_2 + P_c z_1 + B\sigma_0^2) \log_e 2} \right\}. \quad (7.27)$$

Table 7.2: Algorithm 7.2

SCA algorithm for power allocation in the uplink reuse mode

1. If $\mathbf{R}_B < \mathbf{R}_{Bmin}$, set $\mathbf{R}_{sum} = 0$, end process.
2. Find the lower bound of P_d by solving $\mathbf{R}_D = \mathbf{R}_{Dmin}$, and denote this lower bound by P_{dl}. Find an upper bound of P_d by solving $\mathbf{R}_C = \mathbf{R}_{Cmin}$, and denote this upper bound by $\hat{P}_{du}$. Set $P_{du} = \min\{\hat{P}_{du}, P_{dmax}\}$ as the upper bound of P_d. If $P_{dl} > P_{du}$, set $\mathbf{R}_{sum} = 0$, end process.
3. Set $l = 1$ and select an initial value P_d^0 between P_{dl} and P_{du}.

   Repeat

   (a) Maximize U_l with the constraint $P_{dl} \leq P_d \leq P_{du}$.

   (b) Update P_d^l, and increase l by 1.

   Until convergence

   The optimal P_d value is given by the P_d^l in the last iteration, and the maximum sum rate is given by $\mathbf{R}_{sum} = U_{opt} + \mathbf{R}_B$, where U_{opt} is the optimal U_l value in the last iteration.

For a given τ_1, our SCA algorithm is described in Table 7.2, and the overall resource allocation algorithm for the uplink reuse mode is given in Table 7.3. The optimization algorithm for the downlink reuse mode can be obtained through a similar process.

7.1.3.4 Mode Selection

In previous subsections, we have introduced the throughput formulation and resource allocation algorithms for all 4 possible modes. In its overall operation, the system runs the throughput maximization algorithms for these 4 modes, and selects the one with the highest optimal sum throughput. Since effective capacity is a long-term throughput criterion, the system does not need to conduct mode selection frequently.

Our resource allocation algorithms can also be extended to the case of multiple cellular users without much difficulty. Since the number of cellular users only changes

Table 7.3: Algorithm 7.3

Resource allocation in the uplink reuse mode

1. Set a step $\Delta\tau$ for the searching algorithm, and set the optimal τ_1, P_d and $\mathbf{R}_{sum}$ values as $\tau_1^* = 0$, $P_d^* = 0$ and $\mathbf{R}_{sum}^* = 0$.
2. **For** $\tau_1 = 0 : \Delta\tau : 1$
   (a) Maximize the sum rate using the SCA algorithm. Denote the maximum sum rate and optimal P_d as $\hat{\mathbf{R}}_{sum}$ and $\hat{P}_d$, respectively.
   (b) If $\hat{\mathbf{R}}_{sum} > \mathbf{R}_{sum}^*$, then update $\mathbf{R}_{sum}^* = \hat{\mathbf{R}}_{sum}$, $P_d^* = \hat{P}_d$ and $\tau_1^* = \tau_1$.

   end

the dimensionality of the optimization problem, the algorithms given in the previous subsection can still be employed.

7.1.4 Numerical Results

In this subsection, we further analyze the mode selection for the D2D users in the considered cellular network through numerical results. In all numerical results, we consider Rayleigh fading with path loss $\mathbb{E}\{z\} = d^{-4}$, where d is the distance between the transmitter and receiver, and z is the magnitude square of the corresponding fading coefficient. The position of the base station is fixed at the origin of the coordinate plane, and we set $P_c = P_{dmax} = 500$, $P_b = 1200$. For the rate constraints, we set $\mathbf{R}_{Cmin} = \mathbf{R}_{Dmin} = 0.308$ bit/s and $\mathbf{R}_{Bmin} = 0.341$ bit/s [1]. Also, we set all θ values to 1. In the figures, we use dark blue points to indicate the positions of the base station, $\mathbf{D}_1$ and $\mathbf{D}_2$. The color of each pixel represents the selected mode if the cellular user is in that position, and we denote the cellular mode, dedicated mode, uplink reuse mode and downlink reuse mode by I, II, III and IV, respectively in the color bars.

[1]The values of $\mathbf{R}_{Cmin}$, $\mathbf{R}_{Bmin}$ and $\mathbf{R}_{Dmin}$ are selected as the throughput of each link in the dedicated mode, when the distance between the transmitter and receiver pairs is 4, and $\tau_1 = \tau_2 = \tau_3 = 0.25$.

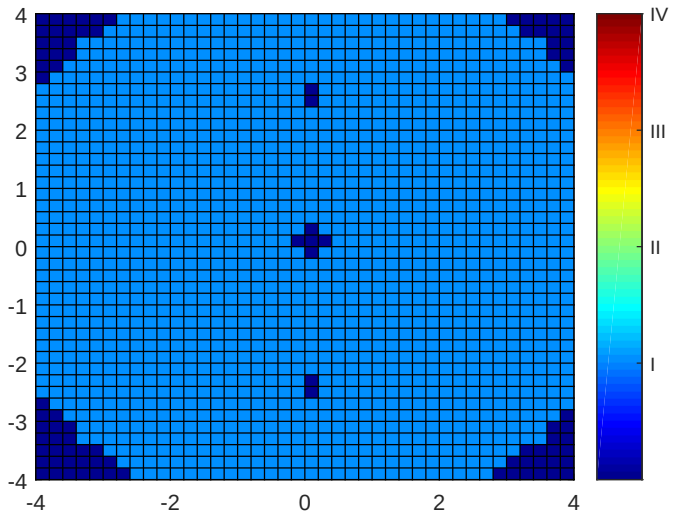

Figure 7.2: Mode selection result when $\mathbf{D}_1$ is placed at $(0, -2.5)$, and $\mathbf{D}_2$ is placed at $(0, 2.5)$. Cellular mode, dedicated mode, uplink reuse mode and downlink reuse mode are denoted by I, II, III and IV, respectively.

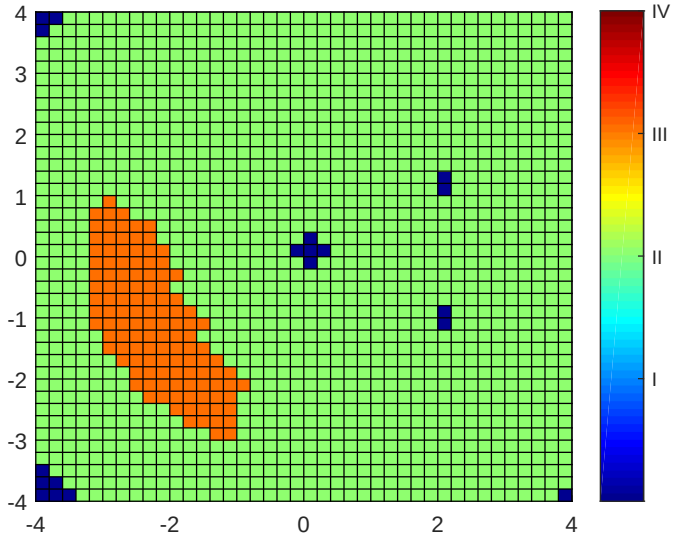

Figure 7.3: Mode selection result when $\mathbf{D}_1$ is placed at $(2, -1)$, and $\mathbf{D}_2$ is placed at $(2, 1)$. Cellular mode, dedicated mode, uplink reuse mode and downlink reuse mode are denoted by I, II, III and IV, respectively.

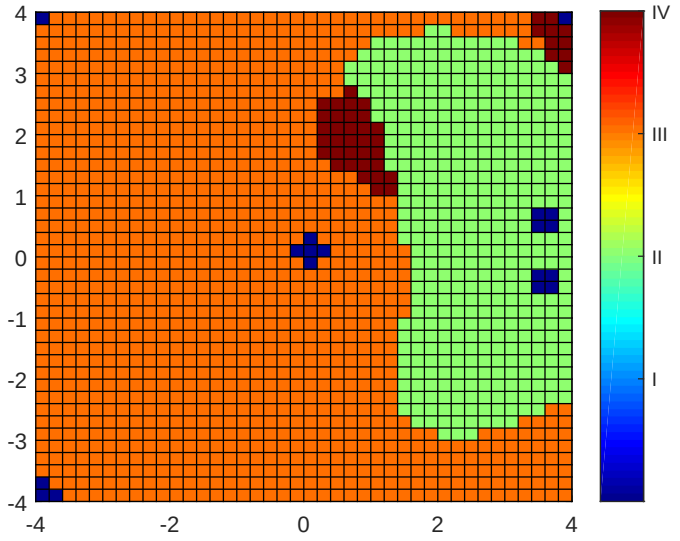

Figure 7.4: Mode selection result when $\mathbf{D}_1$ is placed at $(3.5, -0.5)$, and $\mathbf{D}_2$ is placed at $(3.5, 0.5)$. Cellular mode, dedicated mode, uplink reuse mode and downlink reuse mode are denoted by I, II, III and IV, respectively.

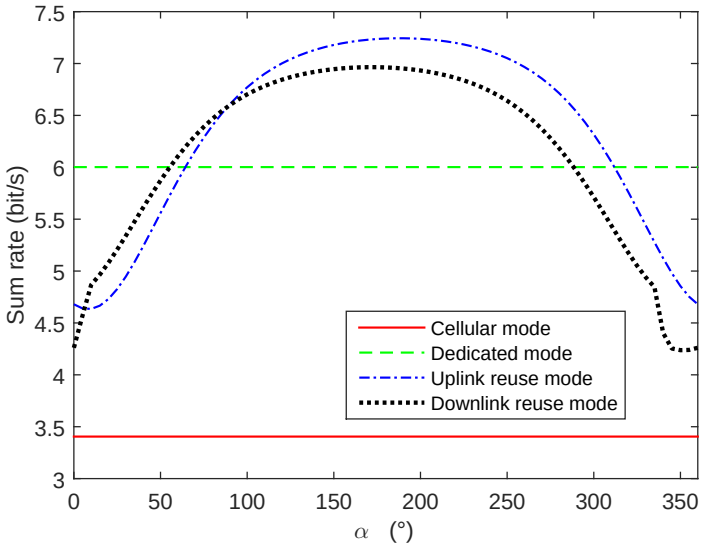

Figure 7.5: Sum rate vs. angle α

In Fig. 7.2, we place $\mathbf{D}_1$ and $\mathbf{D}_2$ at $(0, -2.5)$ and $(0, 2.5)$ respectively. In this situation, only cellular mode can be selected, because the relatively large distance between $\mathbf{D}_1$ and $\mathbf{D}_2$ makes it difficult for the rate constraint of the D2D users to be satisfied. Cellular mode is preferred when the $\mathbf{D}_1 - \mathbf{BS}$ link and $\mathbf{BS} - \mathbf{D}_2$ link are much stronger than direct link $\mathbf{D}_1 - \mathbf{D}_2$. The dark blue regions on the corners represent the region outside the range of the cell.

In Fig. 7.3, $\mathbf{D}_1$ and $\mathbf{D}_2$ get closer to each other, and both of them are shifted to the right. Since the direct link becomes stronger, cellular mode is no longer preferred. In this situation, most of the region prefers dedicated mode. Dedicated mode is preferred when the D2D direct link and interference links are strong. There is a small region on the lower left, in which uplink reuse mode is selected. In this region, the interference link $\mathbf{CU} - \mathbf{D}_2$ becomes weaker, which results in the uplink reuse mode to have higher throughput than the dedicated mode. When $\mathbf{CU}$ further moves towards to the bottom left corner, the uplink becomes too weak to satisfy the rate constraint, and dedicated mode is selected.

In Fig. 7.4, $\mathbf{D}_1$ and $\mathbf{D}_2$ are very close to each other, and both of them are far away from the base station. In this situation, a large area on the left side prefers uplink reuse mode, because the D2D direct link and uplink are stronger than the interference links. Reuse modes are preferred when the direct links are all much stronger than the interference links. When the cellular user is close to the D2D users, there is still a large region, where dedicated mode is preferred. Since P_b is much larger than P_c and P_{dmax}, the interference arising from the base station is very strong, which makes downlink reuse mode unfavorable. We let the cellular user move around the circle $x^2 + y^2 = 4$, and we denote the angle between line $\mathbf{CU} - \mathbf{BS}$ and the positive direction of the x axis as α. We plot the sum rates of these 4 modes as functions of the angle α in Fig. 7.5. When $\alpha = 190°$, the distance between $\mathbf{CU}$ and $\mathbf{D}_2$ is maximal, and the uplink reuse mode achieves the maximum sum rate. When $\alpha = 170°$, the distance between

$\mathbf{CU}$ and $\mathbf{D}_1$ is maximal, and the downlink reuse mode reaches its own maximum sum rate. Moreover, the downlink reuse mode provides the highest sum rate when $60^\circ < \alpha < 90^\circ$, verifying the observation in Fig. 7.4. Around the upper right corner of Fig. 7.4, there is also a small region that prefers downlink reuse mode. In this region, the received SINR in downlink is higher than that in uplink, because P_b is much greater than P_c, and the interference link $\mathbf{CU} - \mathbf{D}_1$ is also relatively weak.

7.2 Joint Mode Selection and Resource Allocation for D2D Communications under Queuing Constraints

7.2.1 System Model and Transmission Modes

7.2.1.1 System Model

We consider a cellular network with one base station ($\mathbf{BS}$), N_1 cellular users $\{\mathbf{CU}_1, \mathbf{CU}_2, \cdots, \mathbf{CU}_{N_1}\}$ and N_2 D2D pairs $\{(\mathbf{DT}_1, \mathbf{DR}_1), (\mathbf{DT}_2, \mathbf{DR}_2), \cdots, (\mathbf{DT}_{N_2}, \mathbf{DR}_{N_2})\}$. We assume that the D2D transmission is one-way, in which $\mathbf{DT}_i$ and $\mathbf{DR}_i$ represent the transmitter and receiver of the i^{th} D2D pair, respectively. Each cellular user transmits to the base station through an uplink channel, and receives data from the base station via a downlink channel. For all transmitters, the data packets are stored in buffers before being sent to the corresponding receiver. We assume that all transmitters operate under statistical queuing constraints, which require the buffer overflow probability to decay exponential fast with QoS exponent θ. The QoS exponent at $\mathbf{CU}_i$ and $\mathbf{DT}_i$ are denoted by θ_{C_i} and θ_{D_i}, respectively. For simplicity, we assume that there are $N_1 + N_2$ separate buffers at the base station. N_1 of them are used for the downlink transmission, and the QoS exponent of the $\mathbf{BS} - \mathbf{CU}_i$ link is θ_{BC_i}. The

remaining N_2 buffers are for the data of the D2D users operating in the cellular mode, in which the base station acts as a relay. The QoS exponent for the base-station buffer storing the data to be sent via the $\mathbf{BS}-\mathbf{DR}_i$ link is θ_{BD_i}. Since not all D2D users operate in the cellular mode, only a portion of the buffers are used at the base station.

The channel fading is assumed to be block fading, in which the fading coefficients h stay constant in one time block and change independently across blocks. In Figs. 7.6, 7.7 and 7.8, the magnitude-square of the fading coefficients are denoted by $z = |h|^2$. At each receiver, the background noise is assumed to follow an independent complex Gaussian distribution with zero mean and variance σ^2, i.e., $n \sim \mathcal{CN}(0, \sigma^2)$.

There are N available orthogonal channels for this cellular network, each of them having a bandwidth of B. In the resource allocation step, these orthogonal channels are allocated to the transmission links, and those links to which channels are not assigned are not activated for communication. There are five assumptions regarding the channel allocation:

1. A D2D pair operating in the cellular mode occupies only one channel, and cannot share its channel with other users.

2. Each cellular uplink or downlink is allocated a single orthogonal channel, and channels cannot be shared by different cellular links.

3. Each D2D direct link, cellular uplink and downlink can occupy at most one channel.

4. Each orthogonal channel can be occupied by two transmission links at most.

5. Cellular links have higher priority to be assigned a channel than D2D links.

In this section, we mainly consider the case in which $N \geq 2N_1$, so that we have sufficient number of channels to guarantee the performance requirements of all cellular

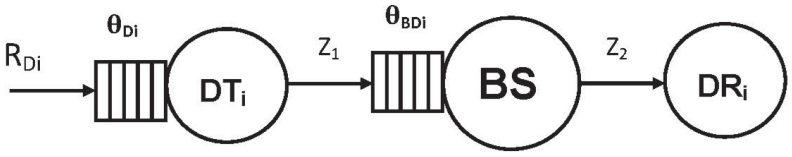

Figure 7.6: System model in the D2D cellular mode

users. The transmission power of a cellular user is set at P_c, and the maximum transmission power of D2D transmitters is P_{dmax}. When acting as a transmitter, the base station transmits with power P_b in each channel. Therefore, the overall transmission power of the base station depends on the number of cellular users and the number of D2D pairs operating in the cellular mode.

The primary objective in this section is to maximize the overall system throughput under statistical queuing constraints by using our joint mode selection and channel allocation algorithm. In the following subsection, we introduce all possible transmission modes and describe the relationship between mode selection and channel allocation.

7.2.1.2 Transmission Modes and Instantaneous Transmission Rate

In this section, our mode selection is done through channel allocation. Depending on how the system uses each channel, we can determine the transmission mode for each user. For instance, if a channel is allocated to a cellular uplink and a D2D direct link, then the corresponding D2D pair and cellular user are in the uplink reuse mode. According to our channel allocation assumptions, there are 7 possible modes for the users in each channel, namely D2D cellular mode, uplink reuse mode, downlink reuse mode, D2D reuse mode, uplink dedicated mode, downlink dedicated mode, and D2D dedicated mode. All of these modes can be summarized into 3 categories.

D2D Cellular Mode

The first category is D2D cellular mode, and the model is shown in Fig. 7.6.

In this mode, the channel is occupied by a D2D pair, transmitting with the help of the base station. In this mode, each transmission period is equally divided into two phases. The first phase is allocated to the $\mathbf{DT}_i - \mathbf{BS}$ link, in which the D2D transmitter sends information to the base station, and the second phase is allocated to the $\mathbf{BS} - \mathbf{DR}_i$ link, in which the base station forwards the information to the D2D receiver. Denote the fraction of time allocated to the first phase as τ, and the fraction of time allocated to the second phase as $1 - \tau$. In each phase, there is only one transmitter and one receiver, and the received signal at each transmitter follows the form

$$y = hx + n, \tag{7.28}$$

where x is the transmitted signal, n is the additive Gaussian noise component, h is the corresponding channel fading coefficient. We can express the instantaneous rates of $\mathbf{DT}_i - \mathbf{BS}$ link and $\mathbf{BS} - \mathbf{DR}_i$ link as

$$r_{D_i,BS}(\tau) = \tau B \log_2 \left(1 + \frac{P_{dmax}}{B\sigma^2} z_1\right) \tag{7.29}$$

and

$$r_{BS,D_i}(\tau) = (1-\tau) B \log_2 \left(1 + \frac{P_b}{B\sigma^2} z_2\right), \tag{7.30}$$

respectively.

In this two-hop model, to guarantee the stability, the average arrival rate should be less than or equal to the average departure rate from the buffer at the base station, which can be expressed as $\mathbb{E}\{r_{D_i,BS}(\tau)\} \leq \mathbb{E}\{r_{BS,D_i}(\tau)\}$. Suppose that τ_0 is the solution of $\mathbb{E}\{r_{D_i,BS}(\tau)\} = \mathbb{E}\{r_{BS,D_i}(\tau)\}$, then the effective capacity of the two-hop

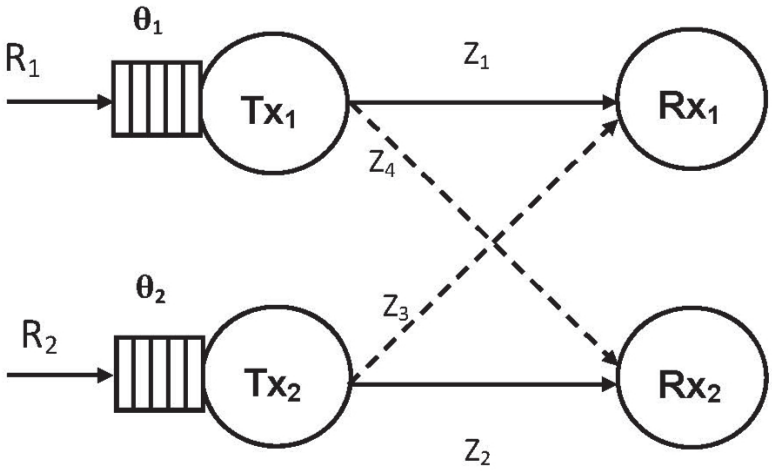

Figure 7.7: System model in the reuse mode (interference links are denoted by the dashed lines)

channel $\mathbf{DT}_i - \mathbf{BS} - \mathbf{DR}_i$ is given by

$$\mathbf{R}_{D_i} = -\frac{1}{\theta_{D_i}} \log \mathbb{E}\{e^{-\theta_{D_i} r_{D_i,BS}(\hat{\tau})}\} \tag{7.31}$$

where $\hat{\tau} = \min\{\tau_0, \tau^*\}$, and τ^* is the solution to

$$-\frac{1}{\theta_{D_i}} \log \mathbb{E}\{e^{-\theta_{D_i} r_{D_i,BS}(\tau)}\} = -\frac{1}{\theta_{BD_i}} \log \mathbb{E}\{e^{-\theta_{BD_i} r_{BS,D_i}(\tau)}\} \tag{7.32}$$

when $\theta_{BD_i} \leq \theta_{D_i}$, or

$$-\frac{1}{\theta_{D_i}} \big[\log \mathbb{E}\{e^{-\theta_{BD_i} r_{BS,D_i}(\tau)}\} + \log \mathbb{E}\{e^{(\theta_{BD_i} - \theta_{D_i}) r_{D_i,BS}(\tau)}\} \big] = -\frac{1}{\theta_{D_i}} \log \mathbb{E}\{e^{-\theta_{D_i} r_{D_i,BS}(\tau)}\} \tag{7.33}$$

when $\theta_{BD_i} > \theta_{D_i}$, which comes from the results in [17].

Reuse Mode

The second category is the reuse mode, and the system model is shown in Fig. 7.7. In this model, two transmitter-receiver pairs share the same channel, and they inflict interference on each other. In Fig. 7.7, two interference links (z_3 and z_4 links) are

denoted by the dashed line. According to the types of users sharing the channel, reuse mode includes uplink reuse mode, downlink reuse mode, and D2D reuse mode. In the uplink reuse mode, a cellular uplink shares the channel with a D2D direct link; in the downlink reuse mode, a cellular downlink shares the channel with a D2D direct link; in D2D reuse mode, two pairs of D2D users transmit in the same channel.

The received signal at each receiver follows the form

$$y = hx + h_{\text{inter}}x_{\text{inter}} + n, \tag{7.34}$$

where x is the desired signal, h is the fading coefficient of the channel between this receiver and its corresponding transmitter, x_{inter} is the interference signal, h_{inter} is the fading coefficient of the interfering link, and n is the Gaussian noise. Treating the interference as noise, we can express the instantaneous transmission rates in these two links as

$$r_1 = B\log_2\left(1 + \frac{P_1}{B\sigma^2 + P_2 z_3}z_1\right) \tag{7.35}$$

and

$$r_2 = B\log_2\left(1 + \frac{P_2}{B\sigma^2 + P_1 z_4}z_2\right), \tag{7.36}$$

respectively.

In (7.35) and (7.36), P_1 and P_2 denote the transmission powers of $\mathbf{Tx_1}$ and $\mathbf{Tx_2}$, respectively. For the base station and cellular users, the transmission powers are fixed at P_b and P_c, respectively. For the D2D users, the transmission power is determined by the average SINR constraints. To control the interference, the users operating in reuse modes have to satisfy the average SINR constraints, which set lower bounds on the SINR values at the receivers. If two transmission links cannot satisfy the average SINR constraints, then they are not allowed to share the same channel, i.e., the reuse mode will not be permitted for these links.

In the uplink reuse mode, we suppose that $\mathbf{Tx_1}$ is the cellular user $\mathbf{CU}_i$, $\mathbf{Rx_1}$ is

the base station, $\mathbf{Tx_2}$ is the D2D user $\mathbf{DT}_j$, and $\mathbf{Rx_2}$ is the D2D user $\mathbf{DR}_j$. Then, the SINR constraints can be expressed as

$$\begin{cases} \mathbb{E}\left\{\frac{P_c}{B\sigma^2+P_{d_j}z_3}z_1\right\} \geq \gamma_c \\ \mathbb{E}\left\{\frac{P_{d_j}}{B\sigma^2+P_C z_4}z_2\right\} \geq \gamma_d \\ P_{d_j} \leq P_{dmax} \end{cases} \tag{7.37}$$

where P_{d_j} is the transmission power of $\mathbf{DT}_j$, γ_c and γ_d are the SINR thresholds of cellular and D2D users, respectively. Inequality group (7.37) provides us upper and lower bounds on P_{d_j}, hence describing an interval of P_{d_j} values that satisfy the average SINR constraints. Similar formulations and characterizations can be obtained for the downlink reuse mode.

In the D2D reuse mode, we have to determine the transmission power for the two D2D transmitters. Suppose that $\mathbf{Tx_1}$ is the D2D user $\mathbf{DT}_i$, $\mathbf{Rx_1}$ is the D2D user $\mathbf{DR}_i$, $\mathbf{Tx_2}$ is the D2D user $\mathbf{DT}_j$, and $\mathbf{Rx_2}$ is the D2D user $\mathbf{DR}_j$. Then, the SINR constraints can be expressed as

$$\begin{cases} \mathbb{E}\left\{\frac{P_{d_i}}{B\sigma^2+P_{d_j}z_3}z_1\right\} \geq \gamma_d \\ \mathbb{E}\left\{\frac{P_{d_j}}{B\sigma^2+P_{d_i}z_4}z_2\right\} \geq \gamma_d \\ P_{d_i} \leq P_{dmax} \\ P_{d_j} \leq P_{dmax} \end{cases} \tag{7.38}$$

where P_{d_i} and P_{d_j} are the transmission powers of $\mathbf{DT}_i$ and $\mathbf{DT}_j$, respectively. Inequality group (7.38) provides a region on the $P_{d_i} - P_{d_j}$ plane, and each (P_{d_i}, P_{d_j}) pair in this region satisfies the average SINR constraints.

For the uplink, downlink and D2D reuse modes, the optimal transmission power of the D2D transmitter has to be identified by searching over the region determined by

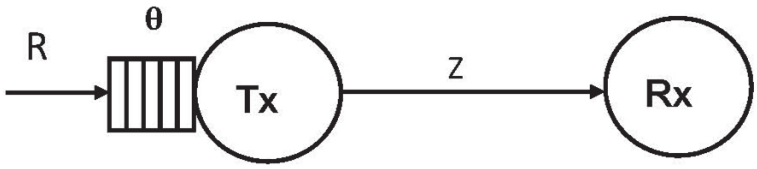

Figure 7.8: System model in the dedicated mode

the average SINR constraints. Inserting the rate expressions given in (7.35) and (7.36) with the optimal transmission power into (2.7), we can get the effective capacities of these two links.

Dedicated Mode

The third category is the dedicated mode, which is depicted in Fig. 7.8. In this model, one transmitter-receiver pair occupies a channel without sharing it with others. Depending on the type of the transmission link occupying the channel, this category includes uplink dedicated mode, downlink dedicated mode, and D2D dedicated mode, in which the channel is occupied by a cellular uplink, a cellular downlink, and a direct D2D link, respectively. Since there is no interference, the received signal at the receiver also follows the form given in (7.28), and the instantaneous rate can be expressed as

$$r = B \log_2 \left(1 + \frac{P}{B\sigma^2} z \right), \tag{7.39}$$

where P is the transmission power. In this mode, the transmission powers of cellular users, base station and D2D transmitters are fixed at P_c, P_b and P_{dmax}, respectively.

For the uplink, downlink and D2D dedicated modes, inserting the instantaneous rates given by (7.39) into (2.7), we obtain the effective capacity for the corresponding transmission link.

7.2.2 Channel Allocation via Maximum-Weight Matching Approach

From the above analysis, we can determine that the essence of mode selection is channel allocation. To identify the optimal modes for all users, we only need to come up with a channel matching rule that maximizes the overall throughput. This problem can be modeled as a maximum-weight matching problem. In our setting, we seek to match the transmission links to the channels. The weight/gain of matching a transmission link with a channel is given by the effective capacity of that link.

There are four main challenges to solve this maximum-weight matching problem as described below:

1. This problem involves both one-to-one matching and two-to-one matching.
    - This is due to the fact that there are three kinds of reuse modes, in which two transmission links are matched to a single channel.
2. The weights are not independent of each other in the case of two-to-one matching.
    - Due to the interference in three reuse modes, the throughput of each link also depends on the other link with which it shares the channel.
3. Even for the case of one-to-one matching, the weights are not constant.
    - This occurs when a D2D pair uses the channel exclusively. The D2D pair has two choices, which are the D2D cellular mode and D2D dedicated mode. The link throughput varies depending on which mode the user is in.
4. Cellular links have higher priority than D2D links.
    - In conventional matching problems, there are no priorities.

To solve these problems, we propose a new channel matching algorithm, which transforms the original problem into a one-to-one maximum-weight matching problem. In the following subsections, we introduce our new algorithm step by step. Before the matching process, we enumerate the N_1 cellular users, N_2 D2D users and N channels to distinguish them.

7.2.2.1 Channel Allocation for Cellular links

The first step is to assign the cellular uplink and downlink to the channels. Since cellular links have higher priority, we can give each of them a channel at the very beginning. In this step, we allocate the 1^{st} to N_1^{th} channel to the cellular uplink, and allocate the $(N_1+1)^{\text{th}}$ to $2N_1^{\text{th}}$ channel to the cellular downlink. After this step, there are $N-2N_1$ free channels left. Now the problem becomes matching N_2 D2D pairs to the N channels, which is a one-to-one matching problem if the D2D reuse mode is excluded. If a D2D pair gets a channel assigned also to a cellular link, then they operate in the corresponding reuse mode; if a D2D pair gets a free channel, it can choose from the D2D reuse mode, D2D dedicated mode and D2D cellular mode. Now, the first and last challenges above are overcome. In the next step, we construct the throughput matrix.

7.2.2.2 Constructing the Throughput Matrix

The structure of the $(N_2+2N_1)\times N$ throughput matrix is shown in Fig. 7.9. Each row corresponds to a D2D pair, and each column corresponds to a channel. Each element of this matrix equals to the channel throughput if the corresponding channel and D2D pair are matched together. To include the uplink and downlink dedicated modes, we introduce $2N_1$ dummy D2D pairs with enumerations varying from N_2+1 to N_2+2N_1, and assume that they do not perform transmissions. If a dummy D2D pair is matched to the channel with a cellular uplink/downlink, then the cellular link

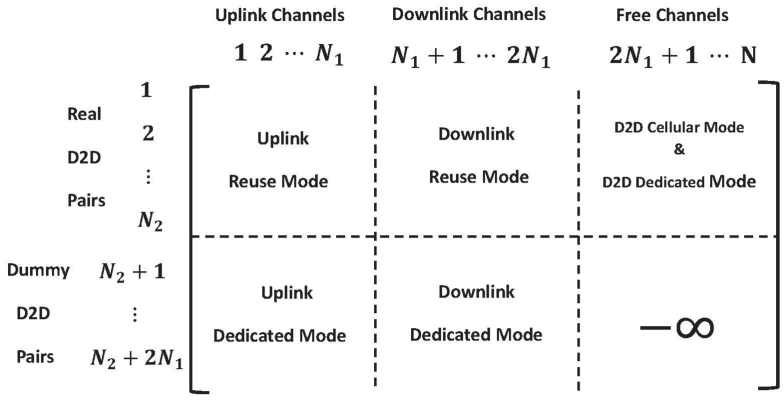

Figure 7.9: Structure of the throughput matrix

works in uplink/downlink dedicated mode. The matrix can be divided into 6 regions, which correspond to the 6 modes.

Suppose $R_{i,j}$ denotes the element of the throughput matrix in the i^{th} row and j^{th} column, then $R_{i,j}$ is given by the maximum throughput that can be achieved in channel j if we match D2D pair i to it. Here, the throughput of a channel is the sum effective capacity of the transmission links in this channel. For the reuse modes, we have to search for the optimal transmission power for the D2D transmitters. If a D2D pair gets a free channel, then it compares the effective capacities of D2D dedicated and D2D cellular modes, and chooses the one giving a higher throughput. Since we do not want to match a dummy D2D pair to a free channel, we set the elements in the corresponding region (in the lower right corner) to negative infinity. Since the D2D reuse mode involves two-to-one matching, we exclude the D2D reuse mode in this step, and consider it in the next step.

At the end of this step, we have successfully transformed the original problem to a conventional maximum-weight matching problem, and all matching weights are independent from each other. Hence, the second and third challenges are addressed. Assume that $\zeta_{i,j}$ is a parameter that indicates whether the i^{th} row and j^{th} column are matched. If they are matched, then $\zeta_{i,j} = 1$, otherwise $\zeta_{i,j} = 0$. Then, the matching

problem can be formulated as

$$\max_{\zeta_{i,j}} \quad \sum_i \sum_j R_{i,j}\zeta_{i,j} \tag{7.40}$$

$$\text{Subject to} \quad \sum_i \zeta_{i,j} \leq 1, \forall j \tag{7.41}$$

$$\sum_j \zeta_{i,j} \leq 1, \forall i \tag{7.42}$$

This problem can be solved by the Hungarian algorithm (Kuhn-Munkres algorithm) [92]. After applying the Hungarian algorithm, we get a result for both mode selection and channel allocation without considering the D2D reuse mode. We pick out all D2D pairs working in the D2D dedicated mode, enumerate them from 1 to M_1, and also enumerate the D2D pairs without any channel assignments from 1 to M_2, where M_1 is the number of the D2D pairs in D2D dedicated mode, and M_2 is the number of D2D pairs which were not assigned channels. In the next step, we conduct the matching just for the D2D reuse mode.

7.2.2.3 Channel Allocation for the D2D Reuse Mode

In this last step, we just seek to match the D2D pairs in D2D dedicated mode with the D2D pairs without any channel assignments. Similar to the previous step, we form another $(M_1 + M_2) \times M_1$ throughput matrix, which is shown in Fig. 7.10. In this matching step, those D2D pairs in the D2D dedicated mode choose whether to share their channel with another D2D pair. In order to give the right to the D2D users in the D2D dedicated mode to refuse sharing channel with others for the purposes of throughput maximization, we have to include D2D dedicated mode by introducing M_1 dummy D2D pairs with enumerations from $M_2 + 1$ to $M_2 + M_1$. After applying the Hungarian algorithm to this throughput matrix, the channel allocation and mode selection are accomplished.

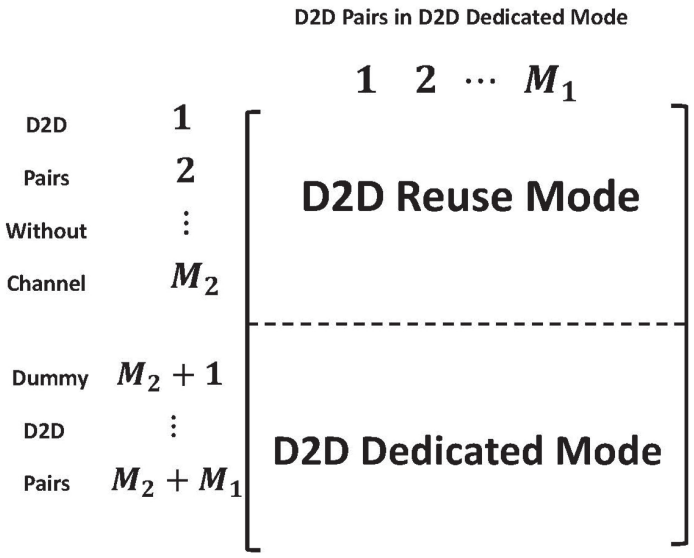

Figure 7.10: Structure of the throughput matrix for the D2D reuse mode

7.2.3 Numerical Results

In this subsection, we further investigate the performance of our proposed channel matching algorithm via numerical results. In the simulation, we set the overall channel number as $N = 6$, the number of cellular users $N_1 = 2$, all $\theta_C = \theta_D = 1$, $\theta_{BD} = \theta_{BC} = 2$, $P_c = P_{dmax} = 500$, $P_b = 600$. For each cellular uplink, the SINR threshold is $\gamma_c = 1.95$; for each downlink, the SINR threshold is $\gamma_c = 2.34$; and for each D2D link, the SINR threshold is $\gamma_d = 0.39$. We assume Rayleigh fading with path loss $\mathbb{E}\{z\} = d^{-4}$, where d is the distance between the corresponding transmitter and receiver. In the numerical results, we choose random channel matching as the benchmark, and compare its performance with our improved channel matching algorithm. In the random matching algorithm, we randomly allocate the channel to the D2D users after assigning $2N_1$ channels to the cellular links, and repeat N_2^3 times and pick the best matching results. Hence, in the results of random matching, there is still an attempt to optimize the performance by finite attempts but without using any particular structure.

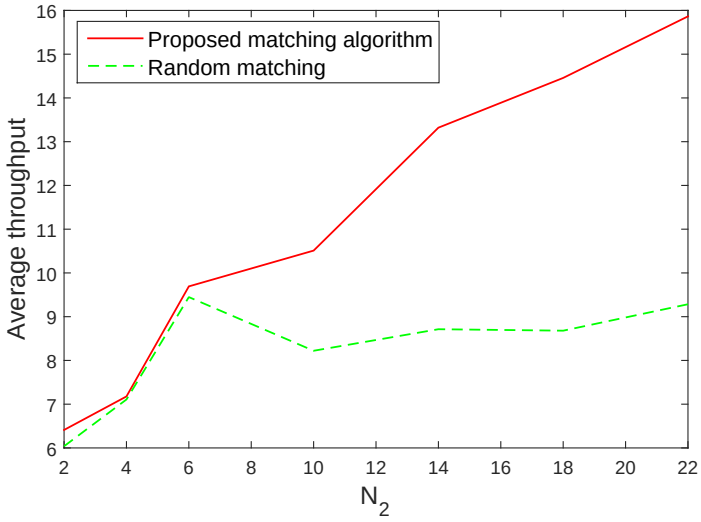

Figure 7.11: Average throughput vs. N_2

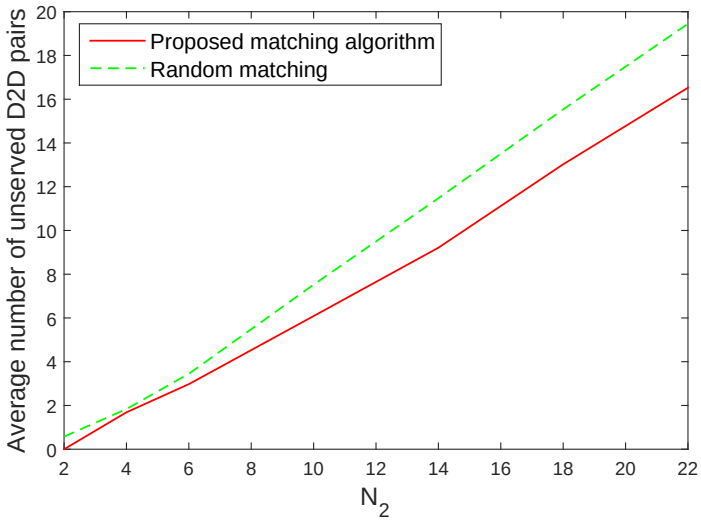

Figure 7.12: Average number of unserved D2D pairs vs. N_2

In Figs. 7.11 and 7.12, we plot the system throughput and the number of unserved D2D pairs as functions of the number of D2D pairs N_2. Since the results vary as we move the position of users for each N_2 value, we take the average over 100 different systems, which are generated randomly. In Fig. 7.11, we find that the system throughput of our algorithm is higher, especially for high N_2 values. As the number of D2D pairs increases, the random allocation algorithm has less chance to obtain the best result via random searching. Therefore, the throughput of random allocation algorithm reaches the limit when $N_2 > 8$. On the contrary, the throughput of our algorithm keeps increasing fast as we increase N_2 to very high values. This is because our algorithm can more effectively match the D2D pairs to the channels and efficiently identify their modes when the number of D2D pairs is large. In addition, Fig. 7.12 shows that our algorithm can serve more users compared to the random allocation algorithm.

7.3 A Joint Mode Selection and Resource Allocation Algorithm for D2D Communications via Vertex Coloring

7.3.1 System Model and Assumptions

In this section, as shown in Fig. 7.13, we consider a D2D underlaid cellular network, which has one base station (**BS**), N_c cellular users $\{\mathbf{CU}_1, \mathbf{CU}_2, \cdots, \mathbf{CU}_{N_c}\}$ and N_d D2D pairs $\{(\mathbf{DT}_1, \mathbf{DR}_1), (\mathbf{DT}_2, \mathbf{DR}_2), \cdots, (\mathbf{DT}_{N_d}, \mathbf{DR}_{N_d})\}$. We assume that the D2D transmission is one-way, in which $\mathbf{DT}_i$ and $\mathbf{DR}_i$ represent the transmitter and receiver of the i^{th} D2D pair, respectively. Each D2D pair can choose between the cellular mode and D2D mode. In D2D mode, D2D users transmit through D2D direct links, while in cellular mode, they transmit via D2D two-hop links through

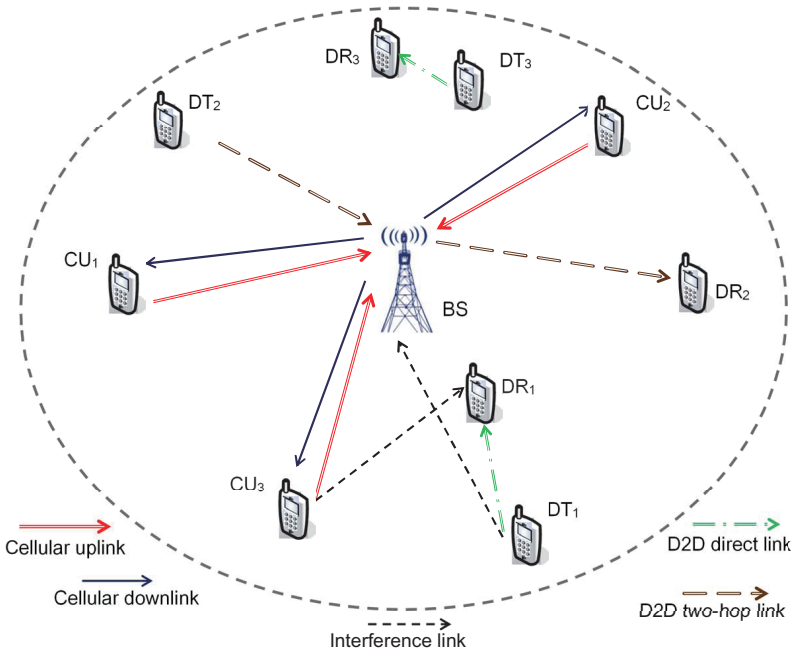

Figure 7.13: System model

the base station. Each cellular user transmits to the base station through an uplink channel, and receives data from the base station via a downlink channel. Hence, there are overall N_c uplinks, N_c downlinks and N_d D2D links. The maximum transmission power of a cellular user and D2D transmitter are set at P_c and P_d, respectively. When acting as a transmitter, the maximum transmission power of the base station is P_b in each channel. Therefore, the overall transmission power of the base station depends on the number of cellular users and the number of D2D pairs operating in the cellular mode.

There are N available orthogonal channels for this cellular network, each of them having a bandwidth of B. For simplicity, there are four assumptions regarding the channel allocation, which were also made in many related works such as [43] and [93]:

1. A D2D pair operating in the cellular mode cannot share its channel with other users.

2. Each cellular link, including both uplink and downlink, is allocated a single orthogonal channel, and channels cannot be shared by different cellular links.

3. It is necessary for a pair of direct links to satisfy the pair-wise interference constraints given below in (7.43) in Section 7.3.2.1 to reuse the same channel.

4. Each link, including D2D direct link, D2D two-hop link, cellular uplink and downlink, can operate in one channel at most.

5. The base station has the knowledge of the distributions of all channel fading coefficients, i.e., has statistical channel side information.

The first assumption helps to protect the performance of those D2D users that select the cellular mode. In general, D2D users that select the cellular mode usually have weak connections to their corresponding receivers, i.e., the distances between D2D transmitter, D2D receiver and the base station are relatively large. Therefore, assigning these D2D two-hop channels dedicated transmission resources provides a certain level of QoS guarantee. The second assumption guarantees the performance of cellular users, which have higher priorities than D2D users. The third assumption controls the interference among the users that reuse the same transmission resource. The last assumption implies that our resource allocation algorithm is performed at the base station, and our algorithm only requires the knowledge of the fading distributions, i.e., statistical channel side information. In general, fading distributions depend on the environment and distance between the transmitter and receiver. If a certain fading model is considered, such as Rayleigh, Rician or Nakagami-m fading, then the fading distributions are mainly determined by the location of the users.

According to these assumptions, a channel can be assigned to a single D2D link, a single cellular link, a group of D2D direct links or a group of D2D direct links together with a cellular link. For the last two cases, the users transmitting in the same channel cause interference to each other. Note that in some systems, cellular downlinks do

not share transmission resources with D2D users. In such cases, we just need to first assign transmission resources to those cellular downlinks before applying our algorithm. In this section, we assume that $2N_c \leq N \leq 2N_c + N_d$, which implies that we have sufficient number of channels to guarantee the performance requirements of all cellular users. However, having dedicated channels for all D2D users is not feasible in such a situation, and reusing/sharing of channel resources has to be considered in order to serve as many D2D users as possible.

The channels are assumed to experience ergodic fading, and the fading coefficients are denoted by h. Fading coefficients in different frequency bands are assumed to be i.i.d.. In the following analysis, the magnitude-squares of the fading coefficients are denoted by $z = |h|^2$. At each receiver, the background noise is assumed to follow an independent complex Gaussian distribution with zero mean and variance σ^2, i.e., $n \sim \mathcal{CN}(0, \sigma^2)$. Therefore, the SNR of each transmitter can be defined as $\text{SNR} = \frac{P}{B\sigma^2}$, where P represents the transmission power.

In this section, we consider mode selection, power optimization and channel allocation jointly to maximize the throughput as well as the number of users served in the network. In the next subsection, we introduce our algorithm step by step.

7.3.2 Joint Mode Selection and Resource Allocation Algorithm

In this subsection, we introduce our three-step joint mode selection and resource allocation algorithm in detail. In the first step, we divide the transmission links into groups via the vertex coloring method. In the second step, we conduct power optimization for each group, and perform mode selection between D2D mode and cellular mode for those D2D links which form groups. In the last step, we assign channels to those groups.

Before applying the algorithm, we enumerate cellular uplinks from 1 to N_c, cellular

downlinks from $N_c + 1$ to $2N_c$, and D2D direct links from $2N_c + 1$ to $2N_c + N_d$. D2D two-hop links are only considered in the mode selection part in the second step. With the given link indices, we can denote the magnitude-square of the fading coefficient between the transmitter of link i and the receiver of link j by $z_{i,j}$, and we can represent the *expected* values of z collectively in a channel fading matrix $\mathbf{Z}$.

Two main objectives of our algorithm are to maximize the sum rate and to maximize the number of users served in the network. Most of the time, these two goals cannot be achieved simultaneously because of the presence of interference. In the following discussion, we illustrate how to balance these two goals via parameter selection.

7.3.2.1 Partition via Vertex Coloring Method

The first step of our algorithm is transmission link partition. The partition algorithm divides transmission links into small groups, greatly reducing the dimensionality of the power optimization problem in the second step. According to our channel assignment assumptions, multiple cellular links cannot be in the same group, and any two links in the same group have to satisfy the pair-wise interference constraints given by

$$\begin{cases} P_{imax}\bar{z}_{ii}/(P_{jmax}\bar{z}_{ji}) \geq \gamma \\ P_{jmax}\bar{z}_{jj}/(P_{imax}\bar{z}_{ij}) \geq \gamma \end{cases} \tag{7.43}$$

where P_{imax} and P_{jmax} are the maximum transmission powers over links i and j respectively, $\bar{z}$ represents the expected value of z, and γ is the interference threshold. These pair-wise interference constraints provide QoS guarantees for both cellular and D2D users from the perspective of interference control.

The key steps of our partition algorithm are to construct a graph while regarding these $2N_c + N_d$ transmission links as vertices, and to perform the partition using the minimum vertex coloring algorithms from graph theory. Note that these algorithms

Table 7.4: Algorithm 7.4

Partition Algorithm

Input: interference threshold γ, channel fading matrix $\mathbf{Z}$.

Output: partition $\mathbf{\Pi} = \pi_1, \pi_2, \cdots, \pi_{n_g}$.

For $i = 1 : 2N_c + Nd$
  $\gamma_i = \gamma$;
End
Generate a random permutation of integers from 1 to $2N_c + N_d$, and denote it by $\mathbb{A}_1$;
For each $i \in \mathbb{A}_1$
  Generate a random permutation of integers from $i + 1$ to $2N_c + N_d$, and denote it by $\mathbb{A}_2$;
  For each $j \in \mathbb{A}_2$
    If both links i and j are smaller than $2N_c$
      Create an edge between vertices i and j;
    Elseif links i and j cannot satisfy

$$\begin{cases} P_{imax}\bar{z}_{ii}/(P_{jmax}\bar{z}_{ji}) \geq \gamma_i \\ P_{jmax}\bar{z}_{jj}/(P_{imax}\bar{z}_{ij}) \geq \gamma_j \end{cases}$$

      Create an edge between vertices i and j;
    Else
      Increase both γ_i and γ_j by $\Delta\gamma$;
    End
  End
End
Apply the Welsh-Powell algorithm to get the partition $\mathbf{\Pi}$;

divide all vertices into minimum number of groups such that any two vertices in the same group are not connected. Therefore, we construct the graph by checking each pair of vertices, and connect them if they cannot be in the same group. A detailed description of our partition algorithm is given in Table 7.4. The output of this algorithm is a partition with size n_g, and each element of the partition is a set of vertices that form a group. In order to further control the interference and number of users in a group, we gradually increase the γ values of each link. As we can see in the algorithm, all threshold values are set at γ initially. Each time we find a pair of links that can be in the same group, we increase the thresholds of these two links by

$\Delta\gamma$. This mechanism can effectively limit the received interference at each receiver, and balance the size of each group. Also due to this mechanism, two links may have a higher chance to be in the same group if we check them earlier. In order to let the vertices to have equal chances to connect with each other, we use random orders to choose link pairs in the double for-loop. In the last step of the algorithm, we use the Welsh-Powell algorithm [94] to solve the vertex coloring problem. Welsh-Powell algorithm is a very fast algorithm that can provide good results effectively.

In this step, γ and $\Delta\gamma$ are the parameters to control the tradeoff between sum rate and number of users served by the system. For large values of γ and $\Delta\gamma$, the interference is well controlled, but the system serves potentially small number of users. On the other hand, for small γ and $\Delta\gamma$ values, more users can reuse the same channel resource, but the interference may lower the sum rate.

In practice, the partition algorithm can potentially provide us a partition with size smaller than the number of channels, which means that some of the channels will not be utilized, because each user group π_i is assigned a channel in the third step of our algorithm. In order to avoid this situation, we need to further improve our partition algorithm using a γ-adjusting algorithm described in Table 7.5. In Algorithm 7.5, we find a threshold $\hat{\gamma}$ that makes the partition size $n_g = N$ through bisection search. Notice that the threshold value that can achieve $n_g = N$ is not unique, and the time consumption of this adjusting algorithm is very small.

After obtaining the partition, we conduct power optimization and mode selection in the second step.

7.3.2.2 Power Optimization and Mode Selection

In the second step, we do power optimization for each group. If a group just contains a single D2D direct link, then we perform mode selection for this D2D pair.

Table 7.5: Algorithm 7.5

γ Adjusting Algorithm
Input: interference threshold γ, channel fading matrix $\mathbf{Z}$.
Output: partition $\mathbf{\Pi} = \pi_1, \pi_2, \cdots, \pi_{n_g}$.
Run Algorithm 7.4 with threshold γ; **If** $n_g \geq N$ End process; **End** Set $\hat{\gamma} = \gamma$; **While** $n_g < N$ $\hat{\gamma} = 2\hat{\gamma}$; Run Algorithm 7.4 with threshold $\hat{\gamma}$; **End** Set the upper bound $\gamma_u = \hat{\gamma}$, lower bound $\gamma_l = \hat{\gamma}/2$, and $\hat{\gamma} = (\gamma_u + \gamma_l)/2$; **While** $n_g \neq N$ Run Algorithm 7.4 with threshold $\hat{\gamma}$; **If** $n_g > N$ $\gamma_u = \hat{\gamma}$; **Elseif** $n_g < N$ $\gamma_l = \hat{\gamma}$; **End** $\hat{\gamma} = (\gamma_u + \gamma_l)/2$; **End**

Power Optimization

If a group only contains one direct link, then the transmitter transmits with its maximum power. For the groups containing multiple transmission links, a general expression of the objective function for the power optimization problem in group π_i is

$$Obj(\mathbf{P_i}) = \sum_k \omega_k CRT_k(\mathbf{P_i}), \tag{7.44}$$

where $\mathbf{P_i}$ represents the power vector which consists of the transmission powers of the transmitters in group π_i, the function CRT can be defined based on the criteria selected in the optimization problem, such as the maximization of the sum rate, energy efficiency, or minimum rate, and ω_k is the corresponding weight of CRT_k which indicates the significance of CRT_k. The formulation given in (7.44) can provide QoS and fairness guarantees. For instance, by choosing the energy efficiency as a criterion, a certain energy efficiency performance can be achieved; or by choosing the minimum rate as a criterion, the minimum rate performance of each users can be guaranteed.

In this section, we consider both the sum rate and minimum rate as the criteria, and formulate our power optimization problem for group π_i as

$$\textbf{Maximize}_{\mathbf{P_i}} \quad \sum_{j\in\pi_i} \mathbb{E}\{R_j(\mathbf{P_i})\} + \mu \min_{j\in\pi_i} \left\{\mathbb{E}\{R_j(\mathbf{P_i})\}\right\} \tag{7.45}$$

$$\textbf{Subject to} \qquad 0 \le P_j \le P_{jmax}, \text{ for } j \in \pi_i \tag{7.46}$$

where P_{jmax} represents the maximum transmission power in link j. The evaluation of the average transmission rate $\mathbb{E}\{R_j(\mathbf{P_i})\}$ is discussed in Remark 3 below. In this problem, μ is the weight parameter for the minimum rate. For small μ values, the objective function is mainly determined by the sum rate component, which may sacrifice the rates of some users. On the other hand, for large μ values, the objective function is mainly dominated by the minimum rate component, which may limit the

sum rate. This optimization problem can be transformed into

$$\textbf{Maximize}_{\mathbf{P_i},r} \quad \sum_{j\in\pi_i} \mathbb{E}\{R_j(\mathbf{P_i})\} + \mu r \tag{7.47}$$

$$\textbf{Subject to} \qquad 0 \le P_j \le P_{jmax}, \text{ for } j \in \pi_i \tag{7.48}$$

$$\mathbb{E}\{R_j(\mathbf{P_i})\} \ge r, \text{ for } j \in \pi_i \tag{7.49}$$

for which suboptimal solutions can be obtained via the interior-point method [95]. In order to improve the performance, we need to repeat the algorithm several times with randomly selected initial points.

Remark 3 *In order to determine the average rate of a user accurately and efficiently,we perform numerical integration. To evaluate this high-dimensional integral, we transform it into two single integrals for certain specific fading models:*

$$\begin{aligned}\mathbb{E}\{R_j(\mathbf{P_i})\} =& \mathbb{E}\left\{B\log_2\left(1+\frac{SNR_j z_{jj}}{1+\sum_{k\in\pi_i,k\neq j} SNR_k z_{kj}}\right)\right\} &(7.50)\\ =& \mathbb{E}\left\{B\log_2\left(1+\sum_{k\in\pi_i} SNR_k z_{kj}\right)\right\} \\ & - \mathbb{E}\left\{B\log_2\left(1+\sum_{k\in\pi_i,k\neq j} SNR_k z_{kj}\right)\right\}. &(7.51)\end{aligned}$$

In Rayleigh fading, $SNR\,z$ follows an exponential distribution with probability density function (pdf)

$$f(x) = \frac{1}{SNR\,\bar{z}} e^{-x/(SNR\,\bar{z})}. \tag{7.52}$$

According to the results in [96], the summation of independent exponentially distributed random variables $S_M = \sum_{k=1}^{M} X_k$, where $X_k \sim \exp(\lambda_k)$, has a pdf given by

$$f_{S_M}(s) = \sum_{i=1}^{M} \frac{\prod_{j=1}^{M}\lambda_j}{\prod_{j=1,j\neq i}^{M}(\lambda_j-\lambda_i)} e^{-s\lambda_i}. \tag{7.53}$$

Using this characterization, the sum terms $\sum_{k\in\pi_i} SNR_k z_{kj}$ *and* $\sum_{k\in\pi_i, k\neq j} SNR_k z_{kj}$ *in (7.51) can be regarded as two random variables, and the average rate can be evaluated using two single integrals. Similar approach can be applied to some other fading models as well.*

Mode Selection

If a group just contains a single D2D direct link, then this D2D pair can choose between D2D mode and cellular mode. In cellular mode, D2D users communicate through the base station, and each time block is divided into two phases. In the first phase, the D2D transmitter sends packets to the base station, and base station forwards the packets to the corresponding D2D receiver in the second phase. We assume that the base station decodes and stores the received packets from D2D transmitters in a buffer, and the buffer empty probability is negligible. Let τ_i denote the fraction of time allocated to link $\mathbf{DT}_i - \mathbf{BS}$. If users $\mathbf{DT}_i$ and $\mathbf{DR}_i$ are in cellular mode, then the fraction of time allocated to $\mathbf{BS} - \mathbf{DR}_i$ link is $1 - \tau_i$. Since the throughput of the two-hop link $\mathbf{DT}_i - \mathbf{BS} - \mathbf{DR}_i$ is $\min\{\tau_i \mathbb{E}\{R_{\mathbf{DT}_i-\mathbf{BS}}\}, (1-\tau_i)\mathbb{E}\{R_{\mathbf{BS}-\mathbf{DR}_i}\}\}$, the optimal τ_i value is given by

$$\tau_i^* = \frac{\mathbb{E}\{R_{\mathbf{BS}-\mathbf{DR}_i}\}}{\mathbb{E}\{R_{\mathbf{DT}_i-\mathbf{BS}}\} + \mathbb{E}\{R_{\mathbf{BS}-\mathbf{DR}_i}\}}, \tag{7.54}$$

which leads to $\tau_i \mathbb{E}\{R_{\mathbf{DT}_i-\mathbf{BS}}\} = (1-\tau_i)\mathbb{E}\{R_{\mathbf{BS}-\mathbf{DR}_i}\}$. Above in (7.54), the instantaneous rates of links $\mathbf{DT}_i - \mathbf{BS}$ and $\mathbf{BS} - \mathbf{DR}_i$ are formulated as

$$R_{\mathbf{DT}_i-\mathbf{BS}} = B\log_2\left(1 + \frac{P_d}{B\sigma^2} z_{\mathbf{DT}_i-\mathbf{BS}}\right) \tag{7.55}$$

$$R_{\mathbf{BS}-\mathbf{DR}_i} = B\log_2\left(1 + \frac{P_b}{B\sigma^2} z_{\mathbf{BS}-\mathbf{DR}_i}\right) \tag{7.56}$$

where the subscript of the fading power z denotes the link to which the fading power

is associated. Then, the average transmission rate of the i^{th} D2D pair in cellular mode is

$$\mathbb{E}\{R_{\mathbf{DT}_i-\mathbf{BS}-\mathbf{DR}_i}\} = \tau_i^* \mathbb{E}\{R_{\mathbf{DT}_i-\mathbf{BS}}\} \tag{7.57}$$

$$= \frac{\mathbb{E}\{R_{\mathbf{BS}-\mathbf{DR}_i}\}\mathbb{E}\{R_{\mathbf{DT}_i-\mathbf{BS}}\}}{\mathbb{E}\{R_{\mathbf{DT}_i-\mathbf{BS}}\} + \mathbb{E}\{R_{\mathbf{BS}-\mathbf{DR}_i}\}}. \tag{7.58}$$

In D2D mode, the average transmission rate of link $\mathbf{DT}_i - \mathbf{DR}_i$ is

$$\mathbb{E}\{R_{\mathbf{DT}_i-\mathbf{DR}_i}\} = \mathbb{E}\left\{B\log_2\left(1+\frac{P_d}{B\sigma^2}z_{jj}\right)\right\} \tag{7.59}$$

where $j = 2N_c + i$ is the index of the i^{th} D2D direct link. We compare the average rates in these two modes, and select the one with the higher average rate.

7.3.2.3 Channel Assignment

In the first step, we divide the transmission links into n_g groups, and the optimal transmission power and transmission mode of each user are obtained in the second step. In this third step discussed in this subsection, we allocate channel resources to each group.

We first allocate a channel to each group containing a cellular link, to guarantee that each cellular link is provided a channel. Following this step, there are $N - 2N_c$ channels left for the remaining D2D users. Given these channels, we can choose to maximize the sum rate or maximize the total number of users served by the system.

If we choose to maximize the sum rate, then we need to pick $N - 2N_c$ groups with the highest group sum rates from the remaining $n_g - 2N_c$ groups, and assign each of them a channel. If we choose to maximize the number of users served by the system, then we need to select $N - 2N_c$ groups with the largest group sizes, and assign each of them a channel.

Table 7.6: Algorithm 7.6

Joint Mode Selection and Resource Allocation Algorithm
Run Algorithm 7.5 for a given γ value to obtain a partition with size N_g greater or equal to the number of channels N; **For** $i = 1 : n_g$ Run the power optimization algorithm for the i^{th} group; **If** the i^{th} group only contains one D2D link Run the mode selection algorithm for this D2D link; **End** **End** Run the channel assignment algorithm to assign channel resources to these groups.

7.3.2.4 Summary

Our joint mode selection and resource allocation algorithm is described in Table 7.6. Via the vertex coloring algorithm, we can quickly divide users into small groups, which greatly lowers the dimensionality of the power optimization problem in the second step and reduces the time consumption. From numerical results, we notice that the majority of the time is spent on solving the power optimization problems in the second step. Therefore, finding a faster algorithm instead of the interior-point method for the power optimization problem is the key to further reduce the time consumption of our algorithm, and we leave a detailed study of this problem as our future work.

7.3.3 Numerical Results

In this subsection, we further investigate the performance and parameter selection of our joint mode selection and resource allocation algorithm via simulations. For our algorithm, we set the initial threshold of the interference constraints as $\gamma = 250$, and $\Delta\gamma \in \{50, 125, 250, 1250, 2500\}$. In the power allocation step, we set the weight for the minimum rate objective as $\mu = 0.2 \times \text{Size}(\pi_i)$, where size$(\pi_i)$ represents the number of links in the i^{th} group. In the channel assignment step, we choose to maximize the

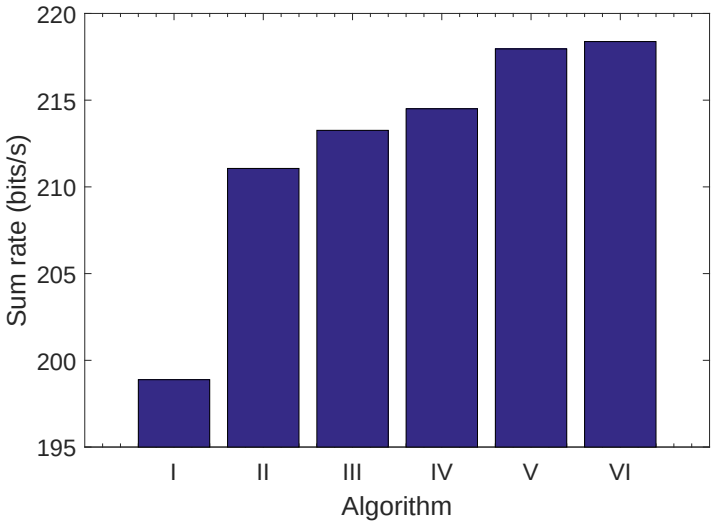

Figure 7.14: Comparison of sum rate

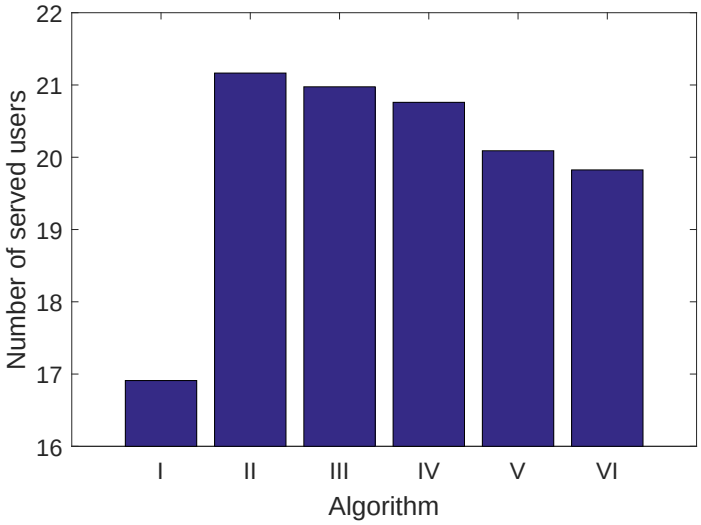

Figure 7.15: Comparison of the number of served users

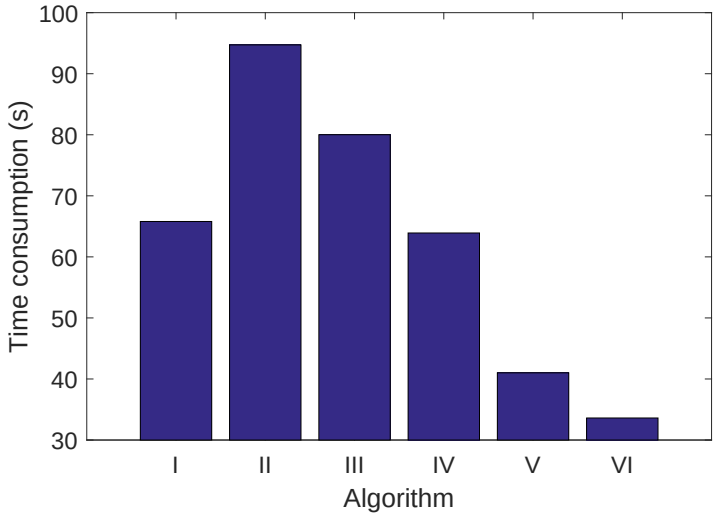

Figure 7.16: Comparison of time consumption

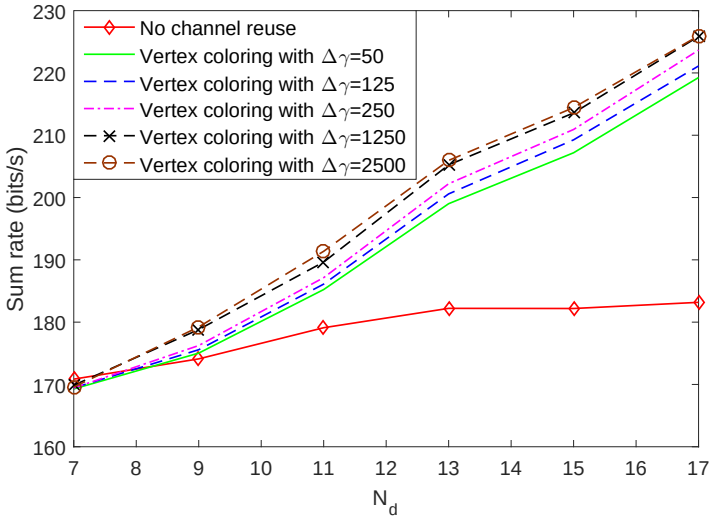

Figure 7.17: Sum rate vs. N_d

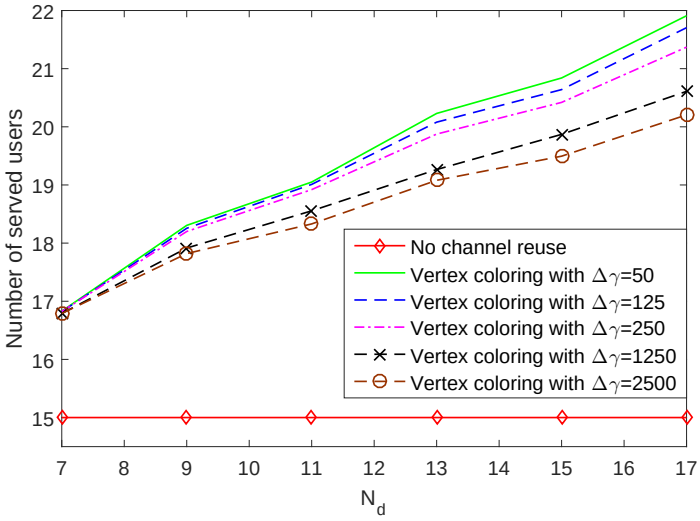

Figure 7.18: Number of users served by the system vs. N_d

sum rate. The number of channels is $N = 25$, and the number of cellular users is $N_c = 10$. The maximum SNRs are $\frac{P_b}{B\sigma^2} = 27.78$ dB, $\frac{P_c}{B\sigma^2} = \frac{P_d}{B\sigma^2} = 26.99$ dB. We consider Rayleigh fading with path loss $\mathbb{E}\{z\} = d^{-4}$, where d is the transmission distance, and users are randomly placed in the cell. We repeat each simulation 200 times, and each point in the numerical plots is averaged over 200 randomly generated systems.

In Figs. 7.14-7.16, we compare the performance of our algorithm with the coalitional game method proposed in [48]. In the coalitional game, each user forms a coalition at the beginning, and link i prefers to join coalition j if the sum objective function increases by moving link i to coalition j. In Figs. 7.14-7.16, algorithms I, II, III, IV, V and VI represent the coalitional game algorithm and the vertex coloring algorithms with $\Delta\gamma = 50$, $\Delta\gamma = 125$, $\Delta\gamma = 250$, $\Delta\gamma = 1250$ and $\Delta\gamma = 2500$, respectively, and the number of D2D pairs is fixed as $N_d = 15$. From the results, we can see that our vertex coloring algorithm provides higher sum rates and serves more users than the coalitional game algorithm. As $\Delta\gamma$ increases, the sum rate increases due to less interference, but the number of users being served decreases due to stricter interference constraints, as noted in Section 7.3.2.1. Also, we notice that as $\Delta\gamma$ increases,

the time consumption of our algorithm reduces, and when $\Delta\gamma = 2500$, our algorithm is much faster than the coalitional game algorithm[2]. For larger values of $\Delta\gamma$, the maximum group size is small, reducing the dimensionality and time consumption of the power allocation problems in the second step.

In Figs. 7.17 and 7.18, we plot the sum rate and number of served users as functions of the number of D2D pairs N_d. In these two figures, we consider the results without channel reuse as the benchmark, in which all users transmit with their maximum power and only 25 links (10 cellular uplinks, 10 cellular downlinks and 5 D2D links) with the highest rates are allocated dedicated channels. We can see that as N_d increases, the advantages of allowing channel reuse become more obvious. The sum rate and number of served users increase much faster when channel reuse is allowed. Similarly, larger $\Delta\gamma$ improves the sum rate while sacrificing the number of served users.

In summary, our algorithm has high performance and low time consumption. When the sum rate is more important, we can choose relatively high values for γ and $\Delta\gamma$, and assign channels to the groups with higher group rates. In this case, the time consumption can also be reduced. However, if the values of γ and $\Delta\gamma$ are too large, then our algorithm leads to the cases in which channel reuse is not allowed. On the other hand, we can choose relatively low values for γ and $\Delta\gamma$, and assign channels to the groups with more users if we choose to maximize the number of served users. However, if the values of γ and $\Delta\gamma$ are too small, then the interference limits the transmission rate of each user, and the service quality may degrade. Therefore, avoiding such extreme values and optimizing parameter selection are preferred.

[2]These time consumption measurements are obtained for codes in Matlab 2015b running on a 2.40GHz Intel i7-4700MQ CPU.

Chapter 8

Resource Allocation for Content Delivery over Wireless Cellular Networks

In this chapter, we focus on the delay performance of content delivery over wireless cellular networks. Three types of network models are considered, including D2D caching network, D2D cellular network, and C-RAN.

By storing parts of the popular files at the mobile users, users can locate some of their requested files in their own caches or the caches at their neighbors. In the latter case, when a user receives files from its neighbors, D2D communication is enabled. D2D communication underlaid with cellular networks is also a new paradigm for the upcoming 5G wireless systems. In Section 8.1, we propose a very efficient caching algorithm for D2D-enabled cellular networks to minimize the average transmission delay. Instead of searching over all possible solutions, our algorithm finds out the best <file,user> pairs, which provide the best delay improvement in each loop to form a caching policy with very low transmission delay and high throughput. This algorithm is also extended to address a more general scenario, in which the distributions of fading

coefficients and values of system parameters potentially change over time.

In Section 8.2, we develop a scheduling algorithm for D2D cellular networks with deadline constraints via the convex delay cost approach. At the beginning of each time slot, the algorithm allocates all available channels to the users, and each user can choose to transmit in different modes. After characterizing the transmission rates and defining the utility for each possible scheduling decision, we propose power optimization algorithms to maximize the utility for each type of decision. Our scheduling algorithm allocates each channel according to the decision that provides the maximum utility value, and it manages mode selection, channel allocation and power optimization. Via simulation results, we discuss the parameter selection for our algorithm and verify the performance improvements by allowing D2D users to share channels with other users.

In Section 8.3, we formulate the utility of each user using a convex delay cost function, and design a two-step scheduling algorithm with good delay performance for the C-RAN architecture. C-RAN architecture is a new mobile network architecture that enables cooperative baseband processing and information sharing among multiple cells and achieves high adaptability to nonuniform traffic by centralizing the baseband processing resources in a virtualized BBU pool. In the first step, all users in multiple cells are grouped into small user groups, according to their interference levels and estimated utilities. In the second step, channels are matched to the user groups to maximize the system utility. The performance of our algorithm is further studied via simulations, and the advantages of C-RAN architecture is verified.

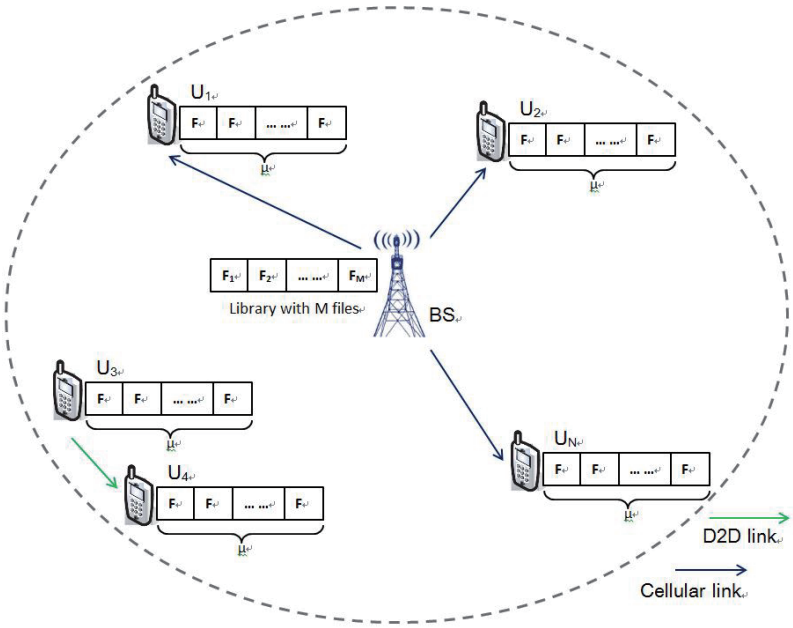

Figure 8.1: System model of a D2D cellular network with caches

8.1 A Delay-Aware Caching Algorithm for Wireless D2D Caching Networks

8.1.1 System Model and Problem Formulation

8.1.1.1 System Model and Channel Allocation

As shown in Fig. 8.1, we consider a cellular network with one base station (**BS**), in which a library with M files ($\mathcal{F}_1$, $\mathcal{F}_2$, $\cdots$, $\mathcal{F}_M$) is stored, and we assume that the size of each file is fixed to F bits [1]. There are N users (U_1, U_2, $\cdots$, U_N) in the network who seek to get the content files from the library. Each user is equipped with a cache of size μF bits, and therefore can store μ content files. The caching state is described by an $N \times M$ matrix $\mathbf{\Phi}$, whose (i, j)-th component has a value of $\phi_{i,j} = 1$ if file $\mathcal{F}_j$ is

[1] In the literature, it is noted that the base station may only store a portion of the library contents, and needs to acquire the remaining files from the content server [58]. Since we focus on the wireless transmission delay, we do not explicitly address the link between the base station and content server. Also, the content files may not have the same size in practice, but we can further divide them into sub-files with equal size.

cached at user U_i, and $\phi_{i,j} = 0$ when the user U_i does not have file $\mathcal{F}_j$ in its cache.

In general, users request files with different probabilities, which are characterized by an $N \times M$ popularity matrix $\mathbf{P}$, in which the entry on the i^{th} row and j^{th} column, $P_{i,j}$, represents the probability of user U_i requesting file $\mathcal{F}_j$. Each row of the popularity matrix corresponds to a popularity vector of a user. Although the popularity matrix may change over time in practice, we can assume that the popularity stays constant within a certain period, and our caching algorithm needs to be repeated when the popularity matrix is updated. In the literature, Zipf distribution is generally considered as a good statistical model for the popularity. The pmf of this distribution is given by

$$P_{i,j} = \frac{f_{i,j}^{-\beta}}{\sum_{k=1}^{M} k^{-\beta}}, \tag{8.1}$$

where $f_{i,j}$ is the popularity index that user U_i gives to file $\mathcal{F}_j$, and $\beta \geq 0$ is the Zipf exponent. Each user enumerates the files with popularity index from 1 to M, where the most popular file gets index 1, and the least popular file gets index M. As the Zipf exponent β increases, the difference in the popularity of different files increases, while all files have the same popularity when $\beta \to 0$. Although we use Zipf distribution for our numerical results in Section 8.1.4, our proposed algorithm works for any type of popularity model. At each user, the generated requests are buffered in a queue before getting served, and it is assumed that these request queues are not empty at any time.

In a D2D-enabled wireless network, users can choose to transmit in cellular mode or D2D mode. In the cellular mode, users request and receive information from the base station, while in the D2D mode, a user requests and receives information from another user through a D2D direct link. In our model, the users first check their local cache when a file is requested. If the user does not have the corresponding file

in its own cache, it sends a request to the base station. We assume that the base station has knowledge of all fading *distributions* (i.e., only has statistical information regarding the channels) and the cached files at each user. After receiving the request, the base station identifies the source node from which the file request can be served and allocates channel resources to the corresponding user. Therefore, the result of mode selection is determined by the result of source selection. If the source node is another user, then the requested file is sent over the direct D2D link and hence the communication is in D2D mode, otherwise the receiving user works in cellular mode and receives files from the base station. In source selection, among all the nodes (including the base station) who have the requested file, the node with the lowest average transmission delay to the receiver is selected as the transmitter.

In this section, we consider an OFDMA system with N_c orthogonal channels, and the bandwidth of each channel is B. We assume that the background noise samples follow i.i.d. circularly-symmetric complex Gaussian distribution with zero mean and variance σ^2 at all receivers in all frequency bands, and the fading coefficients of the same transmission link are i.i.d. in different frequency bands. The fading coefficients are assumed to stay constant within one time block of duration T_0, and change across different time blocks. We summarize the resource allocation assumptions for the discussions in Sections 8.1.1 and 8.1.2 as follows:

1. Each channel can be used for the transmission of one requested file at most, and the transmission of a file cannot occupy multiple channels.

2. D2D transmitters are not allowed to transmit to multiple receivers simultaneously. In other words, the file requests whose best source node is a user who is already transmitting cannot be assigned a channel resource by the base station.

3. The probability of a channel being allocated to a request generated by user i is $\hat{p}_i$.

4. After a request is served, the corresponding transmitter keeps silent in the remaining time block, and the base station allocates the channel resource to other requests at the beginning of the following time block.

5. If the i^{th} user is selected as a D2D transmitter, its maximum transmission power is P_i.

6. The base station can serve multiple requests simultaneously using different channels, and its maximum transmission power is P_b for each request.

The first four assumptions describe a class of simple scheduling algorithms, in which only point to point transmission without spectrum reusing is considered. At the beginning of each time block, base station assigns available channels to the requests, and each transmission link gets one channel at most, and uses the assigned channel exclusively. The transmitter transmits until the request is served and then releases the channel resource. The behavior of the scheduling algorithm is described by a set of probabilities $\hat{p}_i$ defined in the third assumption. Although our delay characterizations in Sections 8.1.1.2 and 8.1.1.3 are only valid for this type of scheduling algorithms, we further extend our results for more complicated scheduling algorithms in Section 8.1.3. In that case, only the last two assumptions are required, which describe the maximum power constraints. With a more complicated scheduling algorithm, we can only estimate the average delay of each request at each user through simulation or learning methods. A detailed discussion is provided in Section 8.1.3.

In this section, only the distributions of the fading coefficients are required at the base station, which mainly depend on the environment and the location of each user. A centralized computation scheme is used, and the base station sends the results of caching and scheduling algorithms to the users through additional control channels. Since the base station knows the distributions of all fading coefficients and the cached files at each user, the average delay between each user pair and the best source node

for each request can be obtained at the beginning and stored in tables at the base station. A detailed discussion on delay calculation and the determination of the best source node is provided in the next subsection.

8.1.1.2 Transmission Delay

In this section, we use the transmission delay, which is defined as the number of time blocks used to transmit a content file, as the performance metric. From the above discussion, the instantaneous channel capacity a transmission link in the k^{th} time block is

$$C[k] = B \log_2 \left(1 + \frac{P_t}{B\sigma^2} z_k\right) \quad \text{bits/s} \tag{8.2}$$

where P_t is the transmission power, and z_i is the magnitude square of the corresponding fading coefficient in the k^{th} time block. In order to maximize the transmission rate, all transmitters transmit at the maximum power level. Therefore,

$$P_t = \begin{cases} P_b & \text{if the transmitter is the base station} \\ P_i & \text{if the transmitter is the } i^{\text{th}} \text{ user} \end{cases}, \tag{8.3}$$

and the duration to send a file is

$$T = \min \left\{ t : F \leq \sum_{k=1}^{t} T_0 C[k] \right\} \tag{8.4}$$

where F is the size of each file, T_0 is the duration of each block, and $C[k]$ is the instantaneous channel capacity in the k^{th} time block. When the fading distribution is available, the average transmission delay of the link $U_i - U_j$, which is denoted by $\mathbb{E}\{T_{i,j}\}$, can be obtained through numerical methods or Monte-Carlo simulations. These average delay values can be stored in an $N \times N$ symmetric matrix $\mathbf{T}_{\text{avg}}$, whose

component on the i^{th} row and j^{th} column is given by $\mathbb{E}\{T_{i,j}\}$ when $i \neq j$, and the diagonal element $T_{i,i}$ is the average delay between U_i and the base station. According to our channel assumptions, the average delays of a transmission link are the same in every channel. Therefore, we only need to analyze the performance in a single channel.

The best source node of the request, which is generated by user U_i requesting file $\mathcal{F}_j$, is the node which has file $\mathcal{F}_j$ and the smallest average transmission delay to U_i, and this minimum average delay is denoted by $D_{i,j}$ [2]. The best source of each possible request can be stored in an $N \times M$ table $\mathbf{S}$, in which each row corresponds to a user who generates the request, and each column corresponds to a file being requested. Also, these $D_{i,j}$ values can be collected in an $N \times M$ matrix $\mathbf{D}$.

Using the above results, the average transmission delay of the requests generated by user U_i can be obtained as

$$D_i = \sum_{j=1}^{M} P_{i,j} D_{i,j}, \tag{8.5}$$

where $P_{i,j}$ is the (i,j)-th component of the popularity matrix $\mathbf{P}$.

8.1.1.3 Problem Formulation

In the previous subsection, we have determined and expressed the average delay. In this subsection, we formulate and discuss our caching problem. In this section, our goal is to minimize the weighted sum of the average delays of the users, which is expressed as

$$\eta = \sum_{i=1}^{N} \omega_i D_i = \sum_{i=1}^{N} \omega_i \sum_{j=1}^{M} P_{i,j} D_{i,j} \tag{8.6}$$

[2] If U_i has cached $\mathcal{F}_j$, then the best source node is U_i itself, and $D_{i,j} = 0$.

where $\omega_i \in [0,1]$ is the weight for user U_i. We assume that the values of the weights are predetermined. In practice, ω values can be determined according to the priorities of users, so that users with higher priority have higher weights.

Our caching problem is formulated as

$$\textbf{P1:} \quad \text{Minimize}_{\ \mathbf{\Phi}} \quad \eta \tag{8.7}$$

$$\text{Subject to} \quad \sum_{j=1}^{M} \phi_{i,j} = \mu \tag{8.8}$$

$$\phi_{i,j} \in \{0,1\} \tag{8.9}$$

where $\mathbf{\Phi}$ is the caching result indicator matrix. The constraint in (8.8) arises due to the maximum cache size. It is obvious that the optimal caching policy must use all caching space.

In a special case, if we choose $\omega_i = \hat{p}_i$, where $\hat{p}_i$ is the probability that a channel is allocated to user U_i, then η expresses the average delay of the system. In this situation, the throughput of the system can be expressed as

$$R = N_c \frac{F}{\eta}. \tag{8.10}$$

Therefore, in this special case, minimizing η is equivalent to maximizing the throughput of the system.

8.1.2 Caching Algorithm

In this subsection, we propose our caching algorithm that solves problem **P1**. Note that the objective in problem **P1** is not convex, and the solution space is a discrete set with size $(\frac{M!}{(M-\mu)!\,\mu!})^N$. Therefore, the globally optimal solution can only be obtained via exhaustive search. In this section, we propose an efficient algorithm to determine a caching policy with delay performance close to the optimal solution. At the end of

Table 8.1: Algorithm 8.1

Find the delay improvement for a <file,user> pair
Input : user index i, file index j, caching indicator $\phi_{i,j}$, weight vector $\boldsymbol{\omega} = (\omega_1, \cdots, \omega_N)$, popularity matrix $\mathbf{P}$, source table $\mathbf{S}$, delay matrices $\mathbf{T}_{\text{avg}}$ and $\mathbf{D}$. **Output :** delay improvement $g_{i,j}$, updated source table $\hat{\mathbf{S}}$, updated optimal delay matrix $\hat{\mathbf{D}}$.
Initialization : $\hat{\mathbf{S}} = \mathbf{S}$ and $\hat{\mathbf{D}} = \mathbf{D}$ **If** $\phi_{i,j} = 1$ $g_{i,j} = 0$, end process. **Else** $g_{i,j} = \omega_i P_{i,j} D_{i,j}$ and update $\hat{S}_{i,j} \leftarrow U_i$, $\hat{D}_{i,j} = 0$. **End** **For** $k = 1 : N$ **If** $D_{k,j} > T_{i,k}$ and $i \neq k$ $g_{i,j} = g_{i,j} + \omega_k P_{k,j}(D_{k,j} - T_{i,k})$ update $\hat{D}_{k,j} = T_{i,k}$ and $\hat{S}_{k,j} \leftarrow U_i$ **End** **End**

this section, we show that our algorithm has the potential to be extended to more complicated scenarios.

8.1.2.1 Caching Algorithm

Our algorithm is a greedy algorithm, which searches over a subset of the solution space with smaller size. At the beginning, we assume that all caches are empty, and every user has to operate in cellular mode, in which they only receive files from the base station. Then, in each step, we find the best <file,user> pair, which provides the maximum delay improvement (or equivalently reduction in delay) if the selected file is stored in the cache of the corresponding user. This process needs to be repeated $N\mu$ times, in order to fill all cache space, and the final caching policy is obtained.

In Table 8.1, we describe Algorithm 8.1 in detail, which calculates the delay improvement and determines the updated $\mathbf{S}$ and $\mathbf{D}$ matrices accordingly when we cache file $\mathcal{F}_j$ at user U_i. First, we check if $\mathcal{F}_j$ has already been cached at U_i. If so, we end

Table 8.2: Algorithm 8.2

Find the optimal <file,user> pair to be added in the updated caching result, leading to maximum delay improvement
Input : weight vector $\boldsymbol{\omega} = (\omega_1, \cdots, \omega_N)$, popularity matrix $\mathbf{P}$, caching indicator matrix $\mathbf{\Phi}$, source table $\mathbf{S}$, delay matrices $\mathbf{T}_{\text{avg}}$ and $\mathbf{D}$. **Output :** new source table $\mathbf{S}$, new optimal delay matrix $\mathbf{D}$, and new caching indicator matrix $\mathbf{\Phi}$.
Initialization : set optimal delay improvement $g^* = 0$, and set the corresponding $\mathbf{S}^* = \mathbf{S}$, $\mathbf{D}^* = \mathbf{D}$.

For $i = 1 : N$
  If $\sum_{j=1}^{M} \phi_{i,j} < \mu$
  For $j = 1 : M$
  run Algorithm 8.1 for $< U_i, \mathcal{F}_j >$, to obtain
  the gain $g_{i,j}$ and the corresponding $\hat{\mathbf{S}}$ and $\hat{\mathbf{D}}$.
  IF $g_{i,j} > g^*$
  update $g^* = g_{i,j}$, $\mathbf{S}^* = \hat{\mathbf{S}}$, $\mathbf{D}^* = \hat{\mathbf{D}}$,
  $\widetilde{i} = i$, and $\widetilde{j} = j$.
  End
  End
  End
End
update $\phi_{\widetilde{i},\widetilde{j}} = 1$, $\mathbf{S} = \mathbf{S}^*$ and $\mathbf{D} = \mathbf{D}^*$.

the process, and return the delay improvement $g_{i,j} = 0$; if not, we set $g_{i,j} = \omega_i P_{i,j} D_{i,j}$ because that is the reduction in η at user U_i if it adds $\mathcal{F}_j$ to its cache. Then, we need to sum up all reductions at each user. At user U_k, if $D_{k,j} > T_{i,k}$, then D2D link $U_i - U_k$ has the lowest average delay for U_k to receive $\mathcal{F}_j$ and the reduction at U_k is $\omega_k P_{k,j}(D_{k,j} - T_{i,k})$; if not, then caching $\mathcal{F}_j$ at U_i does not help to improve the delay performance at U_k.

Based on Algorithm 8.1, Algorithm 8.2 described in Table 8.2 helps to find the optimal <file,user> pair to be added to the updated caching result, which leads to the maximum delay reduction. In Algorithm 8.2, $\widetilde{i}$ and $\widetilde{j}$ record the optimal user index and file index, respectively. g^* tracks the maximum delay improvement, and $\mathbf{S}^*$ and $\mathbf{D}^*$ record the new source table and minimum delay matrix, respectively, after

Table 8.3: Algorithm 8.3

Caching Algorithm
Input : weight vector $\boldsymbol{\omega} = (\omega_1, \cdots, \omega_N)$, popularity matrix $\mathbf{P}$, and delay matrix $\mathbf{T}_{\text{avg}}$.
Output : caching indicator matrix $\mathbf{\Phi}$, source table $\mathbf{S}$.
Initialization : for all requests, $S_{i,j} \leftarrow \mathbf{BS}$, $D_{i,j} = T_{i,i}$. Set all $\phi_{i,j} = 0$.
For $loop = 1 : N\mu$
run Algorithm 8.2 to cache a file and update the result.
End

caching $\mathcal{F}_{\widetilde{j}}$ at $U_{\widetilde{i}}$. We search over all NM possible <file,user> combinations, find their delay improvements and update g^*, $\widetilde{i}$, $\widetilde{j}$, $\mathbf{S}^*$ and $\mathbf{D}^*$ accordingly. At user U_i, we check if there is empty space in its cache. If its cache is full, we directly jump to the next user U_{i+1}. For each <file,user> pair, we run Algorithm 8.1 to calculate the corresponding delay improvement, and compare it with g^*. If a <file,user> pair exceeds the maximum delay improvement up to that point, we perform the update accordingly. Every time we run Algorithm 8.2, we cache one more file at a user. Therefore, we need to run Algorithm 8.2 $N\mu$ times to obtain the final caching result, and this process is described in Algorithm 8.3 in Table 8.3.

For our proposed caching algorithm, we initially have all caches empty, and all users work in cellular mode, in which they only receive files from the base station at first. We assume that the system has calculated the average delay between every two nodes, and stored the delay matrix $\mathbf{T}_{\text{avg}}$ at the base station. Then, base station runs Algorithm 8.2 $N\mu$ times, and in each time we cache one more file and update the caching indicator $\mathbf{\Phi}$, source table $\mathbf{S}$, and minimum delay matrix $\mathbf{D}$ accordingly. Finally, the base station sends the caching files to the users when the traffic load is low.

8.1.2.2 Complexity Analysis

In the l^{th} iteration, Algorithm 8.2 searches over $NM-(l-1)$ possible <file,user> pairs, where the term $l-1$ corresponds to the $l-1$ <file,user> pairs that have been selected in previous iterations. Therefore, the size of the search space of our algorithm is $\sum_{l=1}^{l=N\mu} NM-(l-1) = N^2M\mu - \frac{1}{2}N^2\mu^2 + \frac{1}{2}N\mu$, which is much smaller than the size of the entire solution space $(\frac{M!}{(M-\mu)!\,\mu!})^N$.

In order to test the performance of our algorithm, we compare our algorithm with the brute-force exhaustive search algorithm. We apply both algorithms to a system, in which there are 5 users, 10 files in the library and each user can cache 2 files. These two algorithms obtain the same caching result, however the time consumption of the exhaustive search algorithm is 1.28×10^5 seconds, while our algorithm only takes only 2.7×10^{-3} seconds.

8.1.3 Extensions and Future Work

In this subsection, we consider a more general case, in which the delay matrices $\mathbf{T}_{\text{avg}}$ and $\mathbf{D}$, weight vector $\boldsymbol{\omega}=(\omega_1,\cdots,\omega_N)$, popularity matrix $\mathbf{P}$ and transmission powers P_i change over time. For simplicity, we assume that all these parameters stay constant within one update cycle, and we use κ as the index of cycles. The duration of the κ^{th} cycle, denoted by τ^κ, depends on how fast the parameters vary. Then, we can formulate our caching problem in the κ^{th} cycle as

$$\textbf{P2:}\quad \text{Minimize}_{\ \mathbf{\Phi}^\kappa}\quad \sum_{i=1}^{N}\omega_i^\kappa\sum_{j=1}^{M}P_{i,j}^\kappa D_{i,j}^\kappa \tag{8.11}$$

$$\text{Subject to}\quad \sum_{j=1}^{M}\phi_{i,j}^\kappa=\mu \tag{8.12}$$

$$\sum_{j=1}^{M}\left|\phi_{i,j}^\kappa-\phi_{i,j}^{\kappa-1}\right|\le 2\xi_i^\kappa \tag{8.13}$$

$$\phi_{i,j}^\kappa\in\{0,1\}. \tag{8.14}$$

If we define the weight of user i as $\omega_i^\kappa = \mathbb{E}\{\mathrm{NPK}_i^\kappa/\mathrm{NPK}^\kappa\}$, where NPK_i^κ and NPK^κ represent the number of received packets in the κ^{th} cycle at user i and at all users, respectively, then the objective function in the optimization problem **P2** represents the expected packet delay in the κ^{th} cycle. The transmission power P_i^κ is determined according to the battery budget of user i. Due to the changes in transmission powers and the distributions of channel fading, the delay matrices ${\mathbf{T}_{\text{avg}}}^\kappa$ and $\mathbf{D}^\kappa$ also vary over time. Compared with **P1**, **P2** includes an additional constraint given by (8.13). In (8.13), ξ_i^κ is the upper bound of the number of cache files that will be replaced in the current update cycle. Due to requirements regarding energy efficiency and current traffic load, each user may be able to update only a few cache contents.

The solution of **P2** is described below:

1. At the beginning of the κ^{th} cycle, the system estimates the delay matrix ${\mathbf{T}_{\text{avg}}}^{\kappa-1}$, weight vector $\boldsymbol{\omega}^{\kappa-1}$, and popularity matrix $\mathbf{P}^{\kappa-1}$ according to the samples obtained in the previous cycle. The base station receives the transmission powers P_i^κ from the users, determine the cycle period τ^κ and the upper bound ξ_i^κ, and then predicts ${\mathbf{T}_{\text{avg}}}^\kappa$, $\boldsymbol{\omega}^\kappa$ and $\mathbf{P}^\kappa$.

2. Algorithm 8.2 is repeated $N\mu$ times to determine the caching result in the κ^{th} cycle.

3. At the end of each iteration in the second step, it is checked if the constraint in (8.13) is satisfied with equality at any one of the users. If this constraint is satisfied with equality at a user, then no more cache updating is allowed for this user, meaning that this user can only choose from the files that are already stored in its cache in the remaining iterations.

After this process, the base station sends the cache contents to each user, and conduct regular transmission after updating the cache files at each user.

As we have mentioned in Section 8.1.1, this improved algorithm does not require the first 4 resource allocation assumptions described in Section 8.1.1.1, and works for any resource allocation algorithm, since the delay matrices $\mathbf{T}_{\text{avg}}$ and $\mathbf{D}$ need to be evaluated via estimation or learning methods. Also, we note that this method requires estimation algorithms in the first step. Due to the page limitations, we leave a detailed study of this problem as our future work.

8.1.4 Numerical Results

In this subsection, we investigate the performance of our proposed algorithm via numerical results. Since the estimation and resource allocation components required for the extended algorithm in Section 8.1.3 are beyond our scope, we only consider Algorithm 8.3 and its corresponding system model in this subsection. In the numerical results, the location of each user is randomly generated within a circular cell with the base station placed at the center. Each point in the figures is obtained by taking average over 500 randomly generated systems. The popularity matrix is generated according to the Zipf distribution. When the users have identical popularity, they give the same popularity index to a file, which leads to identical rows in the popularity matrix P. When the users have independent popularity, each user gives popularity indices to the files independently. In other words, identical popularity indicates that all users have the same preference, while independent popularity indicates that each user has an independent preference. The number of files in the library is $M = 100$, and the size of each file is 11.3 bits. We assume Rayleigh fading with path loss $\mathbb{E}\{z\} = d^{-4}$, where d represents the distance between the transmitter and the receiver. The transmission powers are set as $P_b = 23\,dB$ and $P_u = 20\,dB$, and we choose the weights as $\omega_i = \hat{p}_i$ so that η represents the average system delay.

In the numerical results, we compare the performance of our proposed algorithm with a naive algorithm, in which each user just caches the most popular μ files. This

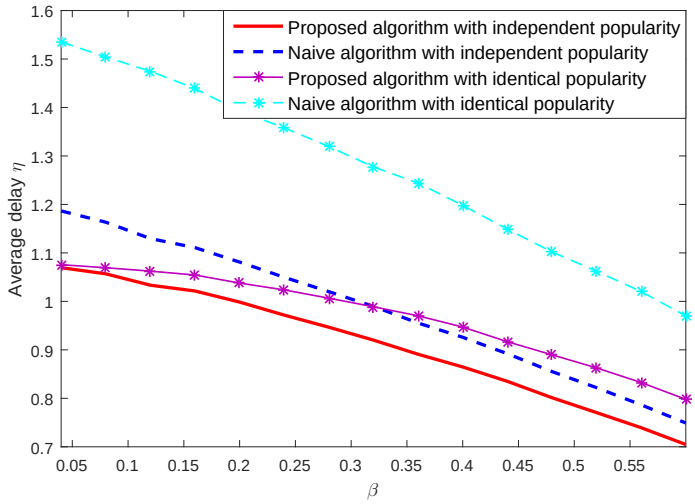

Figure 8.2: Average delay η vs. Zipf exponent β

naive algorithm is efficient when the base station does not have the knowledge of the channel fading statistics and the cached files at each user. In this circumstance, the users just cache files according to their own preference. In the case of naive algorithm with identical popularity, every user caches the same files, and they get the files they do not have via cellular downlink from the base station. Therefore, the gap between the two curves using naive algorithm in Figs. 8.2-8.4 (which will be discussed in detail next) demonstrates the benefit of enabling D2D communications. By allowing D2D transmission, the users far away from the base station can get files from their neighbors, which helps to significantly reduce the delay.

In Fig. 8.2, we set $N = 25$, $\mu = 30$ and plot the average delay η as a function of the Zipf exponent β. As β increases, the popularity difference increases. When $\beta = 0$, the users request all files with equal probability; when $\beta \to +\infty$, each user only requests its most favorite file. Therefore, we only need to concentrate on the delay performance of fewer popular files as β increases, and it becomes easier to achieve better delay performance with limited caching space. That is the reason for having monotonically decreasing curves in Fig. 8.2. Another observation is that our

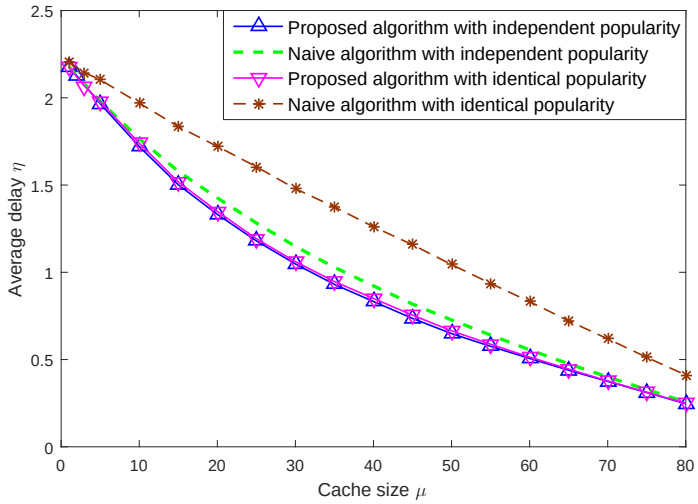

Figure 8.3: Average delay η vs. cache size μ

algorithm is more robust to the popularity setting. Compared to the curves using the naive algorithm, identical popularity model only slightly raises the delay of our algorithm. If a node can get a popular file from its near neighbor, then caching some less popular files might give better delay improvement. Therefore, our algorithm can enable D2D transmission even in an identical popularity model, which guarantees the robustness.

In Fig. 8.3, we select $\beta = 0.1$, $N = 25$ and plot the average delay as a function of the caching size μ. When μ is small, the delay difference between different algorithms and different popularity settings is small. In such a situation, both algorithms cache the most popular files. As μ increases, the difference in performance increases. As we have mentioned in Algorithm 8.2, our algorithm searches for the optimal <file,user> pair that provides the maximum delay improvement, and this mechanism guarantees a very sharp decrease at the beginning. After exceeding a threshold, further increasing the caching size reduces the performance difference, because the system gets enough caching size to cache most of the popular files. Overall, Fig. 8.3 shows that our algorithm can achieve better delay performance with limited caching size.

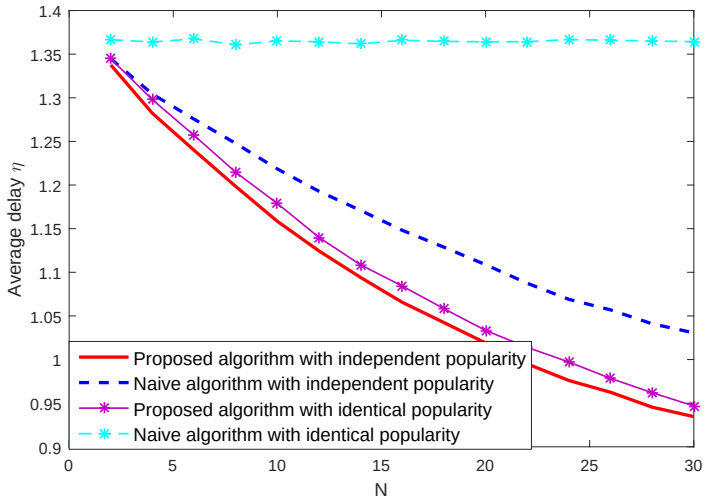

Figure 8.4: Average delay η vs. the number of users N

In Fig. 8.4, we select $\beta = 0.1$, $\mu = 30$ and plot the average delay as a function of the number of users N. For the curve using the naive algorithm with identical popularity model, having more users does not affect the average delay because each user works in cellular mode and receives the files from the base station. For other curves, increased number of users enables more chances for D2D communication, and as a result the average delay decreases. Compared with the naive algorithm, our algorithm can achieve better performance when the number of users is large.

8.2 Scheduling in D2D Underlaid Cellular Networks with Deadline Constraints

8.2.1 System Model and Transmission Modes

8.2.1.1 System Model

We consider an OFDMA cellular network with N available orthogonal channels, one base station (**BS**), N_c cellular users $\{\mathbf{CU}_1, \mathbf{CU}_2, \cdots, \mathbf{CU}_{N_c}\}$ and N_d D2D pairs

$\{(\mathbf{DT}_1, \mathbf{DR}_1), (\mathbf{DT}_2, \mathbf{DR}_2), \cdots, (\mathbf{DT}_{N_d}, \mathbf{DR}_{N_d})\}$. Each channel is assumed to have a bandwidth of B. Each cellular user transmits to the base station through an uplink channel, and receives data from the base station via a downlink channel. D2D transmission is assumed to be one-way between a D2D pair, in which $\mathbf{DT}_i$, the transmitter of the i^{th} D2D pair, sends packets to its corresponding receiver $\mathbf{DR}_i$. The maximum transmission power of cellular users and D2D transmitters are set at $\hat{P}_c$ and $\hat{P}_d$, respectively. When acting as a transmitter, the maximum transmission power of base station is $\hat{P}_b$ in each channel. We assume that the time is slotted, and each time slot has a duration of T. At the beginning of each time slot, the system runs a scheduling algorithm to allocate its channels to the users. Those users to which channels are not assigned are not activated for communication until they get an available channel in another time slot.

In a cellular network with D2D users, D2D users can choose to transmit trough a direct link $\mathbf{DT}_i - \mathbf{DR}_i$ or a two-hop link $\mathbf{DT}_i - \mathbf{BS} - \mathbf{DR}_i$. When it transmits through the base station, a D2D transmitter first sends packets to the base station, and then the base station decodes and forwards the packets to the corresponding D2D receiver. When a pair of D2D users chooses to communicate through the direct link, they can also decide whether to reuse the same channel with an uplink, a downlink or another D2D direct link. In this section, mode selection is performed by the scheduling algorithm, and the mode of each active user is determined by the channel allocation results. There are 4 assumptions for the channel assignment:

1. Each active link can occupy only one channel, and a D2D two-hop link is also regarded as one transmission link.

2. Each channel can be occupied by two transmission links at most.

3. A two-hop link cannot share its channel with other transmission links.

4. Cellular uplinks and downlinks can only share their channels with D2D direct

links.

According to our assumptions, there are overall 7 different modes, namely D2D cellular mode, uplink reuse mode, downlink reuse mode, D2D reuse mode, uplink dedicated mode, downlink dedicated mode, and D2D dedicated mode. A detailed discussion about these possible modes is given in Section 8.2.1.2.

The channel fading is assumed to be block fading, in which the fading coefficients denoted by h stay constant in one time block and change independently across blocks. In Figs. 8.5, 8.6 and 8.7, the magnitude-square of the fading coefficients are denoted by $z = |h|^2$. At each receiver, the background noise is assumed to follow an independent complex Gaussian distribution with zero mean and variance σ^2, i.e., $n \sim \mathcal{CN}(0, \sigma^2)$.

For all transmitters, the data packets are stored in buffers before being sent to the corresponding receiver. For simplicity, we assume that new packets arrive at the beginning of each time slot, and the size of each packet is assumed to be fixed at I_p bits. We further assume that there are totally $2N_c + N_d + 1$ buffers in the system, operating in a first-in first-out (FIFO) manner. At each cellular user and D2D transmitter, there is a buffer storing the packets for cellular uplinks and D2D links, respectively. Although there might be only one physical buffer at the base station in reality, we can decompose it into $N_c + 1$ virtual buffers, in which N_c of them correspond to cellular downlinks, and there is one special buffer corresponding to the D2D cellular mode. We enumerate all these buffers from 1 to $2N_c + N_d$, except the one corresponding to the D2D cellular mode. The system operates under deadline constraints, and the delay upper bound of the i^{th} buffer is set to be D_i. In the following subsection, we introduce all possible transmission modes and describe the relationship between mode selection and channel allocation.

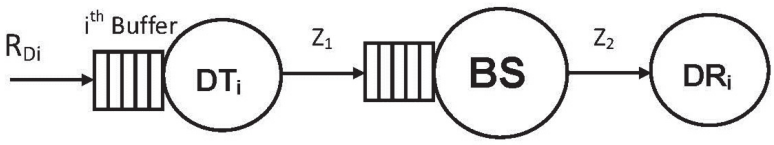

Figure 8.5: System model in the D2D cellular mode

8.2.1.2 Transmission Modes and Instantaneous Transmission Rate

In this section our mode selection is done through scheduling. Depending on how the system uses each channel, we can determine the transmission mode for each user. According to our channel allocation assumptions, there are 7 possible modes, which can be further summarized into 3 categories.

D2D Cellular Mode

The first category is D2D cellular mode, and the model is shown in Fig. 8.5. In this mode, the channel is occupied by a D2D pair, transmitting with the help of the base station, and the packet arrival rate at $\mathbf{DT}_i$ is represented by R_{D_i}. For simplicity, we assume that the whole time slot is divided into two sub-slots with duration τT and $(1-\tau)T$, respectively[3]. In the first sub-slot, D2D transmitter $\mathbf{DT}_i$ sends packets to the base station, and base station stores the received packets in the special buffer corresponding to the D2D cellular mode. In the second sub-slot, base station forwards all the packets received in the first sub-slot to their destination $\mathbf{DR}_i$[4].

In each sub-slot, there is only one transmitter and one receiver, and the received signal at each receiver follows the form

$$y = hx + n, \tag{8.15}$$

[3]In this section, we assume that the time cost for signal processing at the base station is negligible in cellular mode.

[4]No deadline constraints are imposed on this special buffer at the base station, because all packets sent in the first sub-slot arrive at their destination in the second sub-slot

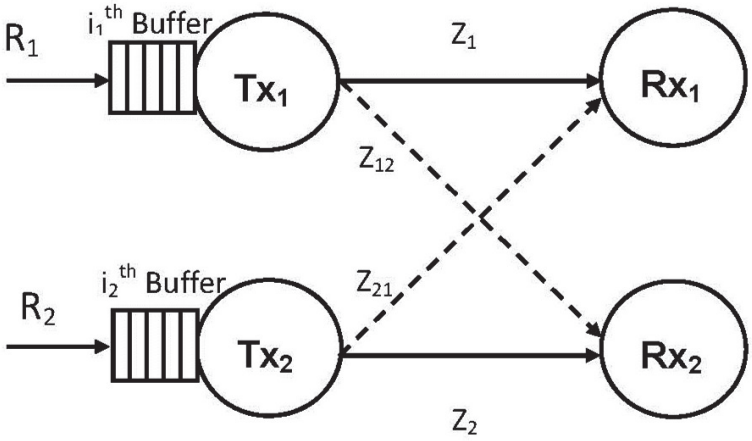

Figure 8.6: System model in the reuse mode

where x is the transmitted signal, n is the additive Gaussian noise component, h is the corresponding channel fading coefficient. We can express the packet transmission rates (in packets/slot) of $\mathbf{DT}_i - \mathbf{BS}$ link and $\mathbf{BS} - \mathbf{DR}_i$ link as

$$r_{D_i,BS}(\tau, P_d) = \left\lfloor \tau \frac{TB}{I_p} \log_2 \left(1 + \frac{P_d}{B\sigma^2} z_1\right) \right\rfloor \tag{8.16}$$

and

$$r_{BS,D_i}(\tau, P_b) = \left\lfloor (1-\tau) \frac{TB}{I_p} \log_2 \left(1 + \frac{P_b}{B\sigma^2} z_2\right) \right\rfloor, \tag{8.17}$$

respectively, where I_p is the packet size, P_d and P_b are the transmission powers of $\mathbf{DT}_i$ and base station respectively, and $\lfloor \bullet \rfloor$ represents the floor function. The parameter τ is determined from $r_{D_i,BS} = r_{BS,D_i}$.

Reuse Mode

The second category is the reuse mode, and the system model is shown in Fig. 8.6. In this model, two transmitter-receiver pairs share the same channel, and they inflict interference on each other. According to the types of users sharing the channel, reuse mode includes uplink reuse mode, downlink reuse mode, and D2D reuse mode. In

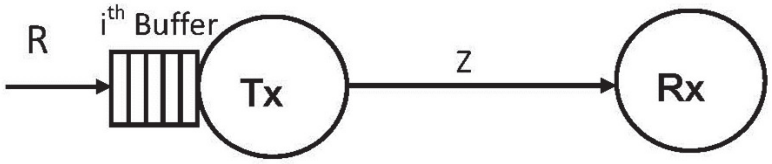

Figure 8.7: System model in the dedicated mode

the uplink reuse mode, a cellular uplink shares the channel with a D2D direct link; in the downlink reuse mode, a cellular downlink shares the channel with a D2D direct link; in D2D reuse mode, two pairs of D2D users transmit in the same channel.

The received signal at each receiver follows the form

$$y = hx + h_{\text{inter}}x_{\text{inter}} + n, \tag{8.18}$$

where x is the desired signal, h is the fading coefficient of the channel between this receiver and its corresponding transmitter, x_{inter} is the interference signal, h_{inter} is the fading coefficient of the interfering link, and n is the Gaussian noise. Treating the interference as noise, we can express the instantaneous packet transmission rates in these two links as

$$r_1(P_1, P_2) = \left\lfloor \frac{TB}{I_p} \log_2 \left(1 + \frac{P_1 z_1}{B\sigma^2 + P_2 z_{21}}\right) \right\rfloor \tag{8.19}$$

and

$$r_2(P_1, P_2) = \left\lfloor \frac{TB}{I_p} \log_2 \left(1 + \frac{P_2 z_2}{B\sigma^2 + P_1 z_{12}}\right) \right\rfloor, \tag{8.20}$$

respectively. In (8.19) and (8.20), P_1 and P_2 denote the transmission powers of $\mathbf{Tx_1}$ and $\mathbf{Tx_2}$, respectively.

Dedicated Mode

The third category is the dedicated mode, which is depicted in Fig. 8.7. In this model, one transmitter-receiver pair occupies a channel without sharing it with others, and

the packet arrival rate is represented by R. Depending on the type of the transmission link occupying the channel, this category includes uplink dedicated mode, downlink dedicated mode, and D2D dedicated mode. Since there is no interference, the received signal at the receiver also follows the form given in (8.15), and the instantaneous rate can be expressed as

$$r(P) = \left\lfloor \frac{TB}{I_p} \log_2 \left(1 + \frac{P}{B\sigma^2} z\right) \right\rfloor, \tag{8.21}$$

where P is the transmission power.

8.2.2 Scheduling with Convex Delay Cost Method

8.2.2.1 Convex Delay Cost Function

The convex delay cost approach was proposed in [55], where a monotonic increasing convex cost function was employed for the packet delay. In our analysis, we use the same convex cost function proposed in [56]. More specifically, for the j^{th} packet in the i^{th} buffer with delay $d_{i,j}$, the delay cost is given by

$$C_{i,j}(d_{i,j}) = \left(\frac{d_{i,j}}{D_i}\right)^{\alpha}, \tag{8.22}$$

where D_i is the corresponding delay threshold for the packets in the i^{th} buffer, and $\alpha \geq 0$ is a relaxation parameter. As α increases, the cost grows faster when the delay increases beyond the threshold.

At the beginning of each slot, the delays for newly arrived packets are set as 0. [5] At the end of each slot, packet delay for all packets that have not been sent increases by 1. If we denote the length of the packet queue in the i^{th} buffer as l_i, then the

[5] When $\alpha = 0$, the cost of a new packet is defined as 1

overall cost of the entire cellular network can be expressed as

$$C = \sum_{i=1}^{2N_c+N_d} \sum_{j=1}^{l_i} C_{i,j}(d_{i,j}). \tag{8.23}$$

8.2.2.2 Scheduling Decisions and Utility

The decision of a single channel assignment can be denoted by a set of active links that occupy this channel. Therefore, for the 3 dedicated modes, the corresponding decisions only contain a single direct link; for the D2D cellular mode, the decision contains a two-hop channel; for the 3 reuse modes, the decisions contain two direct links sharing the same channel. For each channel, there are overall $N_{DC} = 2N_dN_c + 2(N_d + N_c) + N_d(N_d - 1)/2$ different possible decisions at most, and we enumerate all these decisions from 1 to N_{DC}. For each channel, we select the optimal decision from all possible candidates, that minimizes the overall cost.

If the system does not make any decision, then the overall cost would become

$$\tilde{C}_0 = \sum_{i=1}^{2N_c+N_d} \sum_{j=1}^{l_i} C_{i,j}(d_{i,j} + 1), \tag{8.24}$$

at the end of the current slot. If the k^{th} decision is selected, then the overall cost would be

$$\tilde{C}_k = \sum_{i=1}^{2N_c+N_d} \sum_{j=\mu_{i,k}+1}^{l_i} C_{i,j}(d_{i,j} + 1), \tag{8.25}$$

where $\mu_{i,k}$ is the instantaneous departure rate (in packets/slot) from the i^{th} buffer. We define the utility of the k^{th} decision as

$$U_k = \tilde{C}_0 - \tilde{C}_k = \sum_{i=1}^{2N_c+N_d} \sum_{j=1}^{\mu i,k} C_{i,j}(d_{i,j} + 1). \tag{8.26}$$

Since the decision with the highest utility can minimize the overall cost, the scheduling algorithm allocates each channel to the transmission link(s) giving the highest utility. In the next subsection, we discuss the utility maximization for each type of decision.

8.2.3 Utility Maximization and Scheduling Algorithm

In previous subsections, we have characterized the instantaneous transmission rates for all possible modes, and formulated the cost function and utility. In this subsection, we first provide utility maximization algorithms for all possible modes, and propose our scheduling algorithm. In our scheduling algorithm, utility maximization is the same as power optimization, in which we find the optimal transmission powers that maximize the utility for a scheduling decision.

8.2.3.1 Utility Maximization in Dedicated Modes

For uplink, downlink and D2D dedicated mode, there is only one direct transmission link occupying the channel, and the instantaneous transmission rate $r(P)$ is given by (8.21). In order to maximize the utility, the transmitter just transmits with its maximum power. Suppose that Fig. 8.7 describes the k^{th} decision, in which the channel is occupied by a direct transmission link with maximum power P_{max}. The instantaneous departure rate is

$$\mu_{i,k} = \min\left\{r(P_{max}), l_i\right\}. \tag{8.27}$$

Then the maximum utility is given by

$$U_k = \sum_{j=1}^{\mu_{i,k}} C_{i,j}(d_{i,j}+1). \tag{8.28}$$

8.2.3.2 Utility Maximization in D2D Cellular Mode

In D2D cellular mode, a pair of D2D users transmit through a two-hop channel, and the base station works as the relay. In order to maximize the utility, the D2D transmitter and base station transmit with their maximum power. Assume that Fig. 8.5 describes the k^{th} decision. By setting $P_d = \hat{P}_d$ and $P_b = \hat{P}_b$, the instantaneous transmission rates $r_{D_i,BS}$ and r_{BS,D_i} are given by (8.16) and (8.17), respectively. The optimal τ value is given by

$$\tau^* = \frac{\log_2\left(1 + \frac{\hat{P}_b}{B\sigma^2} z_2\right)}{\log_2\left(1 + \frac{\hat{P}_d}{B\sigma^2} z_1\right) + \log_2\left(1 + \frac{\hat{P}_b}{B\sigma^2} z_2\right)}, \tag{8.29}$$

which arises from $r_{D_i,BS} = r_{BS,D_i}$. Inserting the optimal τ value back into (8.16), we get the instantaneous departure rate as

$$\mu_{i,k} = \min\left\{r_{D_i,BS}(\tau^*, \hat{P}_d), l_i\right\}, \tag{8.30}$$

and the maximum utility is also given by (8.28).

8.2.3.3 Utility Maximization in Reuse Modes

In uplink, downlink and D2D reuse modes, two direct links transmit simultaneously in the same channel. Suppose that Fig. 8.6 describes the k^{th} decision, in which the channel is occupied by two direct links, connecting with the i_1^{th} and i_2^{th} buffer, respectively. The instantaneous transmission rates r_1 and r_2 are given by (8.19) and (8.20), respectively, and the departure rates of these two buffers are given by

$$\mu_{i_s,k} = \min\left\{r_s, l_{i_s}\right\}, \tag{8.31}$$

for $s = 1, 2$. The utility is given by

$$U_k = \sum_{j=1}^{\mu i_1,k} C_{i_1,j}(d_{i_1,j} + 1) + \sum_{j=1}^{\mu i_2,k} C_{i_2,j}(d_{i_2,j} + 1). \tag{8.32}$$

Then the optimization problem can be formulated as

$$\begin{aligned} &\text{Maximize}_{P_1,P_2} && U_k \\ &\text{Subject to} && 0 \leq P_1 \leq P_{1max} \qquad (8.33) \\ & && 0 \leq P_2 \leq P_{2max}, \qquad (8.34) \end{aligned}$$

which is not a convex optimization problem. In addition, derivative-based algorithms are not applicable, due to the floor function in (8.19) and (8.20).

Since the transmission rates r_1 and r_2 only take integer values, the optimization problem can be solved by an efficient search algorithm. If we fix the value of r_1, then the maximum utility is achieved when r_2 is maximized. Therefore, we can search over all possible r_1 values, and for any given r_1, the optimal utility is given by the power values P_1 and P_2 which maximize r_2. When we fix r_1, we can get

$$P_1 = \left(2^{\frac{r_1 I_p}{TB}} - 1\right)(P_2 z_{21} + B\sigma^2)/z_1 \tag{8.35}$$

and

$$r_2 = \left\lfloor \frac{TB}{I_p} \log_2 \left(1 + \frac{P_2 z_2}{B\sigma^2 + \frac{z_{12}}{z_1}\left(2^{\frac{r_1 I_p}{TB}} - 1\right)(P_2 z_{21} + B\sigma^2)} \right) \right\rfloor \tag{8.36}$$

from (8.19) and (8.20).

We can easily verify that r_2 given by (8.36) is an increasing function of P_2. Therefore, r_2 is maximized when P_2 achieves its upper bound. From (8.33) and (8.35), we

Table 8.4: Algorithm 8.4

Utility Maximization for Reuse Mode

1. **Initialization**:
   (a) Set $r_2 = 1$ and $P_1 = P_{1max}$. Solve P_2 from 8.20, find r_{1max} and $\mu_{i_1,k\ max}$ using (8.19) and (8.31), respectively. Set the optimal utility value U_k^* as -1.
   (b) If $\mu_{i_1,k\ max} = 0$, , and then end the utility maximization process.
2. **Searching**:
   For $r_1 = 1 : \mu_{i_1,k\ max}$
   (a) Compute the transmission power P_1 and P_2 using (8.35) and (8.38), respectively. Then find r_2 and $\mu_{i_2,k}$ from (8.20) and (8.31), respectively.
   (b) Compute the utility U_k using (8.32). If $U_k > U_k^*$, update the optimal utility value as $U_k^* = U_k$.
   end

get one upper bound on P_2 expressed as

$$P_2 \leq \frac{z_1 P_{1max}}{z_{21}\left(2^{\frac{r_1 I_p}{TB}} - 1\right)} - \frac{B\sigma^2}{z_{21}}. \tag{8.37}$$

Combining this result with (8.34), the optimal P_2 is given by

$$P_2^* = \min\left\{P_{2max}, \frac{z_1 P_{1max}}{z_{21}\left(2^{\frac{r_1 I_p}{TB}} - 1\right)} - \frac{B\sigma^2}{z_{21}}\right\}. \tag{8.38}$$

The overall utility optimization algorithm is given in Table 8.4.

8.2.3.4 Scheduling Algorithm

After providing the utility maximization algorithms for all possible modes, we now propose our scheduling algorithm. At the beginning of each slot, we schedule the channels one by one. For each channel, we compute the maximum utility values that each decision can achieve, and assign the channel according to the decision that gives the maximum utility value. The algorithm is described in Table 8.5. In order to have

Table 8.5: Algorithm 8.5

Scheduling Algorithm
Randomly enumerate the channels from 1 to N. **For** $n = 1 : N$ 1. Using the fading coefficients in the n^{th} channel, compute the maximum utility values for all decisions in the possible decision set. In the first loop, there are N_{DC} possible decisions. 2. Assign the n^{th} channel according to the decision that gives the maximum utility. 3. Remove all decisions that contain the links in the selected decision from the possible decision set. Therefore, the links in the selected decision would not be selected by other channels. **end**

lower complexity and high performance, we randomly enumerate the channels so that the order of channels would not affect our algorithm.

Assume the number of users is $N_u = N_c + N_d$, then it is easy to show that $N_{DC} \leq \frac{16N_u^2+40N_u+1}{24}$, where equality is achieved at $N_d = \frac{4N_u-1}{6}$. Therefore, the complexity of this algorithm is $o(NN_{DC})$, which is also equal to $o\left(NN_u^2\right)$).

8.2.4 Numerical Results

In this subsection we further investigate the performance of our scheduling algorithm via numerical results. In our numerical results, we use Monte Carlo simulations to obtain the delay violation probability, average delay, and average throughput. We consider Rayleigh fading with path loss $\mathbb{E}\{z\} = d^{-4}$, where d is the transmission distance, and we fix $N = N_c = N_d = 3$, $\hat{P}_c = \hat{P}_d = 26.99$ dB, $\hat{P}_b = 30$ dB, $\sigma^2 = 0$ dB. For all users, we select the delay thresholds as 30 time slots. We use constant arrival rate in the simulations, and the arrival rates are set as $R = \rho N\mathbb{E}\{r_{\text{Direct}}\}/(2N_c + N_d)$, where r_{Direct} is the transmission rate of the corresponding direct link, and ρ is the intensity parameter. When generating the position of each node, we fix the base station at the origin of the coordinate axes, and place the D2D and cellular users randomly in the cell with coverage radius equal to 3. In order to eliminate the influence of the random positions, we generate 100 systems randomly, in which all

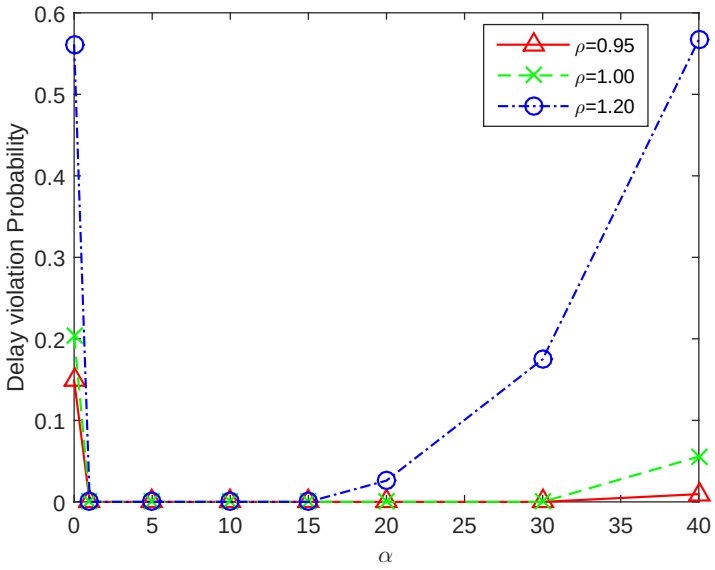

Figure 8.8: Delay violation probability vs. α

nodes are uniformly distributed, and each point we plot in our figures is obtained by averaging the results from these 100 systems. Each simulation is conducted over 2×10^4 time slots.

In Figs. 8.8 and 8.9, we plot the delay violation probability and average delay as functions of the relaxation parameter α. From (8.22), we can see that as α increases, the importance of the maximum packet delay grows. When $\alpha = 0$, the utility just depends on the transmission rate, and our scheduling algorithm becomes a greedy algorithm that maximizes the instantaneous transmission rate. When α is sufficiently large, the utility is mainly decided by the maximum packet delay of each buffer, and our algorithm would allocate the channel to the user(s) with the largest delay. From Figs. 8.8 and 8.9, we can see that $\alpha = 1$ gives the best performance with very small delay and high throughput. The scheduling algorithm has degraded performance when the balance between the importance of instantaneous transmission rate and packet delay is broken. When $\alpha = 0$, the algorithm does not use any delay information; when α is too large, the algorithm may often allocate the channels to the

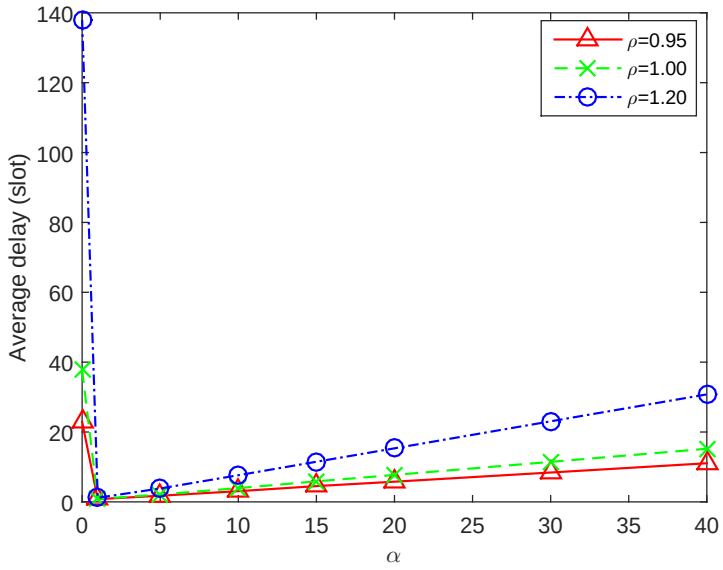

Figure 8.9: Average Delay vs. α

users with large packet delay even when their channel conditions are not favorable, which lowers the throughput and increases the average delay.

Figs. 8.10 and 8.11 show the advantage of using reuse modes. If we do not have reuse modes, the packet delay increases significantly. In order to satisfy the delay constraints and stabilize the system, the systems without reuse modes need to reduce their arrival rates, sacrificing the throughput. Therefore, by allowing D2D users to share channels with other users, D2D communication can have much better performance when deadline constraints are applied.

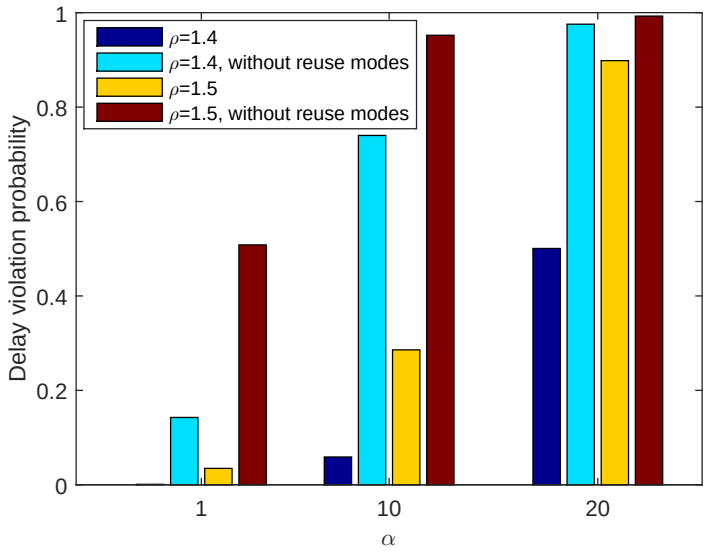

Figure 8.10: Delay violation probability vs. α

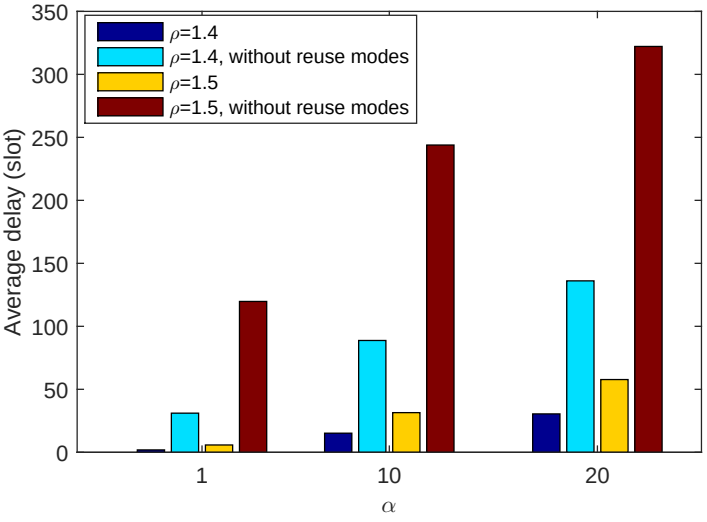

Figure 8.11: Average Delay vs. α

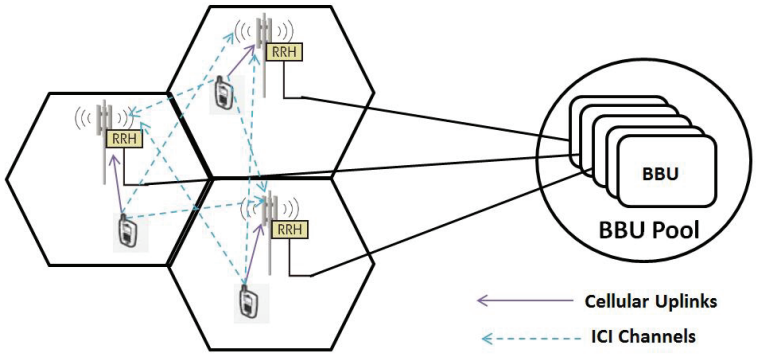

Figure 8.12: System model of C-RAN with ICI

8.3 Intercell Interference-Aware Scheduling for Delay Sensitive Applications in C-RAN

8.3.1 System Model and Preliminaries

8.3.1.1 System Model

In this section, the uplink transmission in a C-RAN within an OFDMA setting is considered as shown in Fig. 8.12. There are N_c cells in this network, and each cell is served by a base station with one RRH. RRHs are connected to a centralized BBU pool with multiple BBUs working cooperatively. All cells reuse N_{ch} frequency bands/channels, and each channel has a bandwidth of B. The total number of mobile users in this network is fixed at N_u, and users are assumed to be associated with their nearest RRHs. Each user is equipped with a buffer storing the arriving packets before sending them through the wireless uplink channels, and the size of each packet is assumed to be I_p bits. All buffers are assumed to operate in a FIFO manner. The system is assumed to operate under delay constraints, and target delay of packets sent by the i^{th} user is denoted by D_i (time frames). Block fading is assumed in this section, in which the fading coefficients stay constant within one time frame with a

duration of T, and change across frames. Also it is assumed that the distributions of the fading coefficients are identical in different channels.

At the beginning of each time frame, BBU pool allocates channel resources to the users using a scheduling algorithm. It is assumed that users keep silent until they get channel resources from the BBU pool, and the channel resources are returned back at the end of each time frame. There are 4 assumptions for the channel assignment:

1. The number of users is much greater than the number of available channels, $N_u \gg N_{ch}$. In such a case, each user transmits using one channel at most.

2. Only the users that can satisfy the pair-wise interference constraints given in (8.47) can reuse the same channel resource.

3. Users associated with the same RRH cannot reuse the same channel resource.

4. The BBU pool is assumed to have perfect CSI, and it is also assumed to keep track of the buffer status (including the queue length and packet delay information) of each user.

The first assumption addresses a heavy load scenario, in which all channels are reused by multiple users and ICI becomes a significant problem. In such a case, the assumption that each user transmits using one channel at most helps to reduce ICI caused by excessive frequency reuse. The second assumption limits the interference, and the third assumption guarantees that all interference comes from neighbouring cells. The last assumption guarantees that the BBU pool has enough information to conduct our scheduling algorithm. CSI is estimated at RRHs and sent to the BBU pool via optical fiber links. Information of the arrival rates at all users is also sent to the BBU pool via special feedback channels[6], and the BBU pool can track the queue status at each user.

[6]We assume ideal feedback without delay and error.

Define $\mathbf{\Psi}_j(t)$ as the set of users that use the j^{th} channel in the t^{th} time frame, and $\xi_{i,j}(t)$ as the indicator function that indicates whether the j^{th} channel is assigned to the i^{th} user in the t^{th} time frame. In other words, $\xi_{i,j}(t) = 1$ if $i \in \mathbf{\Psi}_j(t)$, otherwise $\xi_{i,j}(t) = 0$. According to our first channel assignment assumption, we have $\sum_{j=1}^{N_{ch}} \xi_{i,j}(t) \leq 1$. Then for the t^{th} time frame, the received signal corresponding to user i at its associated base station can be expressed as

$$y_i = h_i^j x_i + \sum_{k \in \mathbf{\Psi}_j(t), k \neq i} h_{k,i}^j x_k + n_i^j \tag{8.39}$$

if $\xi_{i,j}(t) = 1$. Above, x_i represents the transmitted signal of user i, h_i^j denotes the fading coefficient of the channel between user i and its corresponding RRH, $h_{k,i}^j$ denotes the fading coefficient of the interference channel between user k and the RRH associated with user i, and n_i^j is the background noise at the base station associated with user i which is assumed to follow an independent complex Gaussian distribution with zero mean and variance σ^2, i.e., $n_i^j \sim \mathcal{CN}(0, \sigma^2)$. The transmission rate of user i in the t^{th} time frame is given by

$$r_i(t) = TB \log_2 \left(1 + \frac{P_i z_i^j}{B\sigma^2 + \sum_{k \in \mathbf{\Psi}_j(t), k \neq i} P_k z_{k,i}^j} \right) \text{ bits/frame} \tag{8.40}$$

where j is the index of the channel that is assigned to user i, P_i represents the transmission power of user i, T is the duration of each time frame, B is the bandwidth of each channel, $z_i^j = |h_i^j|^2$, and $z_{k,i}^j = |h_{k,i}^j|^2$.

8.3.1.2 Convex Delay Cost and Utility

In the convex delay cost approach, the cost function of a packet is formulated as an increasing convex function of its delay [55]. The high performance of this approach was shown in [56] for a single cell model without any interference. In our previous work [97], we designed a scheduling algorithm using the convex cost function provided in

[56] for a D2D communication setting, and verified via simulations that this approach has very good delay performance. Here, we define the cost of the j^{th} packet in the buffer at user i as

$$C_{j,i} = \frac{d_{j,i}}{D_i}, \tag{8.41}$$

where $d_{j,i}$ is the current delay of this packet, and D_i is the target delay of user i. At user i, the number of packets that can be transmitted in the current time frame is

$$\mu_i = \min\left\{l_i, \lfloor r_i/I_p \rfloor\right\}, \tag{8.42}$$

where l_i is the number of packets waiting in the buffer at user i, I_p is the size of each packet, and $\lfloor \cdot \rfloor$ represents the floor function. The utility of user i is defined as

$$U_i = \sum_{j=1}^{\mu_i} C_{j,i}, \tag{8.43}$$

and the utility of the system is defined as

$$U = \sum_{i=1}^{N_u} U_i = \sum_{i=1}^{N_u} \sum_{j=1}^{\mu_i} C_{j,i}. \tag{8.44}$$

The utility given in (8.44) represents the total cost of the packets that can be transmitted to the base station in the current time frame. At the beginning of each time frame, the BBU pool runs a scheduling algorithm for channel assignment to maximize the utility. In the next subsection, a detailed discussion on our scheduling algorithm is provided.

8.3.2 ICI-Aware Scheduling Algorithm for C-RAN

In this subsection, we introduce our scheduling algorithm. In each time frame, our scheduling algorithm assign channels to the users in a way that maximizes the utility given in (8.44). Since we consider a C-RAN architecture, the BBU pool has the knowledge of all fading distributions and cost functions of each packets, and it can allocate channel resources to all users in different cells together. Our scheduling algorithm can be divided into two steps, namely the user grouping step and channel matching step. In the first step, we divide all users into small groups such that the users in the same group reuse the same channel. In the second step, we match the channels to the user groups to maximize the utility.

8.3.2.1 User Grouping

In the first step of our algorithm, we divide all users into small groups, and each group will be assigned a channel resource in the next step. Before channel assignment, we cannot compute the instantaneous transmission rates because the sets $\mathbf{\Psi}_1$, $\mathbf{\Psi}_2$, $\cdots$, $\mathbf{\Psi}_{N_{ch}}$ have not been determined yet. Therefore, we use a rate estimator

$$\hat{r}_i = \frac{1}{m} \sum_{\tau=t-m}^{t-1} r_i(\tau) \tag{8.45}$$

instead. This rate estimator is essentially the average rate over the most recent m time frames. Plugging (8.45) into (8.42) and (8.43), we obtain the utility estimator of user i as

$$\hat{U}_i = \sum_{j=1}^{\hat{\mu}_i} C_{j,i} = \sum_{j=1}^{\min\{l_i, \lfloor \hat{r}_i / I_p \rfloor\}} C_{j,i}. \tag{8.46}$$

In order to control ICI, we assume that any two users (i_1 and i_2) reusing the same

Table 8.6: Algorithm 8.6

Input: γ, transmission power and utility estimator of each user, the fading coefficients.
Output: User groups GP_1, GP_2, $\cdots$, GP_{N_g}.

Collect the utility estimators $\hat{U}_i$ into a vector $\mathbf{V} = [\hat{U}_1, \hat{U}_2, \cdots, \hat{U}_{N_u}]$.
Set $k = 1$
While $\max(\mathbf{V}) \geq 0$
  Set $\mathbf{V}^* = \mathbf{V}$ and $GP_k = \emptyset$
  While $\max(\mathbf{V}^*) \geq 0$
    $i = \arg\max(\mathbf{V}^*)$
    Add user i into GP_k.
    Set $\mathbf{V}(i) = -1$ and $\mathbf{V}^*(i) = -1$.
    For j from 1 to N_u
      Set $\mathbf{V}^*(j) = -1$ if user i and j cannot satisfy the interference constraints given in (8.47) or they are associated to the same RRH.
    End
  End
  $k = k + 1$
End

channel resource have to satisfy the pairwise interference/SINR constraints given by

$$\begin{cases} \mathbb{E}\left\{\frac{P_{i_1} z_{i_1}}{B\sigma^2 + P_{i_2} z_{i_2,i_1}}\right\} \geq \gamma \mathbb{E}\left\{\frac{P_{i_1} z_{i_1}}{B\sigma^2}\right\} \\ \mathbb{E}\left\{\frac{P_{i_2} z_{i_2}}{B\sigma^2 + P_{i_1} z_{i_1,i_2}}\right\} \geq \gamma \mathbb{E}\left\{\frac{P_{i_2} z_{i_2}}{B\sigma^2}\right\} \end{cases}, \tag{8.47}$$

where the parameter γ is between 0 and 1. Since the distributions of the fading coefficients are identical in different channels, the expected values of the SINRs and SNRs in (8.47) do not depend on the channel assignment result. The details of our user grouping algorithm is given in Table 8.6, and we denote the number of the output user groups as N_g.

At the beginning, we set group GP_k as an empty set. Each time, we select the user with the maximum utility estimator and include it into GP_k. After adding a user into a group, we kick out the users that cannot reuse the same channel resource with this selected user by setting $\mathbf{V}^*(j) = -1$, which can be processed in parallel at

the BBU pool. Our grouping algorithm aims to collect the users with high utility estimators together, which helps to serve these users with less channel resources.

Note that the number of groups N_g might be smaller than the number of channels N_{ch}. In such cases, some of the channels cannot be assigned to users, and we need to break those groups with large sizes into several small groups so that $N_g = N_{ch}$. To divide a big group into two small groups, we select half of the users with smaller utility estimator values within the large group, and let them form a new small group.

8.3.2.2 Channel Matching

In the second step, we assign channels to the user groups via the maximum-weight matching approach. In this step, we find a matching between user groups and channels that maximizes the system utility given in (8.44). Let us define $\eta_{i,j}$ as the indicator of the channel assignment result, i.e., $\eta_{i,j} = 1$ if channel j is assigned to GP_i, and $\eta_{i,j} = 0$ if channel j is not matched to GP_i. Then the matching problem can be formulated as

$$\begin{aligned} \textbf{Maximize} \quad & \underset{\eta_{i,j}}{} \quad U \\ \textbf{Subject to} \quad & \eta_{i,j} \in \{0,1\} \\ & \sum_{j=1}^{N_{ch}} \eta_{i,j} \leq 1 \\ & \sum_{i=1}^{N_g} \eta_{i,j} = 1. \end{aligned}$$

In graph theory, the maximum-weight matching problem can be solved by the Hungarian algorithm (Kuhn-Munkres algorithm) [92]. To use the Hungarian algorithm, we have to first construct the utility matrix $\mathbf{U}$, in which each row corresponds to a user group and each column corresponds to a channel. The element of this matrix $U_{i,j}$ is the sum utility of the users in GP_i if the j^{th} channel is assigned to that

group. The elements of the utility matrix can be computed in parallel at the BBU pool. After constructing the utility matrix, the Hungarian algorithm is applied, and channels are assigned to the users.

8.3.2.3 Summary and Complexity Analysis

In summary, we propose a two-step scheduling algorithm with good delay performance for a multi-cell C-RAN model. In the first step, we group the users to control the ICI and aim to collect the users with high utility estimator values into smaller number of groups. In the second step, we formulate the channel allocation problem as a maximum-weight matching problem, and assign the channel resources to the user groups using the Hungarian algorithm. Although our algorithm only considers an uplink scenario, it can also be easily adapted to a downlink scenario.

Since we consider a C-RAN model, our algorithm is performed considering users in multiple cells, and parallel processing can be performed in some parts of our algorithm at the BBU pool to reduce time consumption. Compared with conventional resource allocation algorithms, in which cooperative processing among multiple cells is not considered, our algorithm has a significant potential to achieve better performance.

Assume that the number of processers at BBU pool is $\Theta(N_c)$, then the time complexity of the user grouping step is $O(N_u^2/N_c)$. In the matching step, the time consumption for constructing the utility matrix is $O(N_g N_{ch}/N_c)$, and the time consumption of the Hungarian algorithm is $O(\max\{N_g, N_{ch}\}^3)$. To further accelerate this process, we can replace the Hungarian algorithm with some heuristic algorithms with time complexity of $O(\min\{N_g, N_{ch}\})$. As an example, in each iteration, we can select the maximum element in the utility matrix, and match its corresponding group and channel together. The overall time consumption of this algorithm depends on the relationship among N_u, N_c, N_g and N_{ch}.

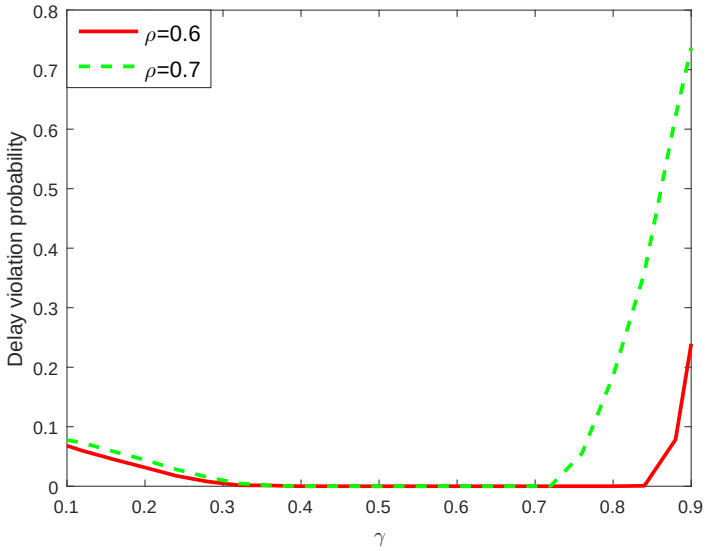

Figure 8.13: Delay violation probability vs. interference control parameter γ

8.3.3 Numerical Results

In this subsection, we further study the performance of our algorithm and the influence of parameters via simulations. In our simulations, we consider a C-RAN with 3 adjacent cells, each with a radius of 2. The coordinates of the RRHs of these three cells are $(-2, 0)$, $(0, 2)$ and $(2, 0)$, respectively. In each cell, there are 5 randomly placed users, and each one has the maximum transmission power $\frac{P_{max}}{B\sigma^2} = 13$ dB. The number of available channels is $N_{ch} = 5$. We assume Rayleigh fading with path loss $\mathbb{E}\{z\} = s^{-4}$, where s represents the distance between the transmitter and the receiver. Each point on the curves is determined by taking the average over the results of 500 systems with randomly placed users, and the performance result of each system is evaluated over 5×10^4 time frames.

In Figs. 8.13 and 8.14, we study the influence of the interference control parameter γ, which is used in the pairwise interference constraints expressed in (8.47). The arrival rate at user i is set as $\lambda_i = \rho\mathbb{E}\{TB\log_2(1 + P_i z_i/B\sigma^2)\}$, where the parameter ρ is the arrival intensity. The target delay is 25 time frames for all users, and all

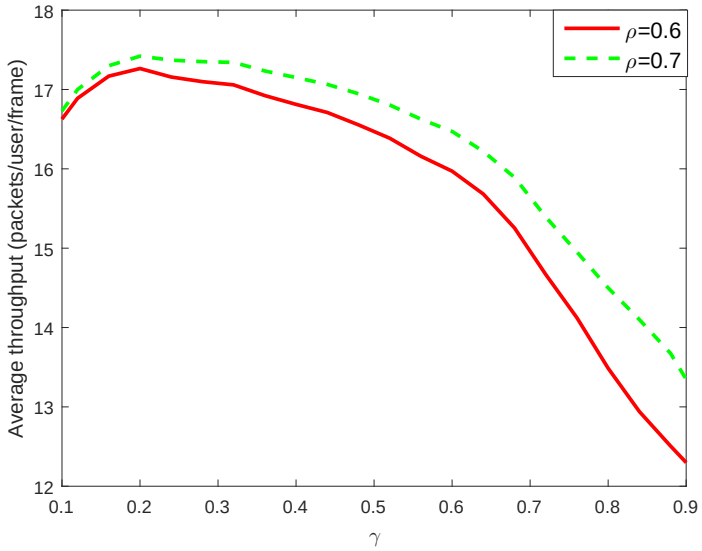

Figure 8.14: Throughput vs. interference control parameter γ

users transmit at their maximum power level. When γ is small, the ICI is not well controlled and the average transmission rate is not maximized. As γ increases, the system achieves lower delay violation probability and higher throughput due to better ICI management. However, when γ is too large, the interference constraints become too strict, which leads to less frequency reuse. In such cases, the throughput becomes smaller and the delay violation probability increases.

In Figs. 8.15 and 8.16, we analyze the influence of power control on our algorithm. In several conventional ICI control algorithms such as SFR, cell center users transmit with small power to reduce the interference they cause to the cell edge users. In these two figures, the transmission power of user i is selected as $P_i = P_{max}(s_i/R_{cell})^{\alpha}$, where s_i is the distance between the user and its corresponding RRH, and R_{cell} is the radius of the cell. As α increases, cell center users are restricted to transmit with smaller power. Also, all arrival rates are set as $\lambda = 1.5\mathbb{E}\{TB\log_2(1+P_{max}z_{edge}/B\sigma^2)\}$, where $\mathbb{E}\{TB\log_2(1+P_{max}z_{edge}/B\sigma^2)\}$ is the average transmission rate of a user at the edge of its associated cell. In Figs. 8.15 and 8.16, we notice that as α increases, both delay and throughput performances become worse. Our algorithm control the interference in the

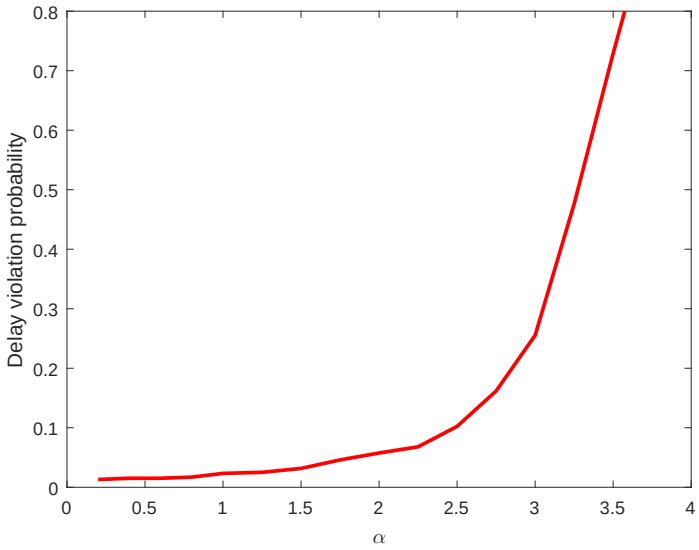

Figure 8.15: Delay violation probability vs. power control parameter α

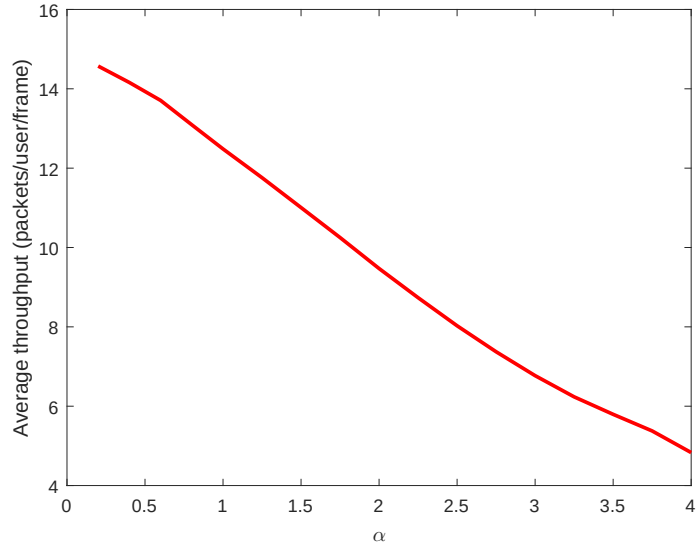

Figure 8.16: Throughput vs. power control parameter α

Table 8.7: Comparison between our algorithm and SFR

	Our Algorithm		SFR	
Arrival intensity ρ	Delay violation probability	Average delay (frame)	Delay violation probability	Average delay (frame)
0.6	0	1.25	0.0018	3.52
0.7	0	1.35	0.0024	8.99
0.8	0.0036	1.95	0.0235	59.04

user grouping step. Users that cannot satisfy the pairwise interference constraints are not allowed to reuse the same channel resource. Further decrease in the transmission power of the cell center users reduces their transmission rates, making it more difficult to stabilize the system.

Finally, we compare our algorithm with the conventional SFR scheme introduced in [72]. The arrival rates are set in the same way as in Figs. 8.13 and 8.14, and the target delay is 40 for all users. In our algorithm, all users transmit with maximum power. In the SFR scheme, users transmit with full power in the edge bands and they use 70% of their maximum power in the center bands. Channel assignment is conducted at the BBU of each cell individually to maximize the sum utility of the users in that cell. The results are provided in Table 8.7. As the arrival intensity increases, the advantage of our algorithm becomes obvious in terms of the average delay. With the C-RAN architecture, cooperative processing over multiple cells enhances the delay performance significantly.

Chapter 9

Conclusion

9.1 Summary

In this thesis, we have studied the delay QoS provisioning and optimal resource allocation for wireless networks. The contributions of this thesis are summarized below.

In Chapter 3, we have analyzed the throughput and energy efficiency of HARQ protocols in the presence of statistical queuing requirements when the QoS exponent θ is sufficiently small via Taylor expansion.

- In Section 3.1, we have investigated the throughput of HARQ-IR in the presence of queuing constraints imposed as limitations on buffer overflow probabilities. Using the statistical properties of the renewal counting process, we have identified the first-order expansion of the effective capacity of HARQ-IR in terms of the QoS exponent θ. We have taken into account hard deadline constraints by imposing an upper bound on the number of HARQ rounds to send a message. We have discussed that the main result on the first-order expansion of the effective capacity holds in the presence of deadline constraints with a modified description of the transmission time. Through numerical results, we have demonstrated that increasing the transmission rate R improves the throughput

monotonically in HARQ-IR and makes it approach the throughput of a system with perfect CSI at the transmitter, while it initially improves and then lowers the throughput of Type-I HARQ and HARQ Chase Combining protocols. We have also observed that increased throughput with larger R comes at the expense of longer transmission time or equivalently larger number of HARQ-IR rounds. We have shown that the throughput degrades when stricter queuing constraints or hard-deadline constraints are imposed. In particular, we have demonstrated that monotonic growth in the throughput with increasing R is not experienced in the presence of deadline limitations.

- In Section 3.2, we have analyzed the energy efficiency of the HARQ-CC scheme under outage, deadline, and statistical queuing constraints in the low-power and low-θ regimes by employing the notions of effective capacity and effective bandwidth from the stochastic network calculus while considering both constant-rate and random data arrivals to the buffer. Two queue models are considered. When outage happens, the transmitter discards the packet in the second queue model, while it transmits the same packet later in the first queue model. First, we have determined the minimum energy per bit and wideband slope achieved with HARQ-CC for fixed outage probability and both constant-rate and Markov source models. Also, we have provided comparisons among different random arrival models. Analyzing the results, we have concluded that source burstiness does not affect the minimum energy per bit when ON-OFF discrete time and Markov fluid sources are considered. On the other hand, due to the Poisson arrivals and the resulting higher level of burstiness, MMPS is shown to have worse energy efficiency compared to the ON-OFF Markov fluid source. Moreover, among the considered arrival models, MMPS is the only source for which the minimum energy per bit depends on the QoS exponent θ and grows with stricter QoS constraints. In contrast to the characterizations

regarding the minimum energy per bit, we have shown that wideband slope in all cases varies with the QoS exponent θ and source statistics. The impact of source burstiness is clearly identified with additional terms introduced in the denominators of the wideband slope expressions.

In Chapter 4, throughput of HARQ under statistical queuing constraints has been studied via the recurrence approach proposed in [15]. Compared with the low-θ approximation used in Chapter 3, recurrence approach is accurate for any QoS exponent value.

- In Section 4.1, we have analyzed the throughput of the HARQ-CC scheme under outage, deadline, and statistical queuing constraints by employing the notions of effective capacity and effective bandwidth from stochastic network calculus, while considering both constant-rate and random data arrivals to the buffer. Two typical queue models are considered. First, we have determined the throughput for the constant-rate arrival model. Then, with the effective capacity expression we obtained for the constant-rate arrival model, we have further formulated the throughput for the ON-OFF discrete-time and fluid Markov sources and MMPS. We have verified our analytical characterizations via Monte Carlo simulations. Finally, with the help of numerical results, we have compared the throughput values for the two considered queue models, and further investigated the impact of the deadline constraints, outage probability, queuing constraints and source randomness on the throughput.

- In Section 4.2, we have studied the throughput of HARQ-IR with finite block-length codes, deadline limits, and statistical queuing constraints by employing the notions of effective capacity and effective bandwidth from stochastic network calculus. Two different arrival models, namely the constant-rate and ON-OFF discrete time Markov arrivals, have been studied, and throughput characteriza-

tions have been obtained for both arrival models. We have first characterized the distribution of the duration of the transmission period and outage probability, and determined the effective capacity using the results from the recurrence relation approach. Subsequently, we obtained the throughput expressions for both constant-rate and ON-OFF discrete time Markov arrival models. Our characterizations have been verified via Monte Carlo simulations. Finally, we have further investigated the impact of the deadline constraints, fixed transmission rate, queuing constraints and blocklength via numerical results.

In Chapter 5, we have investigated the throughput of cooperative relay networks under statistical queuing constraints. Three types of cooperative relay networks are considered, namely two-hop relay channel, two-way relay channel and multi-source multi-destination relay network.

- In Section 5.1, we have investigated the throughput of the buffer-constrainted two-hop relay channel in the finite blocklength regime. We have initially characterized the system throughput through effective capacity analysis. Subsequently, we have formulated the throughput maximization problem, and investigated the properties of the optimal error probabilities for given time allocation parameter τ. Based on these properties, we have proposed an search algorithm in order to determine the optimal parameter setting more efficiently compared to directly searching in the three-dimensional bounded $\tau-\epsilon_1-\epsilon_2$ space. Finally, we have provided numerical results and investigated the impact of the source-relay distances, QoS exponents, and the blocklength on the throughput.

- In Section 5.2, we have investigated the throughput of two-way relaying under queueing constraints at both the source nodes and the relay. We have initially identified the instantaneous transmission/service rates in the multiple-access and broadcast phases of two-way relaying, and considered the stability condi-

tions and their impact on the system parameters (τ, ρ). Subsequently, we have defined the throughput region and provided characterizations of the maximum arrival rates, that can be supported by the sources, in terms of the QoS exponents and resource allocation parameters τ and ρ. We have provided numerical results and investigated the impact of the source-relay distances, signal-to-noise parameters, QoS exponents, and time-sharing between different decoding orders on the throughput.

- In Section 5.3, we have studied the throughput of multi-source multi-destination relay networks under statistical queueing constraints, for both cases of with and without CSI at the transmitter sides. When there is perfect CSI at the transmitter, transmission rates can be varied according to the instantaneous channel conditions. We have characterized the instantaneous channel capacities in different phases as functions of the system parameters τ, ρ and δ. When CSI is not available at the transmitter side, transmissions are performed at fixed rates, and decoding failures lead to retransmission requests via an ARQ protocol. We have modeled the links to be in ON or OFF states depending on the reliability of the reception. We have determined the probabilities of these states. Following these characterizations, we have described, for both perfect and no CSI cases, the stability conditions, and defined the feasible region of the transmission parameters. Finally, we have characterized the arrival rates under queueing constraints at the source and relay nodes as a function of the QoS exponents, channel fading and system parameters for both cases. In addition, the concavity of the throughput function is shown with respect to the system parameters δ and τ for the variable-rate scheme. We have verified the theoretical results via Monte Carlo simulations. Numerically, for the variable-rate model, we have investigated the optimal position of the relay node. Also, the throughput region is obtained via searching over the three dimensional parameter space.

In Chapter 6, we have investigated the mode selection between half-duplex and full-duplex modes in two-way MIMO systems operating under queuing constraints. To have a fair comparison, each antenna in full-duplex mode is assumed to transmit and receive at the same time so that it can have the same number of transmitting antennas as in half-duplex mode. We have characterized the system throughput for both half-duplex and full-duplex modes with given input covariance matrices. In the low-SNR regime, we have proposed an iterative algorithm to find the optimal input covariance matrices, which achieves the smallest $\frac{E_b}{N_0}_{\min}$ of the system. In the numerical results, we have found that full-duplex mode has better performance at low SNRs and short distances because two users can transmit simultaneously and make more efficient use of the resources. On the other hand, half-duplex mode has better performance in the high-SNR regime and at long distances.

In Chapter 7, we have studied the mode selection and resource allocation algorithms for D2D cellular networks.

- In Section 7.1, we have studied the mode selection and resource allocation in a TDM cellular network with one cellular user and one pair of D2D users operating under queueing constraints. For all four possible modes, namely the cellular mode, dedicated mode, uplink reuse mode, and downlink reuse mode, we have first formulated the system throughput using the effective capacity, and proposed efficient throughput maximization algorithms. Via numerical results, we have analyzed the influence of the positions of each node.

- In Section 7.2, we have proposed, for D2D cellular networks, a joint mode selection and channel allocation algorithm that maximizes the system throughput under statistical queueing limitations and average SINR constraints. First, we have characterized the instantaneous rate and effective capacity for all possible modes, namely the D2D cellular mode, D2D dedicated mode, uplink dedicated mode, downlink dedicated mode, uplink reuse mode, downlink reuse mode,

and D2D reuse mode. Then, we have proposed our channel matching algorithm, which selects modes for each user and allocates the channels simultaneously. Finally, we have further studied the performance of our algorithm by comparing its throughput and the numbers of unserved users with the random allocation algorithm through simulation results. In the results, we have demonstrated that our new algorithm can achieve higher throughput and serve more users.

- In Section 7.3, we have proposed a joint mode selection and resource allocation algorithm for D2D underlaid cellular networks. We have decomposed the problem into three subproblems, and designed algorithms for each subproblem. In the first step, we divide the transmission links into small groups using vertex coloring algorithm. In the second step, we solve the power optimization problem using the interior-point method for each group and conduct mode selection for those D2D links which form a group, and we assign channel resources in the final step. Via simulation results, we have compared the performance of our algorithm with that of the coalitional game method, and have shown that our algorithm achieves higher sum rate and serves more users with relatively small time consumption. Also, the influence of the interference threshold step size $\Delta\gamma$ is studied through numerical results, and the tradeoff between sum rate and the number of served users is identified.

In Chapter 8, we have studied the delay performance of content delivery over wireless cellular networks. Three types of network models are considered, including D2D caching network, D2D cellular network, and C-RAN.

- In Section 8.1, we have proposed a caching algorithm for D2D cellular networks, which minimizes the weighted average delay. First, we have characterized the popularity model and average transmission delay of a request. Then, we have formulated the delay minimization problem and developed our algorithm which

can solve the weighted average delay minimization problem efficiently. We have also extended our algorithm for a more general scenario, in which the distributions of fading coefficients and system parameters change over time. Finally, we have further investigated the performance of our algorithm by comparing it with a naive algorithm which simply caches the most popular files at each user. By applying both algorithms to two different popularity models, we have shown that our algorithm is more robust to variations in the popularity models, and can achieve better performance, because the proposed algorithm can more effectively take advantage of D2D communications. Also, the influence of the popularity parameter, caching size and number of users is studied via numerical results.

- In Section 8.2, we have proposed a scheduling algorithm for D2D cellular networks with deadline constraints. First, we have characterized the instantaneous packet rates for all possible modes, namely the D2D cellular mode, D2D dedicated mode, uplink dedicated mode, downlink dedicated mode, uplink reuse mode, downlink reuse mode, and D2D reuse mode. Then, we have formulated the cost function and defined utility for all possible decisions. Each scheduling decision can be regarded as a result of joint mode selection and channel allocation. For each type of decision, we have provided a power allocation algorithm to maximize the utility. Our algorithm allocates each channel according to the decision that provides the maximum utility value. Finally, we have further studied the performance of our algorithm through numerical results. In the results, we have provided characterizations on the optimal value of the system parameter α, and verified the advantages of having D2D users sharing the channel with other users.

- In Section 8.3, we have proposed an ICI-aware scheduling algorithm for the

C-RAN architecture that minimizes the sum delay cost of the system. The procedure is divided into two steps, namely the user grouping step and the channel matching step. In the user grouping step, we have designed a grouping algorithm that partitions all users in the network into small groups by checking their pairwise interference levels. In order to serve those users with high utility values with less channel resources, our grouping algorithm aims to collect users with high utility estimator values into small number of groups. In the channel matching step, we have formulated the channel assignment problem as a maximum-weight matching problem, which can be solved using the Hungarian algorithm. In the second step, user groups are matched to the available channel resources with goal of maximizing the system utility. Finally, we have studied the impact of the interference threshold and power control parameter via simulations, and compared our algorithm to the conventional SFR scheme. With the advantages of cooperative processing and information sharing over multiple cells, it has been verified that our algorithm designed for C-RAN can achieve higher throughput and lower delay.

9.2 Future Research Directions

9.2.1 Popularity Estimation and Scheduling for D2D Caching Systems

In Section 8.1, we assume that the base station has perfect knowledge about the popularity matrix. However, it is very difficult for the base station to predict the popularity of each content for each user. A reasonable popularity estimation and tracking algorithm should be designed, in order to apply our proposed caching algorithm. Also, it is of interest to combine the estimation and caching algorithms with the scheduling algorithm proposed in Section 8.2, to make our work more complete.

Appendix A

A.1 Proof of Theorem 1

The proof of Theorem 1 is based on the Taylor expansion of the cumulant generating function, which expresses C_e as a polynomial function of θ. After deriving the zeroth and first order coefficients of this polynomial, a closed-form expression for the first-order expansion is obtained for small θ. Before finding the polynomial approximation, we need to show that the moments of the random transmission time T are finite.

Let us denote the j^{th} moment of the random transmission time T by

$$\mu_j = \mathbb{E}\{T^j\}. \tag{A.1}$$

The following characterization shows that T has finite support for any fixed transmission rate and therefore we have $\mu_j < \infty$ for all $1 \leq j < \infty$.

Lemma 1 *If the expected value of the instantaneous capacity is strictly greater than zero, then for any fixed transmission rate R, the random transmission time T of HARQ-IR has finite support. Hence, all of its moments are finite.*

Proof 7 *Since $\{z_i\}$ is a sequence of i.i.d. random variables, by the strong law of large numbers [98, Section 7.4], we have $\frac{1}{n}\sum_{i=1}^{n} T_s B \log_2(1 + SNRz_i)$ converge to*

$\mathbb{E}\{T_s B \log_2(1 + SNRz_i)\} = \mathbb{E}\{C_i\}$ *almost surely, i.e., we have*

$$\Pr\left(\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} T_s B \log_2(1 + SNRz_i) = \mathbb{E}\{C_i\}\right) = 1. \tag{A.2}$$

This almost sure convergence implies that with probability one*, for any given* $\varepsilon > 0$*, there exists a positive integer* n_1 *such that for all* $n \geq n_1$

$$\left|\frac{1}{n}\sum_{i=1}^{n} T_s B \log_2(1 + SNRz_i) - \mathbb{E}\{C_i\}\right| \leq \varepsilon. \tag{A.3}$$

or equivalently

$$\mathbb{E}\{C_i\} - \varepsilon \leq \frac{1}{n}\sum_{i=1}^{n} T_s B \log_2(1 + SNRz_i) \leq \mathbb{E}\{C_i\} + \varepsilon. \tag{A.4}$$

Hence, under the assumption that $\mathbb{E}\{C_i\} > 0$*, we have the following lower bound with probability one for some* $0 < \varepsilon < \mathbb{E}\{C_i\}$*:*

$$\frac{1}{n}\sum_{i=1}^{n} T_s B \log_2(1 + SNRz_i) \geq \mathbb{E}\{C_i\} - \varepsilon > 0. \tag{A.5}$$

Next, we consider a bound on $\frac{R}{n}$*. For a fixed transmission rate* R*, we have*

$$\lim_{n\to\infty} \frac{R}{n} = 0. \tag{A.6}$$

Therefore, for any $\varepsilon_2 > 0$*, there exists an integer* $n_2 \geq n_1$ *such that for all* $n \geq n_2$*, we have*

$$\frac{R}{n} \leq \varepsilon_2. \tag{A.7}$$

Choosing $\varepsilon_2 = E\{C_i\} - \varepsilon$ *and using the bound in (A.5), we have for all* $n \geqslant n_2$ *that*

$$\frac{R}{n} \leq \frac{1}{n}\sum_{i=1}^{n} T_s B \log_2(1 + SNRz_i) \tag{A.8}$$

or equivalently

$$R \leq \sum_{i=1}^{n} T_s B \log_2(1 + SNR z_i) \tag{A.9}$$

with probability one for all $n \geq n_2$. According to the condition of successful decoding in equation (3.4), (A.8) implies that the random transmission time T for reliably sending R bits is upper bounded by n_2 with probability one, i.e., $\Pr(T \leq n_2) = 1$.

Hence, for any given fixed transmission rate R, T has finite support as claimed in the lemma. Hence, the moments $\mu_j = \mathbb{E}\{T^j\} \leq n_2^j < \infty$ are finite for all $1 \leq j < \infty$.

For HARQ Chase Combining, we can also show that all the moments of T are finite. Similar approach can be applied to chase combining protocol.

Lemma 2 *If the expected value of z_i is strictly greater than zero, then for any fixed transmission rate R, the random transmission time T of HARQ Chase Combining has finite support. Hence, all of its moments are finite.*

Proof 8 *By the strong law of large numbers [98, Section 7.4], we have $\frac{1}{n}\sum_{i=1}^{n} z_i$ converge to $\mathbb{E}\{z_i\}$ almost surely, i.e., we have*

$$\Pr\left(\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} z_i = \mathbb{E}\{z_i\}\right) = 1. \tag{A.10}$$

Similar to the proof of Lemma 1, for any given $\mathbb{E}\{z_i\} > \varepsilon > 0$, there exists a positive integer n_1 such that for all $n \geq n_1$, we can have

$$\frac{1}{n}\sum_{i=1}^{n} z_i \geq \mathbb{E}\{z_i\} - \varepsilon > 0. \tag{A.11}$$

For a fixed rate R, $\left(2^{\frac{R}{T_s B}} - 1\right)/SNR$ is also a constant. Then we have

$$\lim_{n\to\infty} \left(2^{\frac{R}{T_s B}} - 1\right)/(n\, SNR) = 0. \tag{A.12}$$

Therefore, for any $\varepsilon_2 > 0$, there exists an integer $n_2 \geq n_1$ such that for all $n \geq n_2$, we have

$$\left(2^{\frac{R}{T_s B}} - 1\right) / (n\ SNR) \leq \varepsilon_2. \tag{A.13}$$

Choosing $\varepsilon_2 = E\{z_i\} - \varepsilon$ and using the bound in (A.11), we have for all $n \geq n_2$ that

$$\left(2^{\frac{R}{T_s B}} - 1\right) / (n\ SNR) \leq \frac{1}{n} \sum_{i=1}^{n} z_i \tag{A.14}$$

or equivalently

$$R \leq T_s B \log_2(1 + SNR \sum_{i=1}^{n} z_i) \tag{A.15}$$

with probability one for all $n \geq n_2$. Similar as in the HARQ-IR case, we have shown that for HARQ Chase Combining the transmission time T also has finite support, and hence all the moments of T are finite.

Having shown the finiteness of all moments of T, the rest proof is the same for both HARQ-IR and HARQ Chase Combining. We next consider the cumulant generating function of N_t, which is the logarithm of the moment generating function of N_t, i.e.,

$$g(z) = \log \mathbb{E}\{\mathrm{e}^{zN_t}\}. \tag{A.16}$$

According to the theory of cumulant generating function, it can be expressed as

$$g(z) = \sum_{j=1}^{\infty} \kappa_j(t) \frac{z^j}{j!} \tag{A.17}$$

where $\kappa_j(t)$ is the j^{th} order cumulant of N_t. Examining (3.8), we notice that effective

capacity is proportional to the cumulant generating function of N_t and we can write

$$\frac{1}{\theta t}\log_e \mathbb{E}\{\mathrm{e}^{-\theta R N_t}\} = \frac{1}{\theta t}\sum_{j=1}^{\infty}\kappa_j(t)\frac{(-\theta R)^j}{j!} \tag{A.18}$$

$$= \sum_{j=1}^{\infty}\frac{\kappa_j(t)}{t}\,\frac{(-1)^j R^j}{j!}\,\theta^{j-1}. \tag{A.19}$$

Applying the theory of cumulant generating function to our problem, the effective capacity can be expressed as

$$C_e = -\lim_{t\to\infty}\frac{1}{\theta t}\log_e \mathbb{E}\{\mathrm{e}^{-\theta R N_t}\} \tag{A.20}$$

$$= -\lim_{t\to\infty}\sum_{j=1}^{\infty}\frac{\kappa_j(t)}{t}\,\frac{(-1)^j R^j}{j!}\,\theta^{j-1} \tag{A.21}$$

$$= \sum_{j=1}^{\infty}\left(\lim_{t\to\infty}\frac{\kappa_j(t)}{t}\right)\,\frac{(-1)^{j+1} R^j}{j!}\,\theta^{j-1}. \tag{A.22}$$

(by moving the limit inside the summation)

It has been proven in [99] that if the moments of T are finite, then the j^{th} cumulant of N_t can be written as

$$\kappa_j(t) = a_j t + b_j + o(1) \tag{A.23}$$

for some constants a_j and b_j which depend on the moments of T. From this result, we conclude that

$$\lim_{t\to\infty}\frac{\kappa_j(t)}{t} = a_j \tag{A.24}$$

and hence

$$C_e = \sum_{j=1}^{\infty} a_j \, \frac{(-1)^{j+1} R^j}{j!} \, \theta^{j-1}. \tag{A.25}$$

Furthermore, it has been shown in [99] and [100] that

$$a_1 = \frac{1}{\mu_1} \quad \text{and} \quad a_2 = \frac{\mu_2 - \mu_1^2}{\mu_1^3} = \frac{\sigma^2}{\mu_1^3} \tag{A.26}$$

where $\mu_1 = \mathbb{E}\{T\}$ and $\mu_2 = \mathbb{E}\{T^2\}$ are the first and second moments of T and σ^2 is the variance of T. Plugging in these values into (A.25), we readily obtain

$$C_e = \frac{R}{\mu_1} - \frac{R^2 \sigma^2}{2\mu_1^3} \theta + o(\theta) \tag{A.27}$$

where $o(\theta)$ denote the terms which decay faster than θ, i.e., $\lim_{\theta \to 0} \frac{o(\theta)}{\theta} = 0$. Hence, the desired characterization in Theorem 1 is proved.

A.2 Proof of Theorem 2

Proof 9 *For queue model I, the distribution of T_{Q1} is given by (3.32). Then, the expected value $\mathbb{E}\{T_{Q1}\} = \mu_{Q1}$ can be found as*

$$\mu_{Q1} = \sum_{t=1}^{\infty} t \ \Pr\{T_{Q1} = t\} \tag{A.28}$$

$$= \sum_{v=1}^{M} \sum_{k=0}^{\infty} (kM + v) \ \Pr\{T_{Q1} = kM + v\} \tag{A.29}$$

$$= \sum_{v=1}^{M} \left(\sum_{k=0}^{\infty} (kM + v)\varepsilon^k \ \Pr\{V = v\} \right) \tag{A.30}$$

$$= \sum_{v=1}^{M} \left(v \Pr\{V = v\} \sum_{k=0}^{\infty} \varepsilon^k + M \Pr\{V = v\} \sum_{k=0}^{\infty} k\varepsilon^k \right) \tag{A.31}$$

$$= \frac{1}{1-\varepsilon} \sum_{v=1}^{M} v \Pr\{V = v\} + \frac{M\varepsilon}{(1-\varepsilon)^2} \sum_{v=1}^{M} \Pr\{V = v\} \tag{A.32}$$

$$= \frac{1}{1-\varepsilon} \sum_{v=1}^{M} v \ \Pr\{V = v\} + \frac{M\varepsilon}{1-\varepsilon}. \tag{A.33}$$

Above, in (A.29), we replace t by $kM + v$ and sum over both k and v in order to more explicitly address possible violations of the maximum retransmission limit before successful packet transmission. Noting that $\sum_{k=0}^{\infty} \varepsilon^k = \frac{1}{1-\varepsilon}$ and $\sum_{k=0}^{\infty} k\varepsilon^k = \frac{\varepsilon}{(1-\varepsilon)^2}$, (A.31) can be simplified to (A.32). Notice that $\sum_{v=1}^{M} \Pr\{V = v\} = \Pr\{V \leqslant M\}$ represents the probability that the transmission has been completed before violating the deadline constraint M, and hence is equal to $1 - \varepsilon$. Applying this fact to (A.32), we obtain (3.39).

Similarly, the variance of T_{Q1} is given by

$$\sigma_{Q1}^2 = \mathbb{E}\{T_{Q1}^2\} - \mu_{Q1}^2 \tag{A.34}$$

where

$$\mathbb{E}\{T_{Q1}^2\} = \sum_{t=1}^{\infty} t^2 \ \Pr\{T_{Q1} = t\} \tag{A.35}$$

$$= \sum_{v=1}^{M} \left(\sum_{k=0}^{\infty} \varepsilon^k (kM+v)^2 \ \Pr\{V = v\} \right) \tag{A.36}$$

$$= \frac{1}{1-\varepsilon} \sum_{v=1}^{M} v^2 \ \Pr\{V = v\} + \frac{2M\varepsilon}{(1-\varepsilon)^2} \sum_{v=1}^{M} v \ \Pr\{V = v\} + \frac{M^2\varepsilon(1+\varepsilon)}{(1-\varepsilon)^2}. \tag{A.37}$$

Akin to the steps applied from (A.28) to (A.33), we again sum over $kM+v$ in (A.36), and then compute several summation terms with respect to k. Subsequently, using the fact that $\sum_{v=1}^{M} \Pr\{V = v\} = 1-\varepsilon$, we obtain (A.37).

Similarly, the distribution of T_{Q2} is given by (3.38) for queue model II, and we can directly obtain that

$$\mu_{Q2} = \sum_{t=1}^{M} t \ \Pr\{T_{Q2} = t\} = \sum_{t=1}^{M} t \ \Pr\{V = t\} + M\varepsilon, \tag{A.38}$$

and

$$\sigma_{Q2}^2 = \mathbb{E}\{T_{Q2}^2\} - \mu_{Q2}^2 \tag{A.39}$$

$$= \sum_{t=1}^{M} t^2 \ \Pr\{T_{Q2} = t\} - \mu_{Q2}^2 \tag{A.40}$$

$$= \sum_{t=1}^{M} t^2 \ \Pr\{V = t\} + M^2\varepsilon - \mu_{Q2}^2. \tag{A.41}$$

A.3 Proof of Theorem 3

Proof 10 *In order to derive the minimum energy per bit and wideband slope expressions, we need to obtain the first and second derivatives of the throughput $r_{avg}(SNR)$ with respect to SNR at zero SNR. For the constant-rate arrival model, r_{avg} or equivalently the effective capacity is given by (3.10). In this regard, for both queue models*

I and II, the first and second derivatives of $r_{avg}(SNR)$ *with respect to* SNR*, are given, respectively, by*

$$\dot{r}_{avg}(SNR) = \frac{F_M^{-1}(\varepsilon)}{(1+F_M^{-1}(\varepsilon)SNR)\mu\log_e 2} - \frac{F_M^{-1}(\varepsilon)\theta\sigma^2\left(\log_e(1+F_M^{-1}(\varepsilon)SNR)\right)^2}{(1+F_M^{-1}(\varepsilon)SNR)\mu^3(\log_e 2)^2}, \quad \text{(A.42)}$$

$$\begin{aligned}\ddot{r}_{avg}(SNR) =& \frac{\left(F_M^{-1}(\varepsilon)\right)^2\theta\sigma^2}{(1+F_M^{-1}(\varepsilon)SNR)^2\mu^3(\log_e 2)^2} - \frac{\left(F_M^{-1}(\varepsilon)\right)^2}{(1+F_M^{-1}(\varepsilon)SNR)^2\mu\log_e 2} \\ &+ \frac{\left(F_M^{-1}(\varepsilon)\right)^2\theta\sigma^2\log_e(1+F_M^{-1}(\varepsilon)SNR)}{(1+F_M^{-1}(\varepsilon)SNR)^2\mu^3(\log_e 2)^2}. \end{aligned} \quad \text{(A.43)}$$

For queue model I, the corresponding μ *and* σ *values are given by (3.39) and (3.40) respectively, and for queue model II, the corresponding* μ *and* σ *values are given by (3.41) and (3.42) respectively. Then, taking the limit as* $SNR \to 0$ *results in the following expressions:*

$$\dot{r}_{avg}(0) = \frac{F_M^{-1}(\varepsilon)}{\mu\log_e 2}, \quad \text{(A.44)}$$

and

$$\ddot{r}_{avg}(0) = -\frac{\left(F_M^{-1}(\varepsilon)\right)^2(\theta\sigma^2+\mu^2\log_e 2)}{\mu^3(\log_e 2)^2}. \quad \text{(A.45)}$$

Inserting the expressions in (A.44) and (A.45) into (3.26), (3.27), (3.28) and (3.29), the minimum bit energy and wideband slope for both queue models I and II are readily obtained.

A.4 Proof of Proposition 1

Proof 11 *Comparing (3.39) and (3.41), we obtain that*

$$\mu_{Q1} = \frac{\mu_{Q2}}{1-\varepsilon} \quad \text{(A.46)}$$

by replacing the summation index t with v in (3.41). Inserting (A.46) into (3.43) and comparing with (3.45), we obtain $\frac{E_b}{N_0}_{\min \text{ Q1}} = \frac{E_b}{N_0}_{\min \text{ Q2}}$.

Then, we rewrite (3.40) as

$$\sigma_{Q1}^2 = \frac{1}{1-\varepsilon}\sum_{v=1}^{M} v^2 \Pr\{V=v\} + \frac{2M\varepsilon}{(1-\varepsilon)^2}\sum_{v=1}^{M} v \Pr\{V=v\} + \frac{M^2\varepsilon(1+\varepsilon)}{(1-\varepsilon)^2} - \mu_{Q1}^2 \quad \text{(A.47)}$$

$$= \frac{1}{(1-\varepsilon)^2}\left(\sum_{v=1}^{M} v^2 \Pr\{V=v\} + M^2\varepsilon(1+\varepsilon) - \mu_{Q2}^2\right) + \left(\frac{2M\varepsilon}{(1-\varepsilon)^2}\sum_{v=1}^{M} v \Pr\{V=v\} - \frac{\varepsilon}{(1-\varepsilon)^2}\sum_{v=1}^{M} v^2 \Pr\{V=v\}\right) \quad \text{(A.48)}$$

$$= \frac{1}{(1-\varepsilon)^2}\sigma_{Q2}^2 + \left(\frac{2M\varepsilon}{(1-\varepsilon)^2}\sum_{v=1}^{M} v \ \Pr\{V=v\} - \frac{\varepsilon}{(1-\varepsilon)^2}\sum_{v=1}^{M} v^2 \ \Pr\{V=v\}\right). \quad \text{(A.49)}$$

From (A.47) to (A.48), we use the fact that $\frac{1}{1-\varepsilon} = \frac{1}{(1-\varepsilon)^2} - \frac{\varepsilon}{(1-\varepsilon)^2}$ *to break* $\frac{1}{1-\varepsilon}\sum_{v=1}^{M} v^2 \ \Pr\{V=v\}$ *into two terms. From (A.48) to (A.49), we apply the expression of* σ_{Q2}^2 *in (3.42). Also, for the last two terms in (A.49), we have*

$$\frac{2M\varepsilon}{(1-\varepsilon)^2}\sum_{v=1}^{M} v \Pr\{V=v\} - \frac{\varepsilon}{(1-\varepsilon)^2}\sum_{v=1}^{M} v^2 \Pr\{V=v\} \quad \text{(A.50)}$$

$$\geq \frac{2M\varepsilon}{(1-\varepsilon)^2}\sum_{v=1}^{M} v \Pr\{V=v\} - \frac{\varepsilon}{(1-\varepsilon)^2}\sum_{v=1}^{M} vM \Pr\{V=v\} \quad \text{(A.51)}$$

$$= \frac{M\varepsilon}{(1-\varepsilon)^2}\sum_{v=1}^{M} v \ \Pr\{V=v\} \quad \text{(A.52)}$$

$$\geq 0 \quad \text{(A.53)}$$

From (A.50) to (A.51), we use $v^2 \leq vM$*, because we only consider the summation of* v *from* 1 *to* M*. Using the above results, we get*

$$\sigma_{Q1}^2 \geq \frac{1}{(1-\varepsilon)^2}\sigma_{Q2}^2 \quad \text{(A.54)}$$

from (A.49). Applying (A.46) and (A.54) to (3.44) and (3.46), we conclude that $S_{0\,Q2} \geq S_{0\,Q1}$.

A.5 Proof of Theorem 4

Proof 12 *Plugging the effective capacity formulation obtained in Theorem 1 into (2.13), and then taking the derivative with respect to SNR and evaluating as* $SNR \to 0$*, we have*

$$\dot{r}(0) = \dot{C}_E(0) \Big/ \left[\frac{p_{22}}{2} + \frac{p_{22}(p_{11}+p_{22}) - 2(p_{11}+p_{22}-1)}{2(2-p_{11}-p_{22})} \right] \tag{A.55}$$

$$= \dot{C}_E(0)/P_{ON}. \tag{A.56}$$

In determining (A.55), we have used the fact that $\lim_{SNR\to 0} r(SNR) = 0$ *and* $\lim_{SNR\to 0} C_E(SNR) = 0$*. Note that when the transmit power approaches* 0*, the departure rate should also go to* 0*, which in turn makes the effective capacity approach* 0*. To satisfy the queuing constraints, the arrival rate* r *in the ON state should also diminish to* 0*. In the proof of Theorem 3, we have shown that* $\dot{C}_E(0) = \frac{F_M^{-1}(\varepsilon)}{\mu \log_e 2}$*. Therefore, we can have the first order derivative of the throughput evaluated as SNR goes to* 0 *as*

$$\dot{r}_{avg}(0) = \dot{r}(0)P_{ON} = \dot{C}_E(0) = \frac{F_M^{-1}(\varepsilon)}{\mu \log_e 2}. \tag{A.57}$$

Similarly, by taking the second order derivatives of the arrival rate r *with respect to SNR and evaluating as* $SNR \to 0$*, we obtain*

$$\ddot{r}(0) = \frac{\theta \dot{C}_E(0)^2 + \ddot{C}_E(0) - \dot{C}_E(0)^2\theta(\zeta+1)}{P_{ON}} \tag{A.58}$$

where ζ is defined in (3.49). In the proof of Theorem 3, we show that $\ddot{C}_E(0) = -\frac{F_M^{-1}(\varepsilon)^2(\theta\sigma^2+\mu^2\log_e 2)}{\mu^3(\log_e 2)^2}$. Therefore, we can find

$$\ddot{r}_{avg}(0) = \ddot{r}(0)P_{ON} = \frac{F_M^{-1}(\varepsilon)^2(-\theta\zeta - \frac{\theta\sigma^2+\mu^2\log_e 2}{\mu})}{(\mu\log_e 2)^2}. \tag{A.59}$$

Inserting the results in (A.57) and (A.59) into (3.26) and (3.27), and replacing μ and σ^2 by μ_{Q1} and σ_{Q1}^2 respectively, we get the desired results for queue model I in Theorem 4. Similarly, inserting the results in (A.57) and (A.59) into (3.28) and (3.29), and replacing μ and σ^2 by μ_{Q2} and σ_{Q2}^2 respectively, we obtain the desired results for queue model II.

A.6 Proof of Theorem 5

Proof 13 *The proof is similar to the proof of Theorem 4. Plugging (2.18) into (2.4), taking the first and second order derivatives and evaluating as SNR $\to 0$, we get*

$$\dot{r}(0) = \dot{C}_E(0)/P_{ON}. \tag{A.60}$$

From $\dot{r}(0)$, we get $\dot{r}_{avg}(0)$ as

$$\dot{r}_{avg}(0) = \dot{r}(0)P_{ON} = \dot{C}_E(0) = \frac{F_M^{-1}(\varepsilon)}{\mu\log_e 2}. \tag{A.61}$$

Furthermore, we have

$$\ddot{r}_{avg}(0) = \ddot{r}(0)\frac{\alpha}{\alpha+\beta} \tag{A.62}$$

$$= \ddot{C}_E(0) - \dot{C}_E^2(0)\theta\frac{2\beta}{\alpha(\alpha+\beta)} \tag{A.63}$$

$$= -\left(\frac{F_M^{-1}(\varepsilon)}{\mu\log_e 2}\right)^2\left(\frac{\theta\sigma^2+\mu^2\log_e 2}{\mu} + \frac{2\theta\beta}{\alpha(\alpha+\beta)}\right). \tag{A.64}$$

Inserting the results in (A.61) and (A.64) into (3.26) and (3.27), and replacing μ and σ^2 by μ_{Q1} and σ^2_{Q1} respectively, we obtain the desired results for queue model I in Theorem 5. Similarly, inserting the results in (A.61) and (A.64) into (3.28) and (3.29), and replacing μ and σ^2 by μ_{Q2} and σ^2_{Q2} respectively, we get the desired results for queue model II.

A.7 Proof of Theorem 6

Proof 14 *Using the characterization in (2.24), and taking the first and second order derivatives and evaluating as $SNR \to 0$, we get*

$$\dot{r}_{avg}(0) = P_{ON}\frac{\theta\alpha(\alpha+\beta)}{\alpha^2(e^\theta-1)}\dot{C}_E(0) = \frac{\theta}{e^\theta-1}\dot{C}_E(0), \tag{A.65}$$

and

$$\ddot{r}_{avg}(0) = \frac{\theta}{e^\theta-1}\ddot{C}_E(0) - \frac{2\beta\theta^2}{(\alpha+\beta)(e^\theta-1)}\dot{C}^2_E(0) \tag{A.66}$$

where $\dot{C}_E(0) = \frac{F_M^{-1}(\varepsilon)}{\mu\log_e 2}$ and $\ddot{C}_E(0) = -\frac{F_M^{-1}(\varepsilon)^2(\theta\sigma^2+\mu^2\log_e 2)}{\mu^3(\log_e 2)^2}$. Inserting the results in (A.65) and (A.66) into (3.26) and (3.27), and replacing μ and σ^2 by μ_{Q1} and σ^2_{Q1} respectively, we obtain the desired results for queue model I in Theorem 6. Similarly, inserting the results in (A.65) and (A.66) into (3.28) and (3.29), and replacing μ and σ^2 by μ_{Q2} and σ^2_{Q2} respectively, we get the desired results for queue model II.

A.8 Proof of Theorem 11

Proof 15 *In [17], the throughput of the half-duplex two-hop relay system under queuing constraints is given by*

$$\mathrm{R} = \begin{cases} \min\left\{ -\frac{1}{\theta_1}\Lambda_{\mathbf{S},\mathbf{R}}(-\theta_1), -\frac{1}{\theta_2}\Lambda_{\mathbf{R},\mathbf{D}}(-\theta_2)\right\} & \theta_2 \leq \theta_1 \\ \min\left\{ -\frac{1}{\theta_1}\Lambda_{\mathbf{S},\mathbf{R}}(-\theta_1), -\frac{1}{\theta_1}\Big(\Lambda_{\mathbf{R},\mathbf{D}}(-\theta_2) + \Lambda_{\mathbf{S},\mathbf{R}}(\theta_2-\theta_1)\Big)\right\} & \theta_2 > \theta_1 \end{cases} \tag{A.67}$$

when the stability condition is satisfied. In our finite blocklength regime, the instantaneous rate of the uplink is equal to 0 or $\tau m r_1$ bits per block with probabilities ϵ_1 and $1-\epsilon_1$, respectively. Therefore, we can write the LMGF of the $\mathbf{S}-\mathbf{R}$ link as

$$\begin{aligned} \Lambda_{\mathbf{S},\mathbf{R}}(\theta) &= \log \mathbb{E}_{z_1}\{\epsilon_1 e^{\tau\theta_1 m 0} + (1-\epsilon_1)e^{\tau\theta_1 m r_1}\} \\ &= \log \mathbb{E}_{z_1}\{\epsilon_1 + (1-\epsilon_1)e^{\tau\theta_1 m r_1}\}. \end{aligned} \tag{A.68}$$

Similarly, the LMGF of the $\mathbf{R}-\mathbf{D}$ link can be simplified as

$$\Lambda_{\mathbf{R},\mathbf{D}}(\theta) = \log \mathbb{E}\{\epsilon_2 + (1-\epsilon_2)e^{(1-\tau)\theta_2 m r_2}\}. \tag{A.69}$$

Plugging (A.68) and (A.69) into (A.67), and normalizing over the blocklength m (to change the unit from bits per block to bits per channel use), we obtain the throughput in the finite blocklength regime as in (5.7).

A.9 Proof of Theorem 15

Proof 16 *On the $\epsilon_1-\epsilon_2$ plane, for an arbitrary point $(\hat{\epsilon}_1,\hat{\epsilon}_2)$ inside the stability region, the line segment $(\hat{\epsilon}_1,\hat{\epsilon}_2)-(\epsilon_1^*,\epsilon_2^*)$ has an intersection point with the boundary of the stability region because $(\epsilon_1^*,\epsilon_2^*)$ given in Theorem 13 is assumed to be outside the stability region, and we denote the coordinates of this intersection point as*

$(\hat{\epsilon}_1^*, \hat{\epsilon}_2^*)$. *In [20], it was shown that as* ϵ *increases from* 0 *to* 1, $-\Lambda(-\theta)$ *first increases and then decreases after achieving its maximum value in the single-hop model. This property can be applied directly to* $-\Lambda_{\mathbf{S},\mathbf{R}}(-\theta_1)$ *and* $-\Lambda_{\mathbf{R},\mathbf{D}}(-\theta_2)$ *in our half-duplex two-hop system. Since* $\hat{\epsilon}_1^*$ *is between* $\hat{\epsilon}_1$ *and* ϵ_1^*, *and* $\hat{\epsilon}_2^*$ *is between* $\hat{\epsilon}_2$ *and* ϵ_2^*, *we have* $-\Lambda_{\mathbf{S},\mathbf{R}}(-\theta_1)\big|_{\hat{\epsilon}_1^*} \geq -\Lambda_{\mathbf{S},\mathbf{R}}(-\theta_1)\big|_{\hat{\epsilon}_1}$ *and* $-\Lambda_{\mathbf{R},\mathbf{D}}(-\theta_2)\big|_{\hat{\epsilon}_2^*} \geq -\Lambda_{\mathbf{R},\mathbf{D}}(-\theta_2)\big|_{\hat{\epsilon}_2}$. *In the proof of Theorem 13, we find that the throughput is a non-decreasing function of both* $-\Lambda_{\mathbf{S},\mathbf{R}}(-\theta_1)$ *and* $-\Lambda_{\mathbf{R},\mathbf{D}}(-\theta_2)$. *Therefore, the error probability pair* $(\hat{\epsilon}_1^*, \hat{\epsilon}_2^*)$ *gives the same or higher throughput, compared to* $(\hat{\epsilon}_1, \hat{\epsilon}_2)$. *This implies that for any error probability pair inside the stability region, there exists a point on the boundary of the stability region that achieves the same or higher throughput, when* $(\epsilon_1^*, \epsilon_2^*)$ *is outside the stability region. Therefore, the maximum throughput is achieved on the boundary of the stability region.*

A.10 Proof of Theorem 16

Proof 17 *We know that both* $\mathbf{S}_1 - \mathbf{D}_1$ *and* $\mathbf{S}_2 - \mathbf{D}_2$ *links are restricted by two queuing constraints, one at the corresponding source node, and the other one at the relay node. We consider these two constraints separately, and then combine the results. First, we only consider the constraints at the source nodes. According to (2.25), the maximum arrival rate that can be supported under queuing constraints at a source node is given by*

$$\mathrm{R}_j = -\frac{\Lambda_{\mathbf{S}_j,\mathbf{R}}(-\theta_j)}{\theta_j}, \tag{A.70}$$

for $j = 1, 2$. Similarly, when we only consider the queuing constraint at the relay node, the maximum arrival rates should satisfy

$$\mathrm{R}_j = \begin{cases} -\frac{1}{\theta_r}\Lambda_{\mathbf{R},\mathbf{D}_j}(-\theta_r) & \theta_r \leq \theta_j \\ -\frac{1}{\theta_j}\big(\Lambda_{\mathbf{R},\mathbf{D}_j}(-\theta_r) + \Lambda_{\mathbf{S}_j,\mathbf{R}}(\theta_r - \theta_j)\big) & \theta_r > \theta_j, \end{cases} \tag{A.71}$$

which is obtained from (2.26) and (2.27). Combining these results, the overall maximum arrival rates that can be supported by the system should be the minimum of (A.70) and (A.71), i.e.,

$$\mathrm{R}_j = \begin{cases} \min\Big\{ -\frac{1}{\theta_j}\Lambda_{\mathbf{S}_j,\mathbf{R}}(-\theta_j), -\frac{1}{\theta_r}\Lambda_{\mathbf{R},\mathbf{D}_j}(-\theta_r)\Big\} & \theta_r \leq \theta_j \\ \min\Big\{ -\frac{1}{\theta_j}\Lambda_{\mathbf{S}_j,\mathbf{R}}(-\theta_j), & \\ \quad -\frac{1}{\theta_j}\big(\Lambda_{\mathbf{R},\mathbf{D}_j}(-\theta_r) + \Lambda_{\mathbf{S}_j,\mathbf{R}}(\theta_r - \theta_j)\big)\Big\} & \theta_r > \theta_j, \end{cases} \tag{A.72}$$

for $j = 1, 2$. Using the definition of LMGF, (A.72) can be expressed in terms of the instantaneous rates, which is given by (5.44).

A.11 Proof of Theorem 17

Proof 18 *Depending on the relationship between θ_j and θ_r for $j = 1, 2$, there are two possible cases identified by (5.44).*

Case 1 : $\theta_r \leq \theta_j$.

In this case, the throughput R_j is given by

$$\mathrm{R}_j = \min\Big\{\mathrm{R}_{j,1}, \mathrm{R}_{j,2}\Big\} \tag{A.73}$$

where $\mathrm{R}_{j,1}$ *and* $\mathrm{R}_{j,2}$ *are defined as*

$$\begin{cases} \mathrm{R}_{j,1} = & -\frac{1}{\theta_j}\log\left(\mathbb{E}\left\{e^{-\theta_j\tau(\delta R_{\mathbf{S}_j,\mathbf{R}\{1,2\}}+(1-\delta)R_{\mathbf{S}_j,\mathbf{R}\{2,1\}})}\right\}\right), \\ \mathrm{R}_{j,2} = & -\frac{1}{\theta_r}\log\left(\mathbb{E}\left\{e^{-\theta_r(1-\tau)R_{\mathbf{R},\mathbf{D}_j}}\right\}\right). \end{cases} \tag{A.74}$$

By taking the second order derivative with respect to δ*, we can easily show the concavity of* $\mathrm{R}_{j,1}$ *and* $\mathrm{R}_{j,2}$*. The second order derivative of* $\mathrm{R}_{j,1}$ *is given by*

$$\begin{aligned} \frac{\partial^2 \mathrm{R}_{j,1}}{\partial\delta^2} =& -\frac{\theta_j\tau^2}{\left(\mathbb{E}\{e^{-\theta_j\tau(\delta R_{\mathbf{S}_j,\mathbf{R}\{1,2\}}+(1-\delta)R_{\mathbf{S}_j,\mathbf{R}\{2,1\}})}\}\right)^2} \\ &\times\Bigg\{\mathbb{E}\big\{(R_{\mathbf{S}_j,\mathbf{R}\{1,2\}}-R_{\mathbf{S}_j,\mathbf{R}\{2,1\}})^2 e^{-\theta_j\tau(\delta R_{\mathbf{S}_j,\mathbf{R}\{1,2\}}+(1-\delta)R_{\mathbf{S}_j,\mathbf{R}\{2,1\}})}\big\}\mathbb{E}\{e^{-\theta_j\tau(\delta R_{\mathbf{S}_j,\mathbf{R}\{1,2\}}+(1-\delta)R_{\mathbf{S}_j,\mathbf{R}\{2,1\}})}\} \\ &\quad-\left(\mathbb{E}\big\{(R_{\mathbf{S}_j,\mathbf{R}\{1,2\}}-R_{\mathbf{S}_j,\mathbf{R}\{2,1\}})e^{-\theta_j\tau(\delta R_{\mathbf{S}_j,\mathbf{R}\{1,2\}}+(1-\delta)R_{\mathbf{S}_j,\mathbf{R}\{2,1\}})}\big\}\right)^2\Bigg\}. \end{aligned} \tag{A.75}$$

According to the Cauchy-Schwarz inequality, two random variables U *and* V *should satisfy* $\mathbb{E}^2\{UV\}\leq\mathbb{E}\{U^2\}\mathbb{E}\{V^2\}$*. Assuming that*

$$U = e^{-\frac{1}{2}\theta_j\tau(\delta R_{\mathbf{S}_j,\mathbf{R}\{1,2\}}+(1-\delta)R_{\mathbf{S}_j,\mathbf{R}\{2,1\}})}, \tag{A.76}$$

and

$$V = (R_{\mathbf{S}_j,\mathbf{R}\{1,2\}}-R_{\mathbf{S}_j,\mathbf{R}\{2,1\}})U, \tag{A.77}$$

we can easily determine that the part inside the large curly brackets in (A.75) can be written as $\mathbb{E}\{V^2\}E\{U^2\}-\mathbb{E}^2\{UV\}$ *and hence is nonnegative. Then, we can readily determine that* $\frac{\partial^2\mathrm{R}_{j,1}}{\partial\delta^2}\leq 0$*, which indicates that* $\mathrm{R}_{j,1}$ *is a concave function of* δ*. From (A.74), we notice that the expression of* $\mathrm{R}_{j,2}$ *does not contain* δ*. In other words,* $\mathrm{R}_{j,2}$ *is a constant function in terms of* δ*, and* $\frac{\partial^2\mathrm{R}_{j,2}}{\partial\delta^2}=0$*. Hence, we can still regard* $\mathrm{R}_{j,2}$ *as a concave function of* δ*.*

Since the pointwise minimum of concave functions is concave [95], the concavity

of R_1 *and* R_2 *with respect to the time sharing parameter* δ *follows immediately when* $\theta_r \leq \theta_j$.

Case 2 : $\theta_r > \theta_j$.

In this case, the throughput R_j *is given by*

$$\mathrm{R}_j = \min\left\{\mathrm{R}_{j,1}, \mathrm{R}_{j,3}\right\} \tag{A.78}$$

where $\mathrm{R}_{j,3}$ *is defined as*

$$\mathrm{R}_{j,3} = -\frac{1}{\theta_j}\bigg(\log(\mathbb{E}\{e^{-\theta_r(1-\tau)R_{\mathbf{R},\mathbf{D}_j}}\}) + \log(\mathbb{E}\{e^{(\theta_r-\theta_j)\tau(\delta R_{\mathbf{S}_j,\mathbf{R}\{1,2\}}+(1-\delta)R_{\mathbf{S}_j,\mathbf{R}\{2,1\}})}\})\bigg). \tag{A.79}$$

We have already shown the concavity of $\mathrm{R}_{j,1}$ *in the previous case, and we can show the concavity of* $\mathrm{R}_{j,3}$ *following the same approach. The second order derivative of* $\mathrm{R}_{j,3}$ *is given by*

$$\begin{aligned}\frac{\partial^2 \mathrm{R}_{j,3}}{\partial \delta^2} =& -\frac{(\theta_r-\theta_j)^2\tau^2}{\theta_j\left(\mathbb{E}\{e^{(\theta_r-\theta_j)\tau(\delta R_{\mathbf{S}_j,\mathbf{R}\{1,2\}}+(1-\delta)R_{\mathbf{S}_j,\mathbf{R}\{2,1\}})}\}\right)^2} \\ &\times\Bigg\{\mathbb{E}\big\{(R_{\mathbf{S}_j,\mathbf{R}\{1,2\}}-R_{\mathbf{S}_j,\mathbf{R}\{2,1\}})^2 e^{(\theta_r-\theta_j)\tau(\delta R_{\mathbf{S}_j,\mathbf{R}\{1,2\}}+(1-\delta)R_{\mathbf{S}_j,\mathbf{R}\{2,1\}})}\big\}\mathbb{E}\{e^{(\theta_r-\theta_j)\tau(\delta R_{\mathbf{S}_j,\mathbf{R}\{1,2\}}+(1-\delta)R_{\mathbf{S}_j,\mathbf{R}\{2,1\}})}\} \\ &-\left(\mathbb{E}\big\{(R_{\mathbf{S}_j,\mathbf{R}\{1,2\}}-R_{\mathbf{S}_j,\mathbf{R}\{2,1\}})e^{(\theta_r-\theta_j)\tau(\delta R_{\mathbf{S}_j,\mathbf{R}\{1,2\}}+(1-\delta)R_{\mathbf{S}_j,\mathbf{R}\{2,1\}})}\big\}\right)^2\Bigg\}.\end{aligned} \tag{A.80}$$

Again using the Cauchy-Schwarz inequality, we have $\frac{\partial^2 \mathrm{R}_{j,3}}{\partial \delta^2} \leq 0$, *and the concavity follows. Since* R_j *is the pointwise minimum of* $\mathrm{R}_{j,1}$ *and* $\mathrm{R}_{j,3}$, R_j *is a concave function of* δ. *Now, we have shown in both cases that* R_1 *and* R_2 *are concave functions of* δ.

Finally, since the sum of two concave functions is also a concave function, the

sum rate is concave as well.

A.12 Proof of Theorem 18

Proof 19 *Similar to the proof of Theorem 17, Theorem 18 can be proved easily by evaluating the derivatives with respect to τ. The second order derivatives of $\mathrm{R}_{j,1}$, $\mathrm{R}_{j,2}$ and $\mathrm{R}_{j,3}$ with respect to τ are given, respectively, by*

$$\frac{\partial^2 \mathrm{R}_{j,1}}{\partial \tau^2} = - \frac{\theta_j}{\left(\mathbb{E}\{e^{-\theta_j \tau \mathbf{R}_{\mathbf{S}_j,\mathbf{R}}}\}\right)^2} \left\{ \mathbb{E}\{\mathbf{R}^2_{\mathbf{S}_j,\mathbf{R}} e^{-\theta_j \tau \mathbf{R}_{\mathbf{S}_j,\mathbf{R}}}\} \mathbb{E}\{e^{-\theta_j \tau \mathbf{R}_{\mathbf{S}_j,\mathbf{R}}}\} - \left(\mathbb{E}\{\mathbf{R}_{\mathbf{S}_j,\mathbf{R}} e^{-\theta_j \tau \mathbf{R}_{\mathbf{S}_j,\mathbf{R}}}\}\right)^2 \right\} \tag{A.81}$$

$$\begin{aligned} \frac{\partial^2 \mathrm{R}_{j,2}}{\partial \tau^2} = & - \frac{\theta_r}{\left(\mathbb{E}\{e^{-\theta_r (1-\tau) \mathbf{R}_{\mathbf{R},\mathbf{D}_j}}\}\right)^2} \\ & \times \left\{ \mathbb{E}\{\mathbf{R}^2_{\mathbf{R},\mathbf{D}_j} e^{-\theta_r (1-\tau) \mathbf{R}_{\mathbf{R},\mathbf{D}_j}}\} \mathbb{E}\{e^{-\theta_r (1-\tau) \mathbf{R}_{\mathbf{R},\mathbf{D}_j}}\} - \left(\mathbb{E}\{\mathbf{R}_{\mathbf{R},\mathbf{D}_j} e^{-\theta_r (1-\tau) \mathbf{R}_{\mathbf{R},\mathbf{D}_j}}\}\right)^2 \right\} \end{aligned} \tag{A.82}$$

$$\begin{aligned} \frac{\partial^2 \mathrm{R}_{j,3}}{\partial \tau^2} = & - \frac{\theta_r^2}{\left(\theta_j \mathbb{E}\{e^{-\theta_r (1-\tau) \mathbf{R}_{\mathbf{R},\mathbf{D}_j}}\}\right)^2} \\ & \times \left\{ \mathbb{E}\{\mathbf{R}^2_{\mathbf{R},\mathbf{D}_j} e^{-\theta_r (1-\tau) \mathbf{R}_{\mathbf{R},\mathbf{D}_j}}\} \mathbb{E}\{e^{-\theta_r (1-\tau) \mathbf{R}_{\mathbf{R},\mathbf{D}_j}}\} - \left(\mathbb{E}\{\mathbf{R}_{\mathbf{R},\mathbf{D}_j} e^{-\theta_r (1-\tau) \mathbf{R}_{\mathbf{R},\mathbf{D}_j}}\}\right)^2 \right\} \\ & - \frac{(\theta_r - \theta_j)^2}{\theta_j \left(\mathbb{E}\{e^{-\theta_j \tau \mathbf{R}_{\mathbf{S}_j,\mathbf{R}}}\}\right)^2} \left\{ \mathbb{E}\{\mathbf{R}^2_{\mathbf{S}_j,\mathbf{R}} e^{-\theta_j \tau \mathbf{R}_{\mathbf{S}_j,\mathbf{R}}}\} \mathbb{E}\{e^{-\theta_j \tau \mathbf{R}_{\mathbf{S}_j,\mathbf{R}}}\} - \left(\mathbb{E}\{\mathbf{R}_{\mathbf{S}_j,\mathbf{R}} e^{-\theta_j \tau \mathbf{R}_{\mathbf{S}_j,\mathbf{R}}}\}\right)^2 \right\}. \end{aligned} \tag{A.83}$$

Using the Cauchy-Schwarz inequality and concavity-preserving property of pointwise minimum, the concavity of R_1, R_2 and the sum rate follow readily.

A.13 Proof of Theorem 20

Proof 20 *We first consider user* $\mathbf{A}$ *and assume* $\mathbf{K}_{\mathbf{x}_B}$ *is given. From (6.19), the instantaneous rate of* $\mathbf{A}$ *can be written as*

$$r_A(SNR_A) = \frac{B}{\log_e 2} \sum_i \log_e \left(1 + N_B SNR_A \lambda_i(\mathbf{H}_{AB}\mathbf{K}_{\mathbf{x}_A}\mathbf{H}_{AB}^{\dagger}\mathbf{K}_{zB}^{-1})\right), \tag{A.84}$$

where $\lambda_i(\Delta)$ *denotes the* i^{th} *eigenvalue of the matrix* Δ*. Taking the derivative with respect to* SNR_A*, we get*

$$\dot{r}_A(SNR_A) = \frac{B}{\log_e 2} \sum_i \frac{N_B \lambda_i(\mathbf{H}_{AB}\mathbf{K}_{\mathbf{x}_A}\mathbf{H}_{AB}^{\dagger}\mathbf{K}_{zB}^{-1})}{1 + N_B SNR_A \lambda_i(\mathbf{H}_{AB}\mathbf{K}_{\mathbf{x}_A}\mathbf{H}_{AB}^{\dagger}\mathbf{K}_{zB}^{-1})}. \tag{A.85}$$

From (6.21), the first derivative of C_A *with respect to* SNR_A*, evaluated at* $SNR_A = 0$ *is given by*

$$\dot{C}_A(0) = \mathbb{E}\left\{\dot{r}_A(0)\right\} \qquad = \frac{BN_B}{\log_e 2}\mathbb{E}\left\{\mathrm{tr}(\mathbf{H}_{AB}\mathbf{K}_{\mathbf{x}_A}\mathbf{H}_{AB}^{\dagger}\mathbf{K}_{zB}^{-1})\right\} \tag{A.86}$$

where $\mathrm{tr}(\Delta)$ *denotes the trace of the matrix* Δ*. Note that both of* $\mathbf{K}_{\mathbf{x}_A}$ *and* $\mathbf{K}_{zB}^{-1}$ *are positive definite Hermitian matrices and therefore we can perform eigenvalue decomposition, and express them as* $\mathbf{K}_{\mathbf{x}_A} = \mathbf{V}_A\Lambda_A\mathbf{V}_A^{\dagger} = \sum_{i=1}^{N_A} \lambda_{A,i}\mathbf{v}_{A,i}\mathbf{v}_{A,i}^{\dagger}$ *and* $\mathbf{K}_{zB}^{-1} = \mathbf{V}_{zB}\Lambda_{zB}\mathbf{V}_{zB}^{\dagger}$*, where* $\mathbf{V}_A$ *and* $\mathbf{V}_{zB}$ *are unitary matrices,* Λ_A *and* Λ_{zB} *real diagonal matrices,* $\mathbf{v}_{A,i}$ *is the* i^{th} *column of* $\mathbf{V}_A$*, and* $\lambda_{A,i}$ *is the* i^{th} *eigenvalue of* $\mathbf{K}_{\mathbf{x}_A}$ *corresponding*

to the eigenvector $\mathbf{v}_{A,i}$*. Plugging this decomposition into (A.86), we get*

$$\begin{aligned}
\dot{C}_A(0) =& \frac{BN_B}{\log_e 2}\mathbb{E}\left\{\mathrm{tr}(\Lambda_{zB}^{1/2}\mathbf{V}_{zB}^{\dagger}\mathbf{H}_{AB}\mathbf{K}_{\mathbf{x}_A}\mathbf{H}_{AB}^{\dagger}\mathbf{V}_{zB}\Lambda_{zB}^{1/2})\right\} \\
=& \frac{BN_B}{\log_e 2}\sum_{i=1}^{N_A}\lambda_{A,i}\mathbb{E}\left\{\mathrm{tr}(\Lambda_{zB}^{1/2}\mathbf{V}_{zB}^{\dagger}\mathbf{H}_{AB}\mathbf{v}_{A,i}\mathbf{v}_{A,i}^{\dagger}\mathbf{H}_{AB}^{\dagger}\mathbf{V}_{zB}\Lambda_{zB}^{1/2})\right\} \\
=& \frac{BN_B}{\log_e 2}\sum_{i=1}^{N_A}\lambda_{A,i}\mathbb{E}\left\{\mathrm{tr}(\mathbf{v}_{A,i}^{\dagger}\mathbf{H}_{AB}^{\dagger}\mathbf{K}_{zB}^{-1}\mathbf{H}_{AB}\mathbf{v}_{A,i})\right\} \qquad \text{(A.87)} \\
\leq& \frac{BN_B}{\log_e 2}\mathbb{E}\left\{\lambda_{max}(\mathbf{H}_{AB}^{\dagger}\mathbf{K}_{zB}^{-1}\mathbf{H}_{AB})\right\} \qquad \text{(A.88)}
\end{aligned}$$

where $\lambda_{max}(\Delta)$ *denotes the maximum eigenvalue of matrix* Δ*. The equality is achieved when* $\mathbf{K}_{\mathbf{x}_A} = \mathbf{\Psi}_A\mathbf{\Psi}_A^{\dagger}$*, where* $\mathbf{\Psi}_A$ *is the eigenvector corresponding to the maximum eigenvalue of* $\mathbf{H}_{AB}^{\dagger}\mathbf{K}_{zB}^{-1}\mathbf{H}_{AB}$*. Following the same approach, we can shown a similar result for* $\mathbf{K}_{\mathbf{x}_B}$*, and hence prove the theorem.*

A.14 Proof of Theorem 21

Proof 21 *The second order derivative with respect to* τ_1 *is*

$$\frac{\partial^2\mathbf{R}_{sum}}{\partial\tau_1^2} = -\frac{\theta_C}{\mathbb{E}\{e^{-\tau_1\theta_C r_{C,B}}\}^2}\left[\mathbb{E}\{r_{C,B}^2\, e^{-\tau_1\theta_C r_{C,B}}\}\mathbb{E}\{e^{-\tau_1\theta_C r_{C,B}}\} - \mathbb{E}\{r_{C,B}\, e^{-\tau_1\theta_C r_{C,B}}\}^2\right]. \qquad \text{(A.89)}$$

Applying the Cauchy-Schwarz inequality, we can determine that $\frac{\partial^2\mathbf{R}_{sum}}{\partial\tau_1^2}$ *is negative. Through a similar approach, we can also determine that* $\frac{\partial^2\mathbf{R}_{sum}}{\partial\tau_2^2} \leq 0$ *and* $\frac{\partial^2\mathbf{R}_{sum}}{\partial\tau_3^2} \leq 0$*. Because* τ_1 *only appears in* $\mathbf{R}_C$*,* τ_2 *only appears in* $\mathbf{R}_B$*, and* τ_3 *only appears in* $\mathbf{R}_D$*, the Hessian matrix is diagonal, and can be expressed as*

$$\mathcal{H} = \boldsymbol{Diag}\left(\frac{\partial^2\mathbf{R}_{sum}}{\partial\tau_1^2}, \frac{\partial^2\mathbf{R}_{sum}}{\partial\tau_2^2}, \frac{\partial^2\mathbf{R}_{sum}}{\partial\tau_3^2}\right). \qquad \text{(A.90)}$$

It is readily noted that $\mathcal{H} \preceq 0$*, and the concavity is shown.*

Bibliography

[1] C.-S. Chang, "Stability, queue length, and delay of deterministic and stochastic queueing networks," *IEEE Trans. Automat. Contr.*, vol. 39, no. 5, pp. 913–931, 1994.

[2] C.-S. Chang and T. Zajic, "Effective bandwidths of departure processes from queues with time varying capacities," in *IEEE INFOCOM*, 1995, pp. 1001–1009 vol.3.

[3] C.-S. Chang and J. Thomas, "Effective bandwidth in high-speed digital networks," *IEEE J. Select. Areas Commun.*, vol. 13, no. 6, pp. 1091–1100, 1995.

[4] 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, 2003.

[5] S. Wicker, *Error Control Systems for Digital Communication and Storage*. Prentice Hall, 1995.

[6] S. B. Wicker, *Error Control Systems for Digital Communication and Storage*. Prentice Hall Englewood Cliffs, 1995, vol. 1.

[7] G. Caire and D. Tuninetti, "The throughput of hybrid-ARQ protocols for the Gaussian collision channel," *IEEE Trans. Inform. Theory*, vol. 47, no. 5, pp. 1971–1988, Jul. 2001.

[8] P. Wu and N. Jindal, "Performance of hybrid-ARQ in block-fading channels: A fixed outage probability analysis," *IEEE Trans. Commun.*, Apr. 2010.

[9] ——, "Performance of Hybrid-ARQ in block-fading channels: A fixed outage probability analysis," *IEEE Trans. Commun.*, vol. 58, no. 4, pp. 1129–1141, Apr. 2010.

[10] J. Choi, J. Ha, and H. Jeon, "On the energy delay tradeoff of HARQ-IR in wireless multiuser systems," *IEEE Trans. Commun.*, Aug. 2013.

[11] I. Stanojev, O. Simeone, Y. Bar-Ness, and D. H. Kim, "Energy efficiency of non-collaborative and collaborative Hybrid-ARQ protocols," *IEEE Trans. Wireless Commun.*, vol. 8, no. 1, pp. 326–335, Jan. 2009.

[12] F. Rosas, R. D. Souza, M. E. Pellenz, C. Oberli, G. Brante, M. Verhelst, and S. Pollin, "Optimizing the code rate of energy-constrained wireless communications with harq," *IEEE Trans. Wireless Commun.*, vol. 15, no. 1, pp. 191–205, Jan. 2016.

[13] J. Choi, J. Ha, and H. Jeon, "On the energy delay tradeoff of HARQ-IR in wireless multiuser systems," *IEEE Trans. Commun.*, vol. 61, no. 8, pp. 3518–3529, Aug. 2013.

[14] J. Choi and J. Ha, "On the energy efficiency of AMC and HARQ-IR with QoS constraints," *IEEE Trans. Veh. Technol.*, vol. 62, no. 7, pp. 3261–3270, Sept 2013.

[15] P. Larsson, J. Gross, H. Al-Zubaidy, L. K. Rasmussen, and M. Skoglund, "Effective capacity of retransmission schemes: A recurrence relation approach," *IEEE Trans. Commun.*, vol. 64, no. 11, pp. 4817–4835, Nov. 2016.

[16] J. Tang and X. Zhang, "Cross-layer resource allocation over wireless relay networks for quality of service provisioning," *IEEE J. Select. Areas Commun.*, vol. 25, no. 4, pp. 645–656, 2007.

[17] D. Qiao, M. Gursoy, and S. Velipasalar, "Effective capacity of two-hop wireless communication systems," *IEEE Trans. Inform. Theory*, vol. 59, no. 2, pp. 873–885, Sep. 2012.

[18] Y. Polyanskiy, H. Poor, and S. Verdu, "Channel coding rate in the finite blocklength regime," *IEEE Trans. Inform. Theory*, Apr. 2010.

[19] V. Tan and M. Tomamichel, "The third-order term in the normal approximation for the awgn channel," *IEEE Trans. Inform. Theory*, vol. 61, no. 5, pp. 2430–2438, Mar. 2015.

[20] M. C. Gursoy, "Throughput analysis of buffer-constrained wireless systems in the finite blocklength regime," *EURASIP Journal on Wireless Communications and Networking*, vol. 2013, no. 1, pp. 1–13, Dec. 2013.

[21] Y. Hu, J. Gross, and A. Schmeink, "On the performance advantage of relaying under the finite blocklength regime," *IEEE Commun. Lett.*, vol. 19, no. 5, pp. 779–782, May 2015.

[22] ——, "On the capacity of relaying with finite blocklength," *IEEE Trans. Veh. Technol.*, vol. 65, no. 3, pp. 1790–1794, Mar. 2016.

[23] J. Cui, "The capacity of Gaussian orthogonal multiple-access relay channel," *IEEE Commun. Lett.*, vol. 15, no. 4, pp. 365–367, Apr. 2011.

[24] M. Osmani-Bojd, A. Sahebalam, and G. Hodtani, "Capacity region of multiple access relay channels with orthogonal components," in *ICTC*, Sep. 2011, pp. 74–79.

[25] L. Sankar, Y. Liang, N. B. Mandayam, and H. Poor, "Fading multiple access relay channels: Achievable rates and opportunistic scheduling," *IEEE Trans. Inform. Theory*, vol. 57, no. 4, pp. 1911–1931, Apr. 2011.

[26] S. Salehkalaibar, L. Ghabeli, and M. Aref, "Achievable rate region for multiple-access-relay-networks," *IET Commun.*, vol. 4, no. 15, pp. 1792–1798, Oct. 2010.

[27] Y. Liu and A. Petropulu, "On the sum-rate of amplify-and-forward relay networks with multiple source-destination pairs," *IEEE Trans. Wireless Commun.*, vol. 10, no. 11, pp. 3732–3742, Nov. 2011.

[28] ——, "QoS guarantees in AF relay networks with multiple source-destination pairs in the presence of imperfect CSI," *IEEE Trans. Wireless Commun.*, vol. 12, no. 9, pp. 4225–4235, Sep. 2013.

[29] F. Chen, W. Su, S. Batalama, and J. Matyjas, "Joint power optimization for multi-source multi-destination relay networks," *IEEE Trans. Signal Processing*, vol. 59, no. 5, pp. 2370–2381, May 2011.

[30] C. Luo, S. McClean, G. Parr, P. Ren, and Y. Gong, "Multiple-source multiple-destination relay channels with network coding," *IET Communications*, vol. 7, no. 17, pp. 1958–1968, Nov. 2013.

[31] P. Parag and J. Chamberland, "Queueing analysis of a butterfly network for comparing network coding to classical routing," *IEEE Trans. Inform. Theory*, vol. 56, no. 4, pp. 1890–1908, Mar. 2010.

[32] D. Qiao, M. Gursoy, and S. Velipasalar, "On the achievable throughput region of multiple-access fading channels with QoS constraints," in *IEEE ICC*, May 2010, pp. 1–5.

[33] M. Ozmen and M. Gursoy, "Throughput regions of multiple-access fading channels with Markov arrivals and QoS constraints," *IEEE Wireless Commun. Lett.*, vol. 2, no. 5, pp. 499–502, Oct. 2013.

[34] D. Qiao, M. Gursoy, and S. Velipasalar, "Achievable throughput regions of fading broadcast and interference channels under QoS constraints," *IEEE Trans. Commun.*, vol. 61, no. 9, pp. 3730–3740, Sep. 2013.

[35] A. Sabharwal, P. Schniter, D. Guo, D. W. Bliss, S. Rangarajan, and R. Wichman, "In-band full-duplex wireless: Challenges and opportunities," *IEEE J. Select. Areas Commun.*, vol. 32, no. 9, pp. 1637–1652, Sep. 2014.

[36] T. Riihonen, S. Werner, and R. Wichman, "Mitigation of loopback self-interference in full-duplex mimo relays," *IEEE Trans. Signal Process*, vol. 59, no. 12, pp. 5983–5993, Dec. 2011.

[37] H. Ju, X. Shang, H. Poor, and D. Hong, "Bi-directional use of spatial resources and effects of spatial correlation," *IEEE Trans. Wireless Commun.*, vol. 10, no. 10, pp. 3368–3379, Oct. 2011.

[38] D. Kim, H. Ju, S. Park, and D. Hong, "Effects of channel estimation error on full-duplex two-way networks," *IEEE Trans. Veh. Technol.*, vol. 62, no. 9, pp. 4666–4672, Nov. 2013.

[39] M. Gursoy, "MIMO wireless communications under statistical queueing constraints," *IEEE Trans. Inform. Theory*, vol. 57, no. 9, pp. 5897–5917, Sept 2011.

[40] B. Kaufman and B. Aazhang, "Cellular networks with an overlaid device to device network," in *Asilomar Conference on Signals, Systems and Computers*, Oct. 2008, pp. 1537–1541.

[41] A. Asadi, Q. Wang, and V. Mancuso, "A survey on device-to-device communication in cellular networks," *IEEE Communications Surveys Tutorials*, vol. 16, no. 4, pp. 1801–1819, Fourthquarter 2014.

[42] K. Doppler, C.-H. Yu, C. Ribeiro, and P. Janis, "Mode selection for device-to-device communication underlaying an LTE-advanced network," in *IEEE Wireless Communications and Networking Conference (WCNC)*, Apr. 2010, pp. 1–6.

[43] D. Feng, L. Lu, Y. Yuan-Wu, G. Y. Li, G. Feng, and S. Li, "Device-to-device communications underlaying cellular networks," *IEEE Trans. Commun.*, vol. 61, no. 8, pp. 3541–3551, Jul. 2013.

[44] J. Han, Q. Cui, C. Yang, and X. Tao, "Bipartite matching approach to optimal resource allocation in device to device underlaying cellular network," *Electronics Letters*, vol. 50, no. 3, pp. 212–214, Jan. 2014.

[45] Y. Xu, R. Yin, T. Han, and G. Yu, "Dynamic resource allocation for device-to-device communication underlaying cellular networks," *International Journal of Communication Systems*, vol. 27, no. 10, pp. 2408–2425, Oct. 2014.

[46] G. Yu, L. Xu, D. Feng, R. Yin, G. Li, and Y. Jiang, "Joint mode selection and resource allocation for device-to-device communications," *IEEE Trans. Commun.*, vol. 62, no. 11, pp. 3814–3824, Nov. 2014.

[47] C. Xu, L. Song, Z. Han, D. Li, and B. Jiao, "Resource allocation using a reverse iterative combinatorial auction for device-to-device underlay cellular networks," in *2012 IEEE Global Communications Conference (GLOBECOM)*, Dec. 2012, pp. 4542–4547.

[48] Y. Li, D. Jin, J. Yuan, and Z. Han, "Coalitional games for resource allocation in the device-to-device uplink underlaying cellular networks," *IEEE Trans. Wireless Commun.*, vol. 13, no. 7, pp. 3965–3977, Jul. 2014.

[49] D. Tsolkas, E. Liotou, N. Passas, and L. Merakos, "A graph-coloring secondary resource allocation for D2D communications in LTE networks," in *IEEE 17th International Workshop on Computer Aided Modeling and Design of Communication Links and Networks (CAMAD)*, Sep. 2012, pp. 56–60.

[50] C. Lee, S. M. Oh, and J. S. Shin, "Resource allocation for device-to-device communications based on graph-coloring," in *International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS)*, Nov. 2015, pp. 451–455.

[51] M. Hajiaghayi, C. Wijting, C. Ribeiro, and M. T. Hajiaghayi, "Efficient and practical resource block allocation for LTE-based D2D network via graph coloring," *Wireless networks*, vol. 20, no. 4, pp. 611–624, 2014.

[52] M. Andrews, K. Kumaran, K. Ramanan, A. Stolyar, R. Vijayakumar, and P. Whiting, "Scheduling in a queuing system with asynchronously varying service rates," *Probability in the Engineering and Informational Sciences*, vol. 18, no. 02, pp. 191–217, Apr. 2004.

[53] M. Sharma and X. Lin, "OFDM downlink scheduling for delay-optimality: Many-channel many-source asymptotics with general arrival processes," in *Information Theory and Applications Workshop (ITA)*. IEEE, Feb. 2011, pp. 1–10.

[54] B. Ji, G. Gupta, X. Lin, and N. Shroff, "Low-complexity scheduling policies for achieving throughput and asymptotic delay optimality in multichannel wireless networks," *IEEE/ACM Trans. on Networks*, vol. 22, no. 6, pp. 1911–1924, Dec. 2014.

[55] J. A. Van Mieghem, "Dynamic scheduling with convex delay costs: The generalized $c\mu$ rule," *The Annals of Applied Probability*, pp. 809–833, Aug. 1995.

[56] ——, "Due-date scheduling: Asymptotic optimality of generalized longest queue and generalized largest delay rules," *Operations Research*, vol. 51, no. 1, pp. 113–122, Feb. 2003.

[57] "Cisco visual networking index: Global mobile data traffic forecast update 2012-2017." [Online]. Available: http://www.cisco.com/en/US/solutions/collateral/ns341/ns525/ns537/ns705/ns827/white_paper_c11-520862.html

[58] G. Paschos, E. Bastug, I. Land, G. Caire, and M. Debbah, "Wireless caching: Technical misconceptions and business barriers," *IEEE Commun. Mag.*, vol. 54, no. 8, pp. 16–22, Aug. 2016.

[59] E. Bastug, M. Bennis, and M. Debbah, "Living on the edge: The role of proactive caching in 5G wireless networks," *IEEE Commun. Mag.*, vol. 52, no. 8, pp. 82–89, 2014.

[60] Z. Zhao, M. Peng, Z. Ding, W. Wang, and H. V. Poor, "Cluster content caching: An energy-efficient approach to improve quality of service in cloud radio access networks," *IEEE J. Select. Areas Commun.*, vol. 34, no. 5, pp. 1207–1221, May 2016.

[61] W. Han, A. Liu, and V. K. N. Lau, "Phy-caching in 5G wireless networks: Design and analysis," *IEEE Commun. Mag.*, vol. 54, no. 8, pp. 30–36, Aug. 2016.

[62] M. Ji, A. M. Tulino, J. Llorca, and G. Caire, "On the average performance of caching and coded multicasting with random demands," in *International Symposium on Wireless Communications Systems (ISWCS)*. IEEE, 2014, pp. 922–926.

[63] M. A. Maddah-Ali and U. Niesen, "Coding for caching: Fundamental limits and practical challenges," *IEEE Commun. Mag.*, vol. 54, no. 8, pp. 23–29, Aug. 2016.

[64] M. Ji, G. Caire, and A. F. Molisch, "Wireless device-to-device caching networks: Basic principles and system performance," *IEEE J. Select. Areas Commun.*, vol. 34, no. 1, pp. 176–189, Jan. 2016.

[65] H. J. Kang, K. Y. Park, K. Cho, and C. G. Kang, "Mobile caching policies for device-to-device (D2D) content delivery networking," in *IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, Apr. 2014, pp. 299–304.

[66] N. Golrezaei, A. G. Dimakis, and A. F. Molisch, "Wireless device-to-device communications with distributed caching," in *IEEE International Symposium on Information Theory Proceedings (ISIT)*. IEEE, 2012, pp. 2781–2785.

[67] Y. Lin, L. Shao, Z. Zhu, Q. Wang, and R. K. Sabhikhi, "Wireless network cloud: Architecture and system requirements," *IBM Journal of Research and Development*, vol. 54, no. 1, pp. 4:1–4:12, Jan. 2010.

[68] C. Mobile, "C-RAN: The road towards green RAN," *White Paper, ver*, vol. 2, Oct. 2011.

[69] A. Checko, H. L. Christiansen, Y. Yan, L. Scolari, G. Kardaras, M. S. Berger, and L. Dittmann, "Cloud RAN for mobile networks - A technology overview," *IEEE Communications Surveys Tutorials*, vol. 17, no. 1, pp. 405–426, Firstquarter 2015.

[70] A. Simonsson, "Frequency reuse and intercell interference co-ordination in E-UTRA," in *IEEE 65th Vehicular Technology Conference (VTC) Spring*, Apr. 2007, pp. 3091–3095.

[71] M. C. Necker, "Local interference coordination in cellular OFDMA networks," in *IEEE 66th Vehicular Technology Conference (VTC) Fall*, Sep. 2007, pp. 1741–1746.

[72] Y. Xiang, J. Luo, and C. Hartmann, "Inter-cell interference mitigation through flexible resource reuse in OFDMA based communication networks," in *European wireless*, vol. 2007, Apr. 2007, pp. 1–7.

[73] S. Deb, P. Monogioudis, J. Miernik, and J. P. Seymour, "Algorithms for enhanced inter-cell interference coordination (eICIC) in LTE HetNets," *IEEE/ACM Trans. Networking*, vol. 22, no. 1, pp. 137–150, Feb. 2014.

[74] C.-S. Chang and T. Zajic, "Effective bandwidths of departure processes from queues with time varying capacities," in *INFOCOM '95. Fourteenth Annual Joint Conference of the IEEE Computer and Communications Societies. Bringing Information to People. Proceedings. IEEE*, Apr. 1995, pp. 1001–1009 vol.3.

[75] S. Tanwir and H. Perros, "A survey of VBR video traffic models," *IEEE Commun. Surveys & Tutorials*, vol. 15, no. 4, pp. 1778–1802, 2013.

[76] M. Ozmen and M. Gursoy, "Impact of channel and source variations on the energy efficiency under QoS constraints," in *IEEE International Symposium on Information Theory Proceedings (ISIT)*, Jul. 2012, pp. 806–810.

[77] M. Ozmen and M. C. Gursoy, "Wireless throughput and energy efficiency with random arrivals and statistical queuing constraints," *IEEE Trans. Inform. Theory*, vol. 62, no. 3, pp. 1375–1395, Mar. 2016.

[78] C. Chang, *Performance Guarantees in Communication Networks*, ser. Performance Guarantees in Communication Networks. Springer London, 2000. [Online]. Available: http://books.google.com/books?id=u-2IiZO3rlgC

[79] G. Kesidis, J. Walrand, and C.-S. Chang, "Effective bandwidths for multiclass Markov fluids and other ATM sources," *IEEE/ACM Trans. on Networking*, vol. 1, no. 4, pp. 424–428, Aug. 1993.

[80] G. Caire and D. Tuninetti, "The throughput of hybrid-ARQ protocols for the Gaussian collision channel," *IEEE Trans. Inform. Theory*, Jul. 2001.

[81] 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, Jul. 2003.

[82] R. Serfozo, *Basics of Applied Stochastic Processes*. Springer, 2009.

[83] G. Paschos, E. Bastug, I. Land, G. Caire, and M. Debbah, "Wireless caching: technical misconceptions and business barriers," *IEEE Commun. Mag.*, vol. 54, no. 8, pp. 16–22, Aug. 2016.

[84] S. Verdu, "Spectral efficiency in the wideband regime," *IEEE Trans. Inform. Theory*, vol. 48, no. 6, pp. 1319–1343, Sep. 2006. [Online]. Available: http://dx.doi.org/10.1109/TIT.2002.1003824

[85] Y. Polyanskiy, H. V. Poor, and S. Verdú, "Dispersion of Gaussian channels," in *IEEE International Symposium on Information Theory (ISIT)*. IEEE, Aug. 2009, pp. 2204–2208.

[86] Y. Li, G. Ozcan, M. C. Gursoy, and S. Velipasalar, "Energy efficiency of hybrid-ARQ under statistical queuing constraints," *IEEE Trans. Commun.*, vol. 64, no. 10, pp. 4253–4267, Oct. 2016.

[87] Y. Li, M. Gursoy, and S. Velipasalar, "On the throughput of Hybrid-ARQ under statistical queuing constraints," *IEEE Trans. Veh. Technol.*, vol. 64, no. 6, pp. 2725–2732, Jun. 2015.

[88] L. Li and A. J. Goldsmith, "Capacity and optimal resource allocation for fading broadcast channels-Part I: Ergodic capacity," *IEEE Trans. Inform. Theory*, vol. 47, no. 3, pp. 1083–1102, Mar. 2001.

[89] D. Qiao, M. Gursoy, and S. Velipasalar, "The impact of QoS constraints on the energy efficiency of fixed-rate wireless transmissions," *IEEE Trans. Wireless Commun.*, vol. 8, no. 12, pp. 5957–5969, Dec. 2009.

[90] M. Knox, "Single antenna full duplex communications using a common carrier," in *IEEE 13th Annual Wireless and Microwave Technology Conference (WAMICON)*, Apr. 2012, pp. 1–6.

[91] J. Papandriopoulos and J. Evans, "Low-complexity distributed algorithms for spectrum balancing in multi-user DSL networks," in *2006 IEEE International Conference on Communications (ICC)*, vol. 7, Jun. 2006, pp. 3270–3275.

[92] H. W. Kuhn, "The hungarian method for the assignment problem," *Naval research logistics quarterly*, vol. 2, no. 1-2, pp. 83–97, Mar. 1955.

[93] Y. Li, M. C. Gursoy, and S. Velipasalar, "Joint mode selection and resource allocation for D2D communications under queueing constraints," in *2016 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, Apr. 2016, pp. 490–495.

[94] D. J. Welsh and M. B. Powell, "An upper bound for the chromatic number of a graph and its application to timetabling problems," *The Computer Journal*, vol. 10, no. 1, pp. 85–86, Jan. 1967.

[95] S. P. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge university press, 2004.

[96] M. Akkouchi, "On the convolution of exponential distributions," *J. Chungcheong Math. Soc*, vol. 21, no. 4, pp. 501–510, 2008.

[97] Y. Li, M. C. Gursoy, and S. Velipasalar, "Scheduling in D2D underlaid cellular networks with deadline constraints," in *IEEE Vehicular Technology Conference (VTC) Fall*, Sep. 2016.

[98] G. Grimmett and D. Stirzaker, *Probability and Random Processes*. Oxford University Press, 2001.

[99] W. L. Smith, "On the cumulants of renewal processes," *Biometrika*, vol. 46, no. 1/2, pp. pp. 1–29, 1959. [Online]. Available: http://www.jstor.org/stable/2332804

[100] W. Feller, "Fluctuation theory of recurrent events," *Transactions of the American Mathematical Society*, vol. 67, no. 1, pp. pp. 98–119, 1949. [Online]. Available: http://www.jstor.org/stable/1990420

Vita

Yi Li received his B.S. degree in Electrical Engineering and Information Science in 2011 from University of Science and Technology of China, Hefei, China. He finished the Ph.D. degree in the Department of Electrical Engineering and Computer Science, Syracuse University, in 2017. During his Ph.D. study, he has 3 first author journal publications and 13 first author conference publications. His research interests are in the fields of wireless communications, cooperative communication systems and device-to-device cellular networks.