# ON DELAY-SENSITIVE COMMUNICATION OVER WIRELESS SYSTEMS

A Dissertation

by

LINGJIA LIU

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

DOCTOR OF PHILOSOPHY

May 2008

Major Subject: Electrical Engineering

# ON DELAY-SENSITIVE COMMUNICATION OVER WIRELESS SYSTEMS

A Dissertation

by

LINGJIA LIU

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

DOCTOR OF PHILOSOPHY

Approved by:

| | |
|---|---|
| Co-Chairs of Committee, | Jean-François Chamberland |
| | Scott Miller |
| Committee Members, | Costas Georghiades |
| | Wei Zhao |
| | Takis Zourntos |
| Head of Department, | Costas Georghiades |

May 2008

Major Subject: Electrical Engineering

# ABSTRACT

On Delay-Sensitive Communication Over Wireless Systems. (May 2008)

Lingjia Liu, B.S., Shanghai Jiao Tong University, China

Co–Chairs of Advisory Committee: Jean-François Chamberland
Scott Miller

This dissertation addresses some of the most important issues in delay-sensitive communication over wireless systems and networks. Traditionally, the design of communication networks adopts a layered framework where each layer serves as a "black box" abstraction for higher layers. However, in the context of wireless networks with delay-sensitive applications such as Voice over Internet Protocol (VoIP), on-line gaming, and video conferencing, this layered architecture does not offer a complete picture. For example, an information theoretic perspective on the physical layer typically ignores the bursty nature of practical sources and often overlooks the role of delay in service quality. The purpose of this dissertation is to take on a cross-disciplinary approach to derive new fundamental limits on the performance, in terms of capacity and delay, of wireless systems and to apply these limits to the design of practical wireless systems that support delay-sensitive applications. To realize this goal, we consider a number of objectives.

1. Develop an integrated methodology for the analysis of wireless systems that support delay-sensitive applications based, in part, on large deviation theory.

2. Use this methodology to identify fundamental performance limits and to design systems which allocate resources efficiently under stringent service requirements.

3. Analyze the performance of wireless communication networks that takes advan-

tage of novel paradigms such as user cooperation, and multi-antenna systems.

Based on the proposed framework, we find that delay constraints significantly influence how system resources should be allocated. Channel correlation has a major impact on the performance of wireless communication systems. Sophisticated power control based on the joint space of channel and buffer states are essential for delay-sensitive communications.

To My Family

# ACKNOWLEDGMENTS

The past few years have been one of the most enjoyable periods in my life. I owe my gratitude to all those people who have made this possible. Foremost, I would like to thank my advisor Prof. Jean-François Chamberland. I have been amazingly fortunate to have an advisor who gave me the freedom to explore on my own, and at the same time, the guidance to recover when my steps faltered. Prof. Chamberland has taught me a lot on how to question thoughts and express ideas. His brilliant insight, patient guidance, countless encouragement, and continuous support helped me overcome many crisis situations and finish this dissertation. Working with him has been truly a pleasant, stimulating and rewarding experience. I would like to thank my co-advisor Prof. Scott Miller. He has always been there to listen and give advice. I am deeply grateful to him for the long discussions that inspired my work in the area of user cooperation. I have also benefited greatly from Prof. Miller's courses and his expertise in the field of wireless communications. I would like to thank Prof. Costas Georghiades. He served on my thesis committee, and I have learned a lot from him on the subject of information theory and space-time communications. I wish to thank Prof. Wei Zhao for taking the time to assist me. His comments on the delay analysis of communication systems motivated me to relate the buffer overflow probability to the delay violation probability, which is very useful for delay-sensitive applications over wireless networks. I would also like to thank Prof. Takis Zourntos for serving on my dissertation committee and being supportive in different stages of my doctoral studies.

I am thankful to Prof. Krishna Narayannan, Prof. Henry Pfister, Prof. Tie Liu, and Prof. Natarajan Gautam. It has been an extremely rewarding experience taking their courses and discussing various research topics with them. I am grateful

to my friends and colleagues, Stark Draper, Andy Molisch, Philip Orlik, Jonathan Yedidia, and Jinyun Zhang, during my stay at Mitsubishi Electric Research Laboratories (MERL). I am also indebted to the members of the Wireless Communications Laboratory (WCL) with whom I have interacted during the course of my graduate studies. Particularly, I would like to acknowledge Yongzhe Xie, Guosen Yue, Wenyan He, Janath Peiris, Hari Sankar, Nitin Nangare, Angelos Liveris, Qiang Li, Jing Jiang, Anantharaman Balasaubaramaian, Weiyu Chen, Parimal Parag, Salim El Rouayheb, Markesh Pravin, Nirmal Gunaseelan, and Omar Ali for their various contributions.

Finally, this acknowledgement will not be complete without expressing my profound gratitude to my family. I sincerely thank my mother Yajing, my father Xiyan and my wife Yang for their unconditional love. I am what they made me to be. This dissertation is dedicated to them.

TABLE OF CONTENTS

CHAPTER Page

# LIST OF TABLES

# LIST OF FIGURES

FIGURE Page

FIGURE Page

# CHAPTER I

## INTRODUCTION

Recent years have been marked by a soaring demand for network access. This trend is exemplified by the constant growth of the Internet. The strong demand for network connectivity is fueled, partly, by new software applications, utility computing, and a widespread desire for real-time information access. To bridge the gap between mobile users and established communication infrastructures, wireless technology is being embraced with increasing vigor. Emerging applications for wireless communication and their corresponding service requirements are illustrated in Fig. 1.

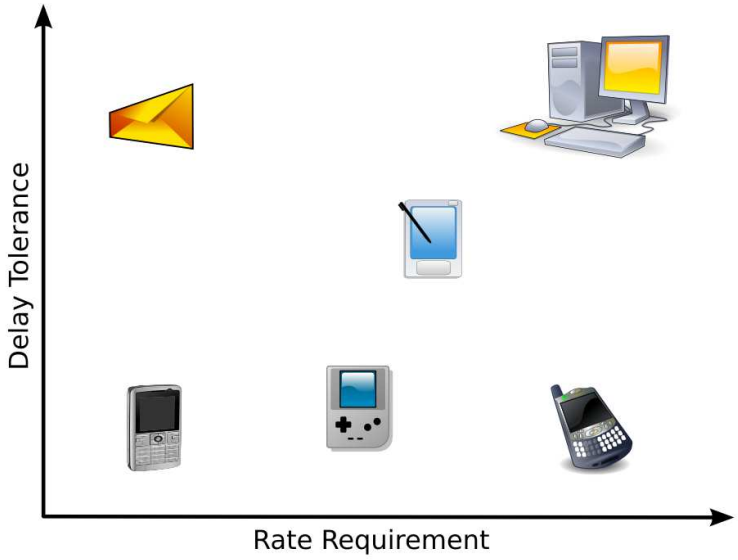


Fig. 1. Challenges in wireless communications.

It is clear that various applications have different service requirements. For example, data communication has a very high rate requirement, but it also has relatively

---

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

high delay tolerance. On the other hand, voice communication requires low communication rate while being very delay-sensitive. Rate and throughput are not the only performance measures; delay plays a very important role in user's satisfaction especially for real-time applications.

Future wireless communication networks will face the dual challenge of supporting large traffic volumes while providing reliable service to delay-sensitive applications such as VoIP, video conferencing, electronic commerce, and gaming. This is a difficult task, especially considering the unreliable nature of wireless environments. Traditionally, the design of wireless networks has borrowed heavily from the design paradigm of wired networks. Functionalities such as resource allocation, signal modulation, routing and error control coding are partitioned into separate network layers with minimum interaction between adjacent layers. Each layer serves as a "black box" abstraction for higher layers. For example, from the medium access control (MAC) and network layers point of view, the overall performance of the underlying physical and data link layers are modeled as "bit pipes" that deliver data at a fixed rate with occasional random errors. This layered framework has somehow caused the research community to split into two distinct groups, namely the networking community and the communication community. The main focus of the communication community is to build better bit pipes while the networking community has concentrated on how to best allocate these bit pipes. Information theory, which lies at the heart of the theory of communication, has played a central role in the design of wireless communication systems at the physical layer. In particular, most of the literature in this field focuses on maximizing the Shannon capacity [1], outage capacity, or spectral efficiency of wireless systems. These approaches give a foundation for improving throughput in wireless networks. However, the stringent service requirements typical of real-time traffic suggest that a classical capacity/throughput analysis alone does

not offer a complete assessment of service quality for the communication infrastructure associated with a wireless network. Real sources of information may be bursty and time-varying. Furthermore, wireless channels are prone to attenuation, fading, and interference. Variations in the sources and channels impair user satisfaction as they negatively impact packet loss probabilities, queue lengths, and delay distributions. For example, the user capacities of VoIP systems over wireless links are delay limited rather than throughput limited.

In many communication systems, power control and error-correcting codes have been employed to mitigate the effects of the channel fluctuations intrinsic to wireless communications. Yet, as the popularity of real-time applications grows, next generation wireless communication systems will have to pay much more attention to the role of delay as an important performance measure, as seen in Fig 1. Various real-time applications have hard delay constraints and various wireless systems are subject to statistical service requirement. For example, the current draft of the *IEEE 802.16m* (WiMax) standard specifies that the maximum delay bound for the voice communication is 50 ms. A user in this system is defined to have experienced voice outage if more than 2% of the VoIP packets are dropped, erased or not delivered successfully. New paradigms that better account for delay profiles and service quality are becoming highly desirable. Much work has been done in the networking community on the topics of scheduling and admission control to provide statistical service guarantees for various applications [2, 3, 4, 5, 6, 7]. Accordingly, the statistical service requirements of various delay-sensitive applications will also have a great impact in the physical layer system design. In particular, the end-to-end delay restrictions imposed on a communication system may preclude the use of error-correcting codes with long block-length, thereby forcing the system to operate away from its Shannon limit. It is clear that, with finite resources and under stringent service requirements,

the maximum throughput of a delay-sensitive system may be much lower than its Shannon capacity. Although several notable contributions to the subject [8, 9, 10, 11] have improved our understanding of delay-sensitive communications, the trade-off between throughput and service quality, and especially delay in wireless environment, is far from fully developed. While the physics governing the transmission of information over wireless channels has been carefully considered, fundamental parts of the relationship between the physical attributes of wireless channels and their impact on delay-sensitive communications remain largely unexplored.

## A. Integrated Framework

In this dissertation, we propose a cross-layer approach and study the interplay between the physical layer infrastructure and the queueing behavior of a wireless communication system. Due to the time-varying nature of wireless channels, it is often impractical or impossible to provide deterministic service guarantees for a specific link. There are two different performance criteria that are commonly considered: buffer overflow probability and delay violation probability. In most communication systems, explicit expressions for probabilities of buffer overflow and delay violation are hard to obtain. It is often easier and equally desirable to evaluate the probability that the system deviates significantly from its expected operating point. The *large deviation principle* (LDP) characterizes the asymptotic behavior, as $\epsilon \to 0$, of a collection of probability measures $\{\mu_\epsilon\}$ in terms of a rate function [12]. The LDP is closely related to the error exponent of random codes, it also forms a basis for Sanov's Theorem and the method of types. In this work, we adopt a statistical performance measure that captures the asymptotic decay-rate of buffer occupancy:

$$\theta = -\lim_{x\to\infty} \frac{\log \Pr\{L > x\}}{x} \tag{1.1}$$

where $L$ is the steady-state queue-length distribution of the buffer present at the transmitter. The parameter $\theta$ is also termed the LDP governing the buffer occupancy or the service exponent, and it reflects the perceived quality of a communication link. A larger $\theta$ represents a more reliable connection or a tighter service constraint. For a specific delay-sensitive application, the service requirement would be of the form $\theta \geq \theta_0$, where $\theta_0$ is the target decay-rate of buffer occupancy for the corresponding application. This performance metric is closely tied to the concept of effective bandwidth and network calculus, which has been studied extensively in the context of wired networks [13, 14, 15]. Given a stochastic arrival process, the effective bandwidth characterizes the minimum service rate required for the communication system to meet a certain service requirement $\theta_0$ [16, 17]. The literature on the effective bandwidth is rich. Comprehensive discussions on the subject and its applications are provided by Kelly [18] and Chang [14]. The decay-rate of (1.1) also forms a basis for the dual concept of effective capacity popularized by Wu and Negi [19, 20, 21]. Unlike wired connections where the service rates are typically constant, wireless channels are inherently unreliable and the associated service rates are time-varying. Assuming a constant flow of incoming data, the effective capacity characterizes the maximum constant arrival rate that a wireless system can support subject to a service requirement specified by $\theta_0$. When the target decay-rate $\theta_0$ approaches zero, the effective capacity converges to the Shannon capacity of the corresponding wireless channel.

The notions of effective bandwidth and effective capacity provide a useful platform to identify system limitations in terms of link-layer service requirements, such as buffer overflow probabilities and delay violation probabilities. Based on concepts from large deviation theory, this dissertation discusses the structure of an evaluation methodology suitable to characterize user dynamics and system limitations under service constraints. This framework leads to achievability results akin to Shannon capac-

ity, albeit in the context of delay-aware systems. Based on the proposed methodology, we are able to analyze the performance of various wireless communication schemes under stringent service constraints.

## B. Resource Allocation

System resources allocation in the context of wireless communication under service constraints $\theta_0$ is considered in [22, 23]. Through a simple example, the interplay between the characteristics of the physical layer, resource allocation, and system performance is exposed. The example presents the important aspects of data transmission over wireless channels using first-order models. Assuming that channel state information (CSI) is not available at the transmitter, a Markov model is introduced to capture the unreliable nature of wireless systems. For a given error correcting code, the behavior of the overall wireless connection is assumed equivalent to a continuous-time Markov chain. Coderate selection does affect system throughput. A higher coderate allows more information to be transmitted when the channel is ON, but it also reduces the probability of this event occurring. Conversely, a lower coderate increases the probability of the channel being ON, yet it decreases the rate at which information flows when the wireless link is ON. The throughput of a system can then be optimized by proper code selection.

As opposed to the throughput analysis, we also characterize the effective capacity of the corresponding Gilbert-Elliott channel as a function of the service constraint $\theta > 0$. The effective capacity is found to decay sharply as a function of $\theta$, which reveals a fundamental trade-off between delay and throughput. Furthermore, the optimal code selection for a wireless system depends on its service requirement. A more stringent constraint on $\theta$ lowers the optimal coderate $R$.

Correlation is also found to have a major impact on performance. The effective capacity of a slowly varying channel can be very small. For delay-sensitive applications over wireless systems, the popular assumption of independent and identically distributed channel realizations results in an over-optimistic assessment of system performance. The impact of correlation on system performance is perhaps best exemplified by the fact that constant arrival at a rate $a$ can only be supported if $\theta < \kappa/a$, where $\kappa$ is the exponential decay rate of the channel memory. That is, even with unlimited amount of physical resources, the maximum constant arrival rate supported under service constraint $\theta$ is bounded.

### C. User-Cooperation Networks

User cooperation is a relatively new concept in wireless communications, and it remains the focus of much research today. Sendonaris, Erkip, and Aazhang [24] demonstrated that cooperative schemes whereby many users pool their resources together to form a virtual antenna array can significantly enlarge the achievable rate region of the corresponding system. We subsequently showed that the achievable rate region of a multi-user system can still be enlarged, despite the lack of side information at the transmitters [25, 26]. Thus, cooperative schemes provide users with higher data-rates and more flexibility in choosing how to best share system resource.

In this dissertation, we seek to quantify the potential benefits of multi-user paradigms where mobile users cooperate to transmit information to an access point. For delay-aware systems, these gains originate from two interconnected aspects. First, user-cooperation augments the achievable rate region of a system and results in a higher sum throughput. Second, it allows the dynamic and fair distribution of system resources among users. The first-order fluid models described in the previous section

can be extended to wireless communications in the context of user-cooperation. In this case, users not only transmit data directly to an access point, but they may also elect to cooperate with one another.

The achievable rate region of a simple user-cooperation scheme is characterized and it is compared to the rate region of a non-cooperative system under different service constraints. Numerical results suggest that user-cooperation yields a large gain over traditional systems. Furthermore, this gain increases as the service requirement becomes more and more stringent. User-cooperation can therefore provide wireless users with the flexibility to better share system resources. Our queueing analysis also hints at the fact that overall performance depends heavily on the time-correlation of the underlying physical channel. In that sense, effective capacity is much more sensitive to higher-order statistics than, say, ergodic capacity or outage capacity.

### D. Multi-Antenna Systems

The recent adoption of wireless systems with multiple antennas has resulted in significant improvements in the capacity of point-to-point wireless links [27]. From an information theoretic point of view, the use of multiple antennas greatly increases the amount of diversity and the number of degrees of freedom in a wireless communication system. In the high signal-to-noise ratio (SNR) regime, the capacity of a Rayleigh channel with $m$ transmit antennas and $n$ receive antennas grows linearly with SNR [27, 28]

$$\text{Capacity}\,(\text{SNR}) = \min\{m, n\} \log \text{SNR} + \text{O}(1) \quad \text{as SNR} \ \to \infty.$$

In general, multi-antenna configurations can be used to increase data rates and to reduce the probability of symbol error at the receiver [29]. In the context of delay-

sensitive communications, the gains of multi-antenna configuration promise to be even more significant. The independence between antenna pairs, which can be met with enough separation within the receiving antennas and the transmitting antennas, decreases the instantaneous variations in the wireless channel. This, in turn, greatly improves the effective capacity of a particular system, as channel variations impairs system performance when operating under stringent service constraints. In this dissertation, we seek to identify the potential benefits of a multi-antenna configuration on the effective capacity of a wireless system. The expressions for the effective capacities of single-antenna systems, vector Gaussian channels, and multi-antenna Gaussian systems are found under a Rayleigh block fading model. The effective capacity of the single-antenna system is compared to those of the vector Gaussian systems in the low SNR regime. Our results suggest that there is a substantial gain in using multiple antennas at the transmitter or receiver for delay-sensitive communications. At low SNR, just as there is a power gain associated with using multiple receive antennas in terms of ergodic capacity [30], there is a statistical gain associated with using multiple transmit antennas in terms of effective capacity. For the general multi-antenna case, asymptotic upper and lower bounds for the effective capacity are derived. The lower bound indicates that the effective capacity of multiple-input-multiple-output (MIMO) systems scales linearly with the minimum number of transmit or receive antennas. An approximation for the effective capacity of the MIMO system is obtained in the low SNR regime when the number of transmit and/or receive antennas is large. Again, the effective capacity expression indicates that in the low SNR regime, a multi-antenna system offers a statistical gain as well as a power gain over a single-antenna system. This suggests that multi-antenna systems are especially suitable for delay-sensitive communication over wireless systems.

# CHAPTER II

## AN INTEGRATED FRAMEWORK

In this chapter, we introduce an integrated methodology for the analysis of wireless systems that support real-time traffic and delay-sensitive applications such as voice, video conferencing, and on-line gaming. A brief mathematical review of limit theorems and large deviations will be presented in Section A. In Section B, we analyze the transmission delay of a single-user wireless communication system. The steady-state delay violation probability is related to the steady-state buffer overflow probability by properly choosing the threshold of the buffer. This result allows us to analyze delay-sensitive applications over wireless systems by focusing on buffer overflow probability. The notions of effective bandwidth and effective capacity are introduced in Section C. We also discuss their corresponding operational meanings.

### A. Limit Theorems and Large Deviation Principles

In this section, we review concepts and theorems from large deviations. For a sequence of independent and identically distributed (iid) random variable $\{X_n, n \geq 1\}$, let $S_n = X_1 + \ldots + X_n$. It is known from the strong law of large numbers (LLN) that the empirical average converges to its mean almost surely, that is,

$$\frac{S_n}{n} = \frac{\sum_{i=1}^{n} X_i}{n} \rightarrow \mathrm{E}\left[\mathrm{X}_1\right].$$

We can further expand $S_n$ around its mean on the order of $\sqrt{n}$ through the central limit theorem. Let $\mathrm{Var}(\mathrm{X}_1)$ be the variance of $X_1$. Then

$$\lim_{n\rightarrow\infty} \Pr\left(\frac{\sum_{\mathrm{i}=1}^{\mathrm{n}}\left(\mathrm{X_i} - \mathrm{nE}\left[\mathrm{X_i}\right]\right)}{\sqrt{\mathrm{nVar}\left(\mathrm{X}_1\right)}} \leq \mathrm{x}\right) = \Phi\left(\mathrm{x}\right),$$

where

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt$$

is the cumulative distribution function (CDF) of the standard normal random variable. Intuitively, one may view the above result as a theorem for small deviation with the order $O(\sqrt{n})$ around the mean. On the other hand, the large deviation principle characterizes the probability of a large deviation with the order $O(n)$ from the mean.

One of the most important concept in the theory of large deviation is the Legendre transform. For a function $\Lambda(\theta) : \mathcal{R} \to \mathcal{R}$, the function

$$\Lambda^*(\lambda) = \sup_{\theta} [\theta\lambda - \Lambda(\theta)]$$

is called the Legendre transform of $\Lambda(\theta)$. Since it is easy to show that $\Lambda^*(\lambda)$ is convex, the Legendre transform sometimes is also called the convex transform. Again, consider the sequence $\{X_n, n \geq 1\}$ with the generic random variable $X$, and let

$$\Lambda(\theta) = \log \mathrm{E}\left[\mathrm{e}^{\theta \mathrm{X}}\right]$$

$$\Lambda^*(\lambda) = \sup_{\theta} [\theta\lambda - \Lambda(\theta)].$$

Assume $\mathrm{E}\left[\mathrm{e}^{\theta \mathrm{X}}\right] < \infty$ for all $\theta$, then Cramér's theorem holds. That is,

1. for every close set $F \subset \mathcal{R}$,

$$\limsup_{n\to\infty} \frac{\log \Pr(\mathrm{S_n}/\mathrm{n} \in \mathrm{F})}{n} \leq -\inf_{\lambda \in F} \Lambda^*(\lambda);$$

2. for every open set $G \subset \mathcal{R}$,

$$\liminf_{n\to\infty} \frac{\log \Pr(\mathrm{S_n}/\mathrm{n} \in \mathrm{G})}{n} \geq -\inf_{\lambda \in G} \Lambda^*(\lambda).$$

The function $\Lambda^*(\lambda)$ is called the rate function. For the special case where $X$ is a real

random variable with $\mathrm{E}[\mathrm{X}] = \bar{\mathrm{X}}$, for any $a > \bar{X}$, applying Cramér's theorem we have

$$-\inf_{\lambda>a} \Lambda^*(\lambda) \leq \lim_{n\to\infty} \frac{\log \Pr(\mathrm{S_n/n} \geq \mathrm{a})}{n} \leq -\inf_{\lambda\geq a} \Lambda^*(\lambda).$$

We say that the empirical average $S_n$ satisfies the LDP with rate function $\Lambda^*(\lambda)$.

Extensions of Cramér's theorem to sequences of not necessarily independent random variables are possible, as illustrated through the Gärtner-Ellis theorem. Consider a sequence of random variables $\{Y_n, n \geq 1\}$. Let

$$\Lambda_n(\theta) = \frac{\log \mathrm{E}\left[\mathrm{e}^{\theta \mathrm{Y_n}}\right]}{n}.$$

Assume

$$\lim_{n\to\infty} \Lambda_n(\theta) = \Lambda(\theta) < \infty \quad \forall \theta \in \mathcal{R},$$

and $\Lambda(\theta)$ is differentiable for all $\theta \in \mathcal{R}$. Let

$$\Lambda^*(\lambda) = \sup_{\theta} [\theta\lambda - \Lambda(\theta)].$$

Then the Gärtner-Ellis theorem holds, i.e.,

1. for every close set $F \subset \mathcal{R}$,

$$\limsup_{n\to\infty} \frac{\log \Pr(\mathrm{Y_n/n} \in \mathrm{F})}{n} \leq -\inf_{\lambda\in F} \Lambda^*(\lambda);$$

2. for every open set $G \subset \mathcal{R}$,

$$\liminf_{n\to\infty} \frac{\log \Pr(\mathrm{Y_n/n} \in \mathrm{G})}{n} \geq -\inf_{\lambda\in G} \Lambda^*(\lambda).$$

Essentially, $\Lambda(\theta)$ plays the role of the logarithmic moment generating function in the iiḋcase.

## B. Statistical Service Guarantee

As mentioned in the previous section, real-time applications such as video conferencing and voice signals require bounded delay service. Once the delay requirement is violated, the data or the packet is discarded. For wireless systems, using deterministic delay bounds is often impractical due to the fading nature of wireless channels. As such, there are two important statistical performance measures for a communication system: probability of buffer overflow and delay violation probability.

Consider a single-user wireless communication system. In this situation, data

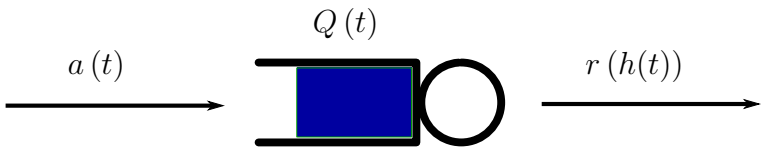


Fig. 2. A wireless queueing system model.

arrives and is placed into a transmission buffer before it gets transmitted. Periodically, the transmitter removes some of the data from the buffer, encodes it and transmits the encoded data over a fading channel. After sufficient information is received, the data is decoded at the receiver. The delay experienced by the data in the system is due to two major components: queueing delay and coding delay. Queueing delay is simply the time data spends in the buffer while the coding delay is the time from when data is encoded until it is decoded. In general, coding delay is related to the codeword length, interleaving, and probability of decoding failure [31, 32, 33]. For most of the practical wireless communication systems, the coding delay is on the order of the channel coherence time which is typically much smaller than the relevant time-scales of the queueing delay. Therefore, in this dissertation we restrict our attention to the queueing delay of a communication system. The joint analysis of these two

components can be regarded as a future research topic listed in Chapter VI.

At time $t$, let $a(t)$ denote the instantaneous arrival rate and $r(h(t))$ be the instantaneous service rate where $h(t)$ is the channel state. $Q(t)$ is the queue length of the buffer and $D(t)$ is the delay experienced by the packet which is about to get serviced. Given a specified delay bound $D_{\max}$, a service constraint may require the delay violation probability to be no greater than a certain threshold $\varepsilon$. That is,

$$\Pr\{D > D_{\max}\} \leq \varepsilon,$$

where $D$ is the steady-state delay experienced by a packet.

For the special case when the service rate is constant, $r(h(t)) = r$, the delay at any time is a constant multiple of the queue length. This is because when a packet arrives at the buffer, we know exactly how long the packet is going to stay in the buffer by looking at the queue length. Therefore, any delay performance criterion can be translated into an equivalent buffer overflow probability requirement.

Consider another special case when the arrival rate is constant, $a(t) = a$. This case is much more complicated than the case when the service rate is constant. Due to the time-varying nature of the wireless channel, we cannot figure out how long a packet will stay in the buffer by looking at the queue length of the buffer when the packet arrives. However, since $D(t)$ denotes the delay experienced by a packet which is about to get serviced at time $t$, $D(t)$ can be related to $Q(t)$ through $Q(t) = aD(t)$. When the communication system is stable, it can be shown (see Appendix A) that

$$\Pr\{D > D_{\max}\} \leq c \cdot \sqrt{\Pr\{Q > Q_{\max}\}},$$

where $c$ is some positive constant, $Q$ is the steady-state queue length of the buffer, and $Q_{\max} = aD_{\max}$. Therefore, the service constraint for the communication system

can be redefined as

$$\Pr\{Q > Q_{\max}\} \leq \varepsilon^2/c^2.$$

The above equation indicates that the delay performance criterion can be upper-bounded by a buffer overflow probability requirement.

Now let us consider a general communication system with a random arrival process $a(t)$ and a stochastic service process $r(h(t))$. Compare this single queueing system with a system that contains two queues. The arrival rate in the first queue is again $a(t)$ and the service offered to the second queue is $r(h(t))$. Moreover, the first queue is serviced at a constant rate $v$ whenever it is non-empty, and the departed packets from the first queue are immediately placed in the second queue. This is illustrated in Figure 3. Because of the additional constraint present in the second scenario, the queue length in the first system is always less than or equal to the sum of the queues in the latter system. The same result can also be applied to delay. Now compare

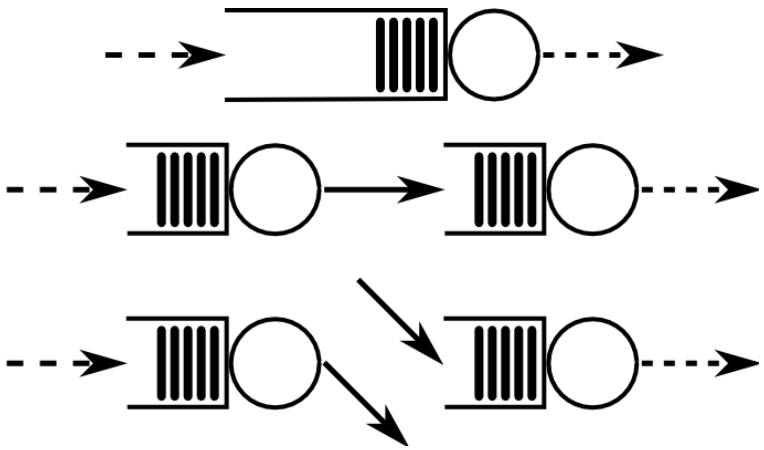

Fig. 3. Illustration of the three systems introduced in the buffer decoupling argument.

the second system with a network composed of two independent queues. The arrival process in the first queue is $a(t)$, and this queue is served at a constant rate $v$ when it

is non-empty. Packets arrive in the second queue at a constant rate $v$, and they are served at a rate $r(h(t))$. Note that the length of the first queue in the third system is always equal to the length of the first queue in the second system. Furthermore, the length of the second queue in the second system is always less than or equal to the length of the second queue in the third system. Clearly, both of the queues in the third system can be analyzed using special cases where either the arrival rate or the service rate is constant. Therefore, we can work directly on the buffer overflow probability to guarantee delay violation probability requirement imposed on the system. Throughout this dissertation we will use requirements on buffer overflow probability to indicate the delay-sensitivity of various applications. Defining the performance metric in terms of queue length rather than delay leads to a much simpler characterization for service constraints. Yet it is instructive to emphasize the relation between the two approaches.

One of the key results in the theory of statistical quality of service (QoS) guarantees [17, page 290] is that for a queueing system with stationary ergodic arrival and service processes, and under sufficient conditions, the queue length process $Q(t)$ converges in distribution to a random variable $Q$ that satisfies

$$-\lim_{Q_{\max}\to\infty} \frac{\log \Pr\left\{Q \geq Q_{\max}\right\}}{Q_{\max}} = \theta,$$

where $\theta$ is called the asymptotic decay-rate of buffer occupancy or, alternatively, the LDP governing buffer overflow. The above result indicates that the probability of the steady-state queue-length exceeding a certain threshold $Q_{\max}$ decays exponentially fast as $Q_{\max}$ increases, i.e.,

$$\Pr\{Q \geq Q_{\max}\} \approx e^{-\theta Q_{\max}}.$$

For a communication system where the value of $Q_{\max}$ is typically large, the above approximation of the buffer overflow probability is very accurate. Therefore, the queueing behavior of a wireless system can be specified by the pair $\{Q_{\max}, \theta\}$. Note that the parameter $\theta$ plays a critical role in meeting the service requirement. A larger $\theta$ leads to a faster decay-rate of the buffer overflow probability and, hence, the delay violation probability. For the special case where $\theta$ approaches zero, the communication system can tolerate arbitrarily long delay. On the other hand, when $\theta$ tends to infinity, the communication system can tolerate no instantaneous delay. In this scenario, the throughput of the system cannot be larger than the minimum instantaneous service rate of the channel.

## C. Effective Bandwidth and Effective Capacity

The performance measure $\theta$ is closely related to the concepts of effective bandwidth and effective capacity. Effective bandwidth is introduced to analyze the system where the service rate is constant. Consider the system in Fig. 2. Assume the service rate is constant with $r(h(t)) = r$. It is clear that, given a random arrival process, the probability of buffer overflow will increase and hence the LDP governing the buffer occupancy will decrease as $r$ decreases. Therefore, for a particular service requirement specified by

$$-\lim_{Q_{\max}\to\infty} \frac{\log \Pr\left\{Q \geq Q_{\max}\right\}}{Q_{\max}} \geq \theta, \tag{2.1}$$

effective bandwidth answers the question: how large does $r$ need to be to satisfy the service constraint. Mathematically, the effective bandwidth can be expressed as

$$\beta(\theta) = \inf\left\{r : -\lim_{Q_{\max}\to\infty} \frac{\log \Pr\{Q > Q_{\max}\}}{Q_{\max}} \geq \theta\right\}.$$

Similarly, we can consider the special case where the arrival rate is constant with

$a(t) = a$. Given a stochastic service process $r(h(t))$, the probability of buffer overflow will increase and hence the LDP governing the buffer occupancy will decrease as $a$ increases. So the effective capacity answers the question: how large can $a$ be to meet the service constraint specified in (2.1). Mathematically, the effective capacity can be written as

$$\alpha(\theta) = \sup\left\{a : -\lim_{Q_{\max}\to\infty} \frac{\log \Pr\{Q > Q_{\max}\}}{Q_{\max}} \geq \theta\right\}.$$

Alternatively, the effective capacity can be formally defined as follows. Let $r\,(t)$ be the instantaneous service rate of the wireless channel at time $t$. Let $\tilde{S}\,(t) = \int_0^t r\,(\tau)\,d\tau$ be the service offered by the wireless channel in the interval from 0 to $t$. Suppose that the service process is stationary and the Gärtner-Ellis limit of $\tilde{S}\,(t)$

$$\Lambda\,(-\theta) = \lim_{t\to\infty} \frac{1}{t} \log E\left[e^{-\theta \int_0^t r(\tau)d\tau}\right] = \lim_{t\to\infty} \frac{1}{t} \log E\left[e^{-\theta \tilde{S}(t)}\right]$$

exists and is differentiable for all $\theta > 0$. Then the effective capacity of the service process is defined by

$$\alpha\,(\theta) = \frac{-\Lambda\,(-\theta)}{\theta} = -\lim_{t\to\infty} \frac{1}{\theta t} \log E\left[e^{-\theta \tilde{S}(t)}\right] \quad \forall\, \theta \geq 0. \tag{2.2}$$

Following the analysis of statistical service guarantees in wired networks [17], it can be shown [34] that if the constant arrival rate $a$ satisfies $a \leq \alpha\,(\theta_0)$, then the LDP governing buffer occupancy $\theta$ defined in (1.1) satisfies $\theta \geq \theta_0$. In other words, the service constraint $\theta_0$ will be fulfilled if and only if $a \leq \alpha\,(\theta_0)$.

Consider a block fading model for the wireless channel. That is, the channel coefficients stay invariant within a block of duration $T$, but vary independently from block to block. In this case, an equivalent discrete-time channel model can then be applied. Assume the service process is stationary and ergodic, let $r$ be a random variable that represents the system throughput during one block, then the Gärtner-

Ellis limit of $\tilde{S}(t)$ and the effective capacity in (2.2) reduces to

$$\begin{aligned} \Lambda(-\theta) &= \frac{1}{T}\log E\left[e^{-\theta r}\right], \\ \alpha(\theta) &= \frac{-\Lambda(-\theta)}{\theta} = -\frac{1}{\theta T}\log E\left[e^{-\theta r}\right]. \end{aligned} \tag{2.3}$$

Since the notion of effective capacity serves as an important performance measure throughout the dissertation, we first show some useful properties of the effective capacity expression in (2.3).

**Lemma 1** *For a random variable $r$, let $g(\theta) = \theta\alpha(\theta) = -\Lambda(-\theta)$, where $\alpha(\theta)$ and $\Lambda(-\theta)$ are defined in* (2.3)*. The function $g(\theta)$ is concave in $\theta$.*

*Proof*: Note that $g(\theta)$ is differentiable, and

$$\begin{aligned} g'(\theta) &= \Lambda'(-\theta) = \frac{\mathrm{E}\left[\mathrm{re}^{-\theta \mathrm{r}}\right]}{T\mathrm{E}\left[\mathrm{e}^{-\theta \mathrm{r}}\right]}, \\ g''(\theta) &= -\Lambda''(-\theta) = \frac{\left(\mathrm{E}\left[\mathrm{re}^{-\theta \mathrm{r}}\right]\right)^2 - \mathrm{E}\left[\mathrm{r}^2\mathrm{e}^{-\theta \mathrm{r}}\right]\mathrm{E}\left[\mathrm{e}^{-\theta \mathrm{r}}\right]}{T\left(\mathrm{E}\left[\mathrm{e}^{-\theta \mathrm{r}}\right]\right)^2} \end{aligned}$$

Applying the Cauchy-Schwartz inequality yields

$$\mathrm{E}\left[\mathrm{r}^2\mathrm{e}^{-\theta \mathrm{r}}\right]\mathrm{E}\left[\mathrm{e}^{-\theta \mathrm{r}}\right] = \mathrm{E}\left[\left(\mathrm{re}^{-\frac{\theta}{2}\mathrm{r}}\right)^2\right]\mathrm{E}\left[\left(\mathrm{e}^{-\frac{\theta}{2}\mathrm{r}}\right)^2\right] \geq \left(\mathrm{E}\left[\mathrm{re}^{-\theta \mathrm{r}}\right]\right)^2.$$

Therefore, $g''(\theta) \leq 0$ and $g(\theta)$ is concave. □

**Lemma 2** *The function $\alpha(\theta)$ is monotonically decreasing in $\theta$ for $\theta > 0$.*

*Proof*: Let $g(\theta) = \theta\alpha(\theta)$. Note that $g(0) = -\Lambda(0) = 0$. From the concavity of $g(\theta)$ and $g(0) = 0$, it follows that for $0 < \theta_1 < \theta_2$,

$$g(\theta_1) = g\left(\left(1 - \frac{\theta_1}{\theta_2}\right)\cdot 0 + \frac{\theta_1}{\theta_2}\,\theta_2\right) \geq \left(1 - \frac{\theta_1}{\theta_2}\right)g(0) + \frac{\theta_1}{\theta_2}\,g(\theta_2) = \frac{\theta_1}{\theta_2}\,g(\theta_2).$$

Clearly, $\alpha(\theta_1) = g(\theta_1)/\theta_1 \geq g(\theta_2)/\theta_2 = \alpha(\theta_2)$. Therefore, $\alpha(\theta) = g(\theta)/\theta$ is monotonically decreasing in $\theta$ for $\theta > 0$. □

Lemma 2 implies that the maximum admissible constant arrival rate decreases as the service requirement $\theta_0$ becomes more and more stringent. This reveals a fundamental trade-off between system throughput and service requirement, delay constraint in particular.

**Lemma 3** *If the service process is stationary, then* $\alpha(\theta) \leq E[r]/T$. *Furthermore, if* $r \geq cT$ *almost surely for some constant* $c$, *then* $\alpha(\theta) \geq c$.

*Proof*: Lemma 2 states us that $\alpha(\theta)$ is monotonically decreasing in $\theta$ for $\theta > 0$. Hence $\lim_{\theta \downarrow 0} \alpha(\theta)$ serves as an upper bound for $\alpha(\theta)$ with $\theta > 0$. Taking the limit, we get

$$\lim_{\theta \downarrow 0} \alpha(\theta) = \lim_{\theta \downarrow 0} \frac{\mathrm{E}\left[\mathrm{re}^{-\theta \mathrm{r}}\right]}{T\mathrm{E}\left[\mathrm{e}^{-\theta \mathrm{r}}\right]} = \frac{\mathrm{E}[\mathrm{r}]}{T}.$$

This upper bound can also be visualized from Jensen's inequality,

$$\alpha(\theta) = -\frac{1}{\theta T} \log \mathrm{E}\left[\mathrm{e}^{-\theta \mathrm{r}}\right] \leq -\frac{1}{\theta \mathrm{T}} \mathrm{E}\left[\log\left(\mathrm{e}^{-\theta \mathrm{r}}\right)\right] = \frac{\mathrm{E}[\mathrm{r}]}{\mathrm{T}} = \lim_{\theta \downarrow 0} \alpha(\theta).$$

To prove the second part of the lemma, we note that $r \geq cT$ almost surely implies $\mathrm{E}\left[\mathrm{e}^{-\theta \mathrm{r}}\right] \leq \mathrm{e}^{-\mathrm{c}\theta \mathrm{T}}$. Thus, if $r \geq cT$ then

$$\alpha(\theta) = -\frac{1}{\theta T} \log \mathrm{E}\left[\mathrm{e}^{-\theta \mathrm{r}}\right] \geq -\frac{1}{\theta \mathrm{T}} \log \mathrm{e}^{-\mathrm{c}\theta \mathrm{T}} = \mathrm{c}. \square$$

Lemma 3 gives two nontrivial bounds for the effective capacity of a wireless system. Namely, the effective capacity is upper bounded by the expected capacity and lower bounded by the minimum instantaneous service rate of the corresponding wireless system.

# CHAPTER III

## RESOURCE ALLOCATION

In this chapter, we study the problem of resource allocation in the context of stringent service constraints. To relate radio resources to system performance and statistical service requirements introduced in Chapter II, we link the behavior of the system to its physical-layer infrastructure. A mobile terminal and its associated wireless connection can be modeled as a single-server queue, provided that the receiver has the ability to acknowledge reception of the data. For example, a simple physical-layer automatic repeat request (ARQ) mechanism may be incorporated in the communication protocol to insure that erroneous data is retransmitted. We assume that such a mechanism is in place throughout. Drawing intuition from information theory and error-control coding, the service offered to a mobile terminal can be modeled as a Markov-modulated fluid process. Previous results on Markov-modulated fluid processes and large deviations can therefore be leveraged to characterize the interplay between system resources at the physical layer and the statistical behavior of queues.

The remainder of this chapter is as follows. In Section A, we describe the generic wireless connection that is used as an abstraction for the physical layer. Based on a Markov assumption, we then construct a mathematical representation for the overall channel behavior. Section B contains a derivation of the equilibrium distribution for a system with a constant arrival rate and a Gilbert-Elliott wireless channel. Specifically, buffer overflow probabilities, the corresponding large deviations, and the effective capacity function are given explicit expressions in terms of physical system parameters. This analysis is subsequently extended to a variable data source. The performance analysis of the Gilbert-Elliott queueing system is presented in Section C. We compare and contrast the statistical characteristics of the Gilbert-Elliott model with the

characteristics of a continuous-state Markov channel using numerical simulations in Section D. This is followed by a discussion of these results of this chapter in the last section.

## A. Wireless Channel

The complex base-band representation of the wireless channel under consideration is shown in Fig. 4. The term $g(d)$ accounts for the mean path attenuation, and $h(t;\zeta)$ represents the small-scale variations due to the motion of the terminals and changes in the environment [35]. The additive noise term $w(t)$ is modeled as a proper complex white Gaussian process. Note that $h(t;\zeta)$ is normalized so that the expected power

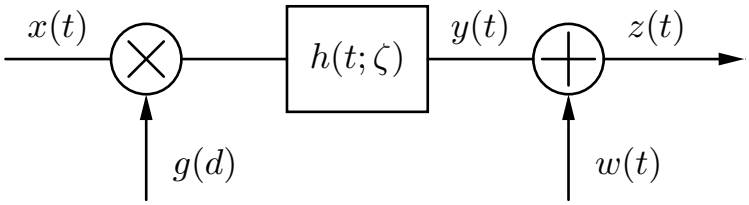


Fig. 4. Block diagram of a wireless communication channel where the transmitted signal is subject to attenuation, fading, and noise corruption.

gain introduced by $h(t;\zeta)$ is equal to one. The bandwidth of the transmitted signal $x(t)$ is assumed to be much smaller than the reciprocal of the delay spread. The channel is therefore purely time-selective, with no frequency distortion [36]. In this case, the standard channel model of Fig. 4 can be written as

$$z(t) = g(d)h(t)x(t) + w(t),$$

where $h(t;\zeta) = h(t)\delta(\zeta)$. Furthermore, we assume that the channel is subject to purely diffuse scattering, i.e., no specular component is present. For a rich scat-

tering environment, the multipath component $h(t)$ is well-modeled as a zero-mean, proper complex Gaussian process. In particular, the envelope process $|h(t)|$ and the phase process are independent, with $|h(t)|$ having a Rayleigh probability distribution function and the phase being uniform over $[0, 2\pi)$.

While it is straightforward to describe the first order statistics of $h(t)$, a complete characterization of this random process requires that joint distributions be specified as well. Under a Gaussian process model, it suffices to describe the correlation between any two sample points of the process. For Rayleigh flat fading, the auto-correlation function of the envelope process can be modeled using the zeroth order Bessel function of the first kind. This function is reasonable over short time horizons corresponding to terminal movements of the order of a few wavelengths. It is derived under the assumption that a mobile terminal is moving in an isotropic environment at a constant velocity. Alternatively, an auto-correlation function can be derived by assuming that the in-phase and quadrature components of $h(t)$ are independent stationary Ornstein-Uhlenbeck processes [37]. The latter model states that the correlation between two samples decays exponentially over time.

These two auto-correlation structures are useful in various contexts. However, for the sake of mathematical tractability, we consider a slightly simplified channel model. We retain the first-order statistics of the channel and assume that the marginal distribution of the envelope process is Rayleigh. Second, we assume that for a fixed threshold $\eta$ the probability of $|h(t)|$ being above or below this threshold is accurately modeled as a continuous-time Markov chain. We refer to the channel envelope exceeding $\eta$ as the "ON" state; otherwise the channel is in its "OFF" state. Such a channel structure is commonly referred to as the Gilbert-Elliott model. It is assumed to provide a sufficiently accurate representation for the statistical behavior of the quantized Rayleigh channel. This quantized channel model appears in Fig. 5. The

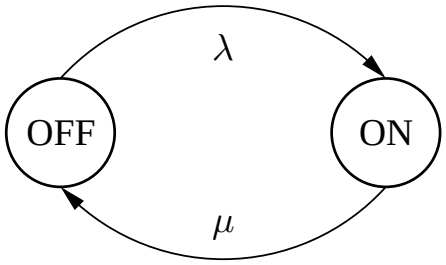


Fig. 5. Continuous-time Gilbert-Elliott Markov representation of a wireless communication link.

transition rate from OFF to ON is denoted by $\lambda$; while the transition rate from ON to OFF, by $\mu$. The generator matrix for this Markov chain is given by

$$Q_s = \begin{bmatrix} -\lambda & \lambda \\ \mu & -\mu \end{bmatrix} = \frac{1}{\lambda+\mu} \begin{bmatrix} 1 & \lambda \\ 1 & -\mu \end{bmatrix} \begin{bmatrix} 0 & 0 \\ 0 & -(\lambda+\mu) \end{bmatrix} \begin{bmatrix} \mu & \lambda \\ 1 & -1 \end{bmatrix}.$$

It is easy to verify that the invariant probability for the ON state is $\lambda/(\lambda+\mu)$, while the invariant probability of being OFF is $\mu/(\lambda+\mu)$. For consistency, the stationary distribution of the Markov chain should agree with the marginal distribution of the underlying channel,

$$\begin{aligned} \Pr\{|\mathrm{h(t)}| \le \eta\} &= \frac{\mu}{\lambda+\mu} = \int_0^{\eta} 2\xi e^{-\xi^2} d\xi = 1 - e^{-\eta^2} \\ \Pr\{|\mathrm{h(t)}| > \eta\} &= \frac{\lambda}{\lambda+\mu} = \int_{\eta}^{\infty} 2\xi e^{-\xi^2} d\xi = e^{-\eta^2} \end{aligned} \tag{3.1}$$

where $f(\xi) = 2\xi e^{-\xi^2}$ with $\xi \ge 0$ is the marginal distribution of the normalized envelope process. To solve for $\lambda$ and $\mu$, two equations are needed. The first one is given by condition (3.1). The second condition can be derived from the Markov structure of the wireless link.

Let $P_t(t) = e^{Q_s t}$ be the probability transition matrix of the the Gilbert-Elliott

channel. More specifically, entry $p_{i,j}(t)$ of the matrix $P_t(t)$ represents the probability of being in state $j$ after $t$ seconds, when starting in state $i$. For a time interval $t$, this probability transition matrix is given by

$$\begin{aligned} P_t(t) &= \frac{1}{\lambda+\mu}\begin{bmatrix} 1 & \lambda \\ 1 & -\mu \end{bmatrix}\begin{bmatrix} 1 & 0 \\ 0 & e^{-t(\lambda+\mu)} \end{bmatrix}\begin{bmatrix} \mu & \lambda \\ 1 & -1 \end{bmatrix} \\ &= \frac{1}{\lambda+\mu}\begin{bmatrix} \mu+\lambda e^{-t(\lambda+\mu)} & \lambda-\lambda e^{-t(\lambda+\mu)} \\ \mu-\mu e^{-t(\lambda+\mu)} & \lambda+\mu e^{-t(\lambda+\mu)} \end{bmatrix}. \end{aligned}$$

We note that the channel memory of this two-state Markov process decays at a rate $\lambda+\mu$. Thus, if the memory of the underlying quantized Rayleigh channel has an exponential decay rate $\kappa$, we must have $\lambda+\mu=\kappa$. This relationship provides the second equation necessary to determine $\lambda$ and $\mu$. Solving for these parameters explicitly in terms of the channel parameters, we get

$$\begin{aligned} \lambda &= \kappa e^{-\eta^2} \\ \mu &= \kappa - \kappa e^{-\eta^2}. \end{aligned}$$

The quantized channel and its associated Markov structure will prove instrumental in computing the probability of buffer overflow and the effective capacity of the associated wireless connection. The more elaborate channel description whose in-phase and quadrature components are independent stationary Orstein-Uhlenbeck processes will be revisited in Section D.

### 1. Coding and Information Theory

A celebrated result from information theory is the Shannon capacity of the Gaussian channel,

$$C = W\log_2\left(1+\frac{P}{N_0 W}\right) \quad \text{bits per second.} \tag{3.2}$$

The variable $P$ represents the power of the signal, $N_0/2$ denotes the power spectral density of the noise process, and $W$ is the channel bandwidth. In theory, error-free communication can be achieved on this channel for any rate below the capacity using asymptotically long codewords [31]. Today, there exists a collection of practical codes that operate close to capacity, with minimal error-rates and small delays. The capacity expression of (3.2) can therefore be employed as an optimistic approximation of code performance. If a code is designed to operate at a rate $R$, the sent information is decoded reliably whenever $R < C$; it is lost otherwise.

A similar performance description can be employed for time-varying channels such as the one introduced at the beginning of Section A. Suppose that a wireless channel varies slowly, data is assumed to reach its destination provided that $R < C(t)$, where

$$C(t) = W \log_2 \left(1 + \frac{P|h(t)|^2}{N_0 W}\right) \quad \text{bits per second} \tag{3.3}$$

is the instantaneous capacity. On the other hand, if $R \geq C(t)$ then information is lost. This simplified characterization is valid provided that the decoding delay is small compared to the coherence time of the wireless channel. It is used in this work for mathematical convenience and because it yields useful guidelines on how to select coderates for specific systems and statistical service requirements. This model can be altered to accommodate real codes and probabilities of link failures.

The state of the Gilbert-Elliott channel is related to the instantaneous capacity and the coderate as follows. Let coderate $R$ be given. The Gilbert-Elliott channel is ON if $R < C(t)$ or, equivalently,

$$|h(t)| > \eta = \sqrt{\frac{N_0 W}{P}\left(2^{\frac{R}{W}} - 1\right)}. \tag{3.4}$$

It is OFF otherwise. We can rewrite the generator matrix for this Gilbert-Elliott

channel as

$$Q_s = \begin{bmatrix} -\kappa e^{-\eta^2} & \kappa e^{-\eta^2} \\ \kappa - \kappa e^{-\eta^2} & -\kappa + \kappa e^{-\eta^2} \end{bmatrix}, \tag{3.5}$$

where $\kappa$ is the exponential decay parameter of the Markov chain and $\eta$ is the threshold defined in (3.4). The corresponding service rate is zero when the channel is OFF and $R$ when the channel is in its ON state.

Under these assumptions, the maximum throughput of this wireless channel is immediately seen to equal

$$R \Pr\{|\mathrm{h(t)}| > \eta\} = \mathrm{Re}^{-\eta^2}.$$

This throughput can be optimized by selecting a proper coderate $R$. A higher rate allows more information to be transmitted when the channel is ON. However, it also implies that the channel is ON less often (larger $\eta$). Conversely, a lower rate increases the probability of the channel being ON but reduces the rate at which data is transferred. Fig. 6 plots the throughput as a function of coderate $R$. The parameter values for the wireless channel used in this example appear in Table I. The maximum average throughput is 508 Kbps, and it is achieved with a coderate $R = 1.33$ Mbps.

Table I. System parameters.

| | |
|---|---|
| $N_0 = 10^{-7}$ W/Hz | Noise power spectral density |
| $W = 11$ MHz | Bandwidth |
| $P = 100$ mW | Received Power |

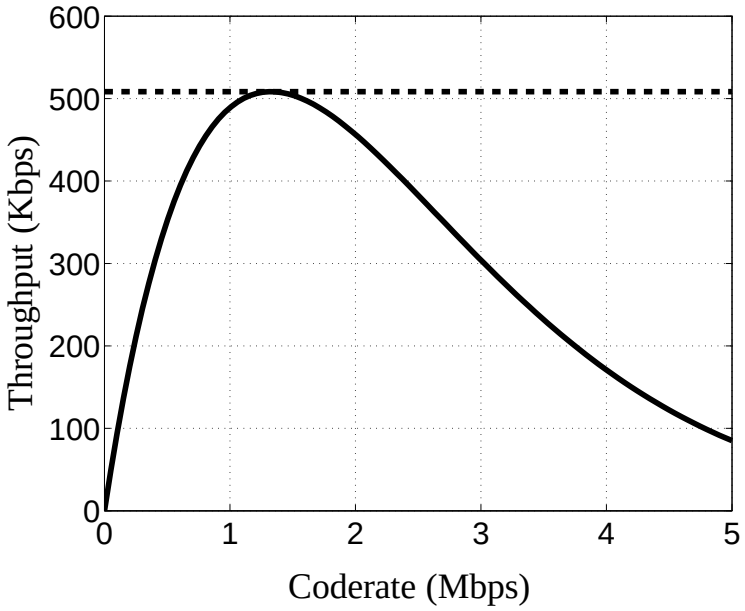


Fig. 6. Mean throughput as a function of coderate $R$ for a Gilbert-Elliott channel model.

## B. Queueing Performance of Markov-Modulated Processes

In a wireless system, much like in a broadband network, the usage of system resources may not be well assessed by the single value of throughput or Shannon capacity. Performance measures such as queue length, packet loss probability, and delay play an instrumental role in user satisfaction. Requirements on these attributes may force a wireless system to operate much below its theoretical Shannon limit. Furthermore, the unreliable link-quality intrinsic to wireless communication along with stringent delay and loss constraints may significantly alter the optimal allocation of system resources. This is exemplified below.

### 1. Markov Fluid Model of a Queue

Consider a simple fluid queueing system with a single queue and one server. Let $a(t)$ denote the instantaneous arrival rate, and let $r(t)$ be the instantaneous service rate. The cumulative arrival function over interval $[0, t]$ is given by

$$A[0,t] = \int_{[0,t]} a(\tau) d\tau.$$

Similarly, the amount of service offered in the interval $[0, t]$ is equal to

$$S[0,t] = \int_{[0,t]} r(\tau) d\tau.$$

Under a work-conserving policy and provided that the queue is initially empty, the state of the queue is governed by the following equation [17],

$$L_t = (A[0,t] - S[0,t]) - \inf_{0<\tau<t} \{A[0,\tau] - S[0,\tau]\}.$$

This generic model provides an appropriate framework for evaluating the performance of a queueing system subject to service constraints.

A natural choice to model the communication system introduced in the previous section is a Markov-modulated fluid process. Consider a queue subject to a Markov-modulated rate process. Let $Q$ be the generator matrix of the underlying finite-state Markov process, and assume that $Q$ is irreducible with state space $\{1, \ldots, M\}$. In particular, the off-diagonal entry $q_{n,m}$ represents the transition rate of going from state $n$ to $m$; and the corresponding diagonal entry is $q_{n,n} = -\sum_{m \neq n} q_{n,m}$, making the total row sum zero. The state $m$ is associated with a rate $d_m$, which represents the difference between the instantaneous arrival rate and the instantaneous service rate. Hence, the net rate of change in the buffer while in state $m$ is $d_m$ when the buffer is not empty, and it is equal to $\max\{0, d_m\}$ when the buffer is empty. In other words, when the buffer is empty and the Markov process is in state $m$ with $d_m \leq 0$ then the buffer simply remains empty. We denote the diagonal matrix $\mathrm{diag}(d_1, \ldots, d_M)$ by $D$.

If we use $L_t$ to denote the level of fluid in the buffer at time $t$ and we let $u_t$ be the state of the underlying Markov chain at time $t$, then $(L_t, u_t)$ forms a continuous-state Markov process. Define the event probability

$$F(x, m, t) = \Pr\left\{\mathrm{u_t} = \mathrm{m}, \mathrm{L_t} \leq \mathrm{x}\right\}.$$

Using the Chapman-Kolmogorov equation, we find that the function $F(x, m, t)$ satisfies

$$\frac{\partial F}{\partial t} = FQ - \frac{\partial F}{\partial x} D \tag{3.6}$$

where $F = (F(x, 1, t), \ldots, F(x, M, t))$. The equilibrium distribution $F(x, m)$ of the continuous state Markov process $(L_t, u_t)$ is subject to $\partial F / \partial t = 0$, which in turn yields

$$FQ = \frac{\partial F}{\partial x} D. \tag{3.7}$$

We denote the invariant probability distribution of the underlying Markov chain by

$w$, with $wQ = 0$. Then

$$\lim_{x\to\infty} F(x,m) = w_m.$$

Since the equilibrium distribution is a bounded solution to (3.7), it has spectral representation

$$F(x,\cdot) = w - \sum_{i=1}^{k} a_i \phi_i e^{xz_i} \tag{3.8}$$

where $\{(\phi_i, z_i)\}$ are the stable eigenvector-eigenvalue pairs of the eigenvalue problem

$$\phi Q = z\phi D. \tag{3.9}$$

If $Q$ is reversible [38], then all such eigenvalues are real numbers [39]. Moreover, there are $k = |\{m : d_m > 0\}|$ strictly negative eigenvalues (counting multiplicity). These values are the ones included in (3.8). The coefficients $\{a_i\}$ are found using the boundary conditions

$$F(0,m) = 0 \quad \forall\, \{m : d_m > 0\}.$$

The unique solution to this boundary value problem is the equilibrium distribution [40]. The reader is referred to Mitra [41] and Meyn [42] for additional information about fluid models.

### 2. Equilibrium Distribution of Gilbert-Elliott Systems

Consider a communication system where data arrives in a buffer at a constant rate $a(t) = a$. Suppose that this buffer is serviced through a wireless connection at a rate $r(t)$, where $r(t)$ is the Markov-modulated process described in Section A. That is, $r(t)$ is equal to $R$ when the channel is ON, and zero otherwise. We assume that the generator matrix of the underlying finite-state Markov chain is the matrix $Q_s$

obtained in (3.5), i.e.,

$$Q = Q_s = \begin{bmatrix} -\lambda & \lambda \\ \mu & -\mu \end{bmatrix}.$$

Note that $Q$ is reversible since $Q$ and $w$ are in detailed balance, $w_1 q_{1,2} = w_2 q_{2,1}$. The net arrival rate in the buffer when the buffer is not empty is

$$D = \begin{bmatrix} a & 0 \\ 0 & a - R \end{bmatrix}.$$

The eigenvalue problem $\phi Q = z\phi D$ has two solution pairs:

$$\begin{aligned} (w, 0) &= \left( \left( \frac{\mu}{\lambda + \mu}, \frac{\lambda}{\lambda + \mu} \right), 0 \right) \\ (\phi, z) &= \left( (R - a, a), \frac{a\lambda + a\mu - R\lambda}{a(R - a)} \right). \end{aligned}$$

The queue will be stable provided that

$$a < \frac{\lambda}{\lambda + \mu} R. \tag{3.10}$$

Under this condition, $L_t$ converges in distribution to a finite random variable $L$. Using the boundary condition $F(0, 1) = 0$, we obtain the equilibrium solution

$$\begin{aligned} F(x, \cdot) &= \left( \frac{\mu}{\lambda + \mu}, \frac{\lambda}{\lambda + \mu} \right) - \left( \frac{\mu}{\lambda + \mu}, \frac{a}{R - a} \frac{\mu}{\lambda + \mu} \right) \exp\left( \frac{a\lambda + a\mu - R\lambda}{a(R - a)} x \right) \\ &= \left( 1 - e^{-\eta^2}, e^{-\eta^2} \right) - \left( 1 - e^{-\eta^2} \right) \left( 1, \frac{a}{R - a} \right) \exp\left( \frac{\kappa a - R\kappa e^{-\eta^2}}{a(R - a)} x \right). \end{aligned} \tag{3.11}$$

Based on this equilibrium distribution, we can compute a number of performance metrics including the probability of buffer overflow, its exponential rate of convergence to zero, and the effective capacity of the system.

The probability of buffer overflow is an important performance metric. For the Gilbert-Elliott system at hand, the probability of the buffer exceeding a threshold $x$

is given by

$$\begin{aligned}\Pr\{\mathrm{L} > \mathrm{x}\} &= 1 - \langle F(x,\cdot),(1,1)\rangle = 1 - \langle F(x,\cdot),\mathbf{1}\rangle \\ &= \frac{R}{R-a}\frac{\mu}{\lambda+\mu}\exp\left(\frac{a\lambda + a\mu - R\lambda}{a(R-a)}x\right) \\ &= \frac{R}{R-a}\left(1-e^{-\eta^2}\right)\exp\left(\frac{\kappa a - R\kappa e^{-\eta^2}}{a(R-a)}x\right).\end{aligned}$$

As seen in (3.11), the eigenvalue problem (3.9) applied to the present two-state system contains only one negative solution. The large deviation principle governing the probability of buffer overflow is therefore immediately seen to equal

$$-\lim_{x\to\infty}\frac{\log\Pr\{\mathrm{L} > \mathrm{x}\}}{x} = -\frac{a\lambda + a\mu - R\lambda}{a(R-a)} = -\frac{\kappa a - R\kappa e^{-\eta^2}}{a(R-a)}.$$

The large deviation principle governing the distribution of a queue is sometimes preferable as a design criterion because it admits a simpler form.

Given specific system parameters and an exponential decay rate $\theta$, the effective capacity is the maximum arrival rate for which the service requirement $\theta$ is fulfilled. Mathematically, this can be expressed as

$$\alpha(\theta) = \sup\left\{a \geq 0 : -\lim_{x\to\infty}\frac{\log\Pr\{\mathrm{L} > \mathrm{x}\}}{x} \geq \theta\right\}. \tag{3.12}$$

For the simple problem considered in this section, (3.12) leads to

$$\begin{aligned}\alpha(\theta) &= \sup\left\{a \geq 0 : \frac{a\lambda + a\mu - R\lambda}{a(R-a)} \leq -\theta\right\} \\ &= \sup\left\{a \geq 0 : \theta a^2 - (\theta R + \lambda + \mu)a + R\lambda \geq 0\right\}.\end{aligned} \tag{3.13}$$

Taking into account condition (3.10), this yields an explicit formula for the effective

capacity of the Gilbert-Elliott channel

$$\begin{aligned}\alpha(\theta) &= \frac{\theta R + \lambda + \mu - \sqrt{(\theta R + \lambda + \mu)^2 - 4\theta R\lambda}}{2\theta} \\ &= \frac{\theta R + \kappa - \sqrt{(\theta R + \kappa)^2 - 4\theta R\kappa e^{-\eta^2}}}{2\theta}.\end{aligned} \tag{3.14}$$

Appendix B shows that (3.14) can also be obtained directly from (2.2).

## 3. On-Off Information Sources

Some traffic sources are better modeled as on-off sources. Voice, for instance, is a good example of an information process that can be accurately modeled as an on-off source. When two people are carrying a conversation, they are unlikely to speak simultaneously. On average, a person involved in a discussion speaks at most half of the time. Other data sources such as instant messaging applications and wireless sensors [43] can also be modeled as on-off sources. As such, we extend the analysis of the previous section to the case where the data source features an on-off behavior.

Suppose that data arrive in the buffer at a rate $a(t)$, where $a(t)$ is a two-state Markov-modulated source. We assume that the arrival rate is equal to $a > 0$ when the source is ON; it is equal to zero otherwise. The generator matrix of the underlying Markov chain for this arrival process can be written as

$$Q_a = \begin{bmatrix} -\lambda_a & \lambda_a \\ \mu_a & -\mu_a \end{bmatrix}.$$

Again, we assume that the service offered through the wireless channel is a Markov-modulated process with generator matrix $Q_s$, as defined in (3.5). The aggregate system is therefore a stochastic fluid process modulated by a four-state Markov chain. The evolution of the buffer content is governed by (3.6), where the generator matrix

$Q$ is equal to

$$Q = \begin{bmatrix} -\lambda_a - \lambda & \lambda_a & \lambda & 0 \\ \mu_a & -\mu_a - \lambda & 0 & \lambda \\ \mu & 0 & -\lambda_a - \mu & \lambda_a \\ 0 & \mu & \mu_a & -\mu_a - \mu \end{bmatrix}$$
$$= Q_s \otimes I + I \otimes Q_a.$$

Throughout, we use $A \otimes B$ to denote the Kronecker product of matrices $A$ and $B$. Again, it is straightforward to verify that the generator matrix $Q$ is reversible. The net arrival rate in the buffer is represented by

$$D = \text{diag}(0, a, -R, a - R)$$
$$= -RE \otimes I + aI \otimes E,$$

where the matrix $E$ is defined by

$$E = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}.$$

We assume that the system is stable; that is, the following condition is satisfied

$$\frac{\lambda_a}{\lambda_a + \mu_a} a < \frac{\lambda}{\lambda + \mu} R.$$

The equilibrium distribution of this system is governed by (3.7), and its spectral representation follows the general form of (3.8).

The generalized eigenvalue problem $\phi Q = z\phi D$ admits three eigenvector-eigenvalue pairs in the present case because $\det(Q - zD)$ is a third order polynomial ($a \neq R$). We note that the underlying Markov chain is reversible since it satisfies the detailed balance condition, $w_i Q_{ij} = w_j Q_{ji}$ for $1 \leq i, j \leq 4$. This property insures that all the eigenvalues of $\phi Q = z\phi D$ are real numbers. The first eigenvector is the invariant

distribution given by

$$\begin{aligned} w &= \left(\frac{\mu}{\lambda+\mu}, \frac{\lambda}{\lambda+\mu}\right) \otimes \left(\frac{\mu_a}{\lambda_a+\mu_a}, \frac{\lambda_a}{\lambda_a+\mu_a}\right) \\ &= \frac{(\mu\mu_a, \mu\lambda_a, \lambda\mu_a, \lambda\lambda_a)}{(\lambda+\mu)(\lambda_a+\mu_a)}. \end{aligned} \tag{3.15}$$

The associated eigenvalue is, of course, zero. To find the remaining two eigenvector-eigenvalue pairs, we make an educated guess based on a standard decomposition technique popularized by Mitra [41]. For any vector of the form $\phi = \phi_s \otimes \phi_a$, we can rewrite (3.9) as

$$(\phi_s Q_s + Rz\phi_s E) \otimes \phi_a = \phi_s \otimes (az\phi_a E - \phi_a Q_a)\,. \tag{3.16}$$

Consider the two vectors defined by

$$\phi_v = (R - v, v) \otimes (a - v, v) \tag{3.17}$$

where $v$ is either solution of the quadratic form

$$\begin{aligned} &v^2(\lambda+\mu+\lambda_a+\mu_a) - vR(\lambda+\lambda_a+\mu_a) \\ &- va(\lambda+\mu+\lambda_a) + aR(\lambda+\lambda_a) = 0. \end{aligned} \tag{3.18}$$

It is straightforward to show that the vectors jointly defined by (3.17) and (3.18) are eigenvectors of (3.16), with corresponding eigenvalues

$$z = \frac{v\lambda + v\mu - R\lambda}{v(R-v)} = \frac{a\lambda_a - v\lambda_a - v\mu_a}{v(a-v)}. \tag{3.19}$$

Incidentally, (3.18) is obtained by equating the above two expressions for $z$. Since (3.18) has two distinct real roots, the two associated eigenvectors along with the invariant distribution described in (3.15) completely characterize the eigenvalue problem $\phi Q = z\phi D$. We can see from the spectral representation of the equilibrium distribution (3.8) that the large deviation principle governing the queue occupancy is

dominated by the largest negative eigenvalue of (3.16).

Consider an exponential decay rate requirement of $\theta > 0$ on the probability of buffer overflow,

$$-\lim_{x\to\infty} \frac{\log \Pr\{\mathrm{L} > \mathrm{x}\}}{x} \geq \theta. \tag{3.20}$$

This requirement will be satisfied provided that the largest negative eigenvalue of (3.19) is less than $-\theta$. In particular, we want the following equations to hold:

$$v^2\theta - v(\theta R + \lambda + \mu) + R\lambda \geq 0$$
$$v^2\theta - v(\theta a - \lambda_a - \mu_a) - a\lambda_a \geq 0.$$

These conditions will be fulfilled if and only if the value of $v$ corresponding to the largest negative eigenvalue of (3.19) is less than $\alpha(\theta)$ but greater than $\beta(\theta)$, where $\alpha(\theta)$ is the effective capacity introduced earlier

$$\begin{aligned}\alpha(\theta) &= \sup\left\{\nu \geq 0 : \frac{\nu\lambda + \nu\mu - R\lambda}{\nu(R-\nu)} \leq -\theta\right\} \\ &= \sup\left\{\nu \geq 0 : \theta\nu^2 - (\theta R + \lambda + \mu)\nu + R\lambda \geq 0\right\}\end{aligned}$$

and $\beta(\theta)$ is the effective bandwidth of a two-state Markov-modulated fluid source [44, 13, 45]

$$\begin{aligned}\beta(\theta) &= \inf\left\{\nu \geq 0 : \frac{a\lambda_a - \nu\lambda_a - \nu\mu_a}{\nu(a-\nu)} \leq -\theta\right\} \\ &= \inf\left\{\nu \geq 0 : \theta\nu^2 - (\theta a - \lambda_a - \mu_a)\nu - a\lambda_a \geq 0\right\}.\end{aligned}$$

Clearly, the service requirement of (3.20) can only be met if $\alpha(\theta) \geq \beta(\theta)$. The fact that this inequality is a necessary and sufficient condition for (3.20) to hold is proved in Chapter IV.

This observation greatly facilitates the performance analysis contained in the next section. In particular, for a service requirement such as (3.20), an on-off source

shares the same service needs as a constant source with rate $\beta(\theta)$. Thus, for a given $\theta > 0$, the allocation of system resources can be studied in terms of fixed arrival rates, whether the source rate is a constant or a Markov modulated fluid model. This is illustrated in the next section.

## C. Performance Analysis of Gilbert-Elliott Systems

We proceed to analyze the performance of the Gilbert-Elliott system as a function of physical resources. Following the literature on effective bandwidth, we use the exponential decay rate as our primary performance measure.

### 1. Effective Capacity Analysis

The effective capacity quantifies the maximum supported arrival rate for a set of system parameters and a service constraint $\theta > 0$. It is an appropriate tool to quantify the optimal operating point of a wireless system. This maximum rate can either be the true rate of a constant source or the effective bandwidth of a time-varying source. Fig. 7 shows the maximal supported arrival rate $\alpha(\theta)$ as a function of the service constraint $\theta$ for the system parameters of Table I and the Markov decay parameters $\kappa \in \{10^2, 10^3, 10^4\}$. This figure also includes the optimal coderate $R$ as a function of $\theta$. We emphasize that for queueing constraint $\theta > 0$, the optimal coderate $R$ differs from the throughput maximizing rate introduced in Section 1. Not surprisingly, more stringent service constraints result in lower effective capacities for fixed system parameters. This is intuitive since a lower arrival rate reduces the expected queue length. More importantly, we see that the optimal coderate $R$ is also a function of the service requirements. Under strict service constraints, an error-control code with a lower rate performs better as it reduces the probability of the channel being OFF.

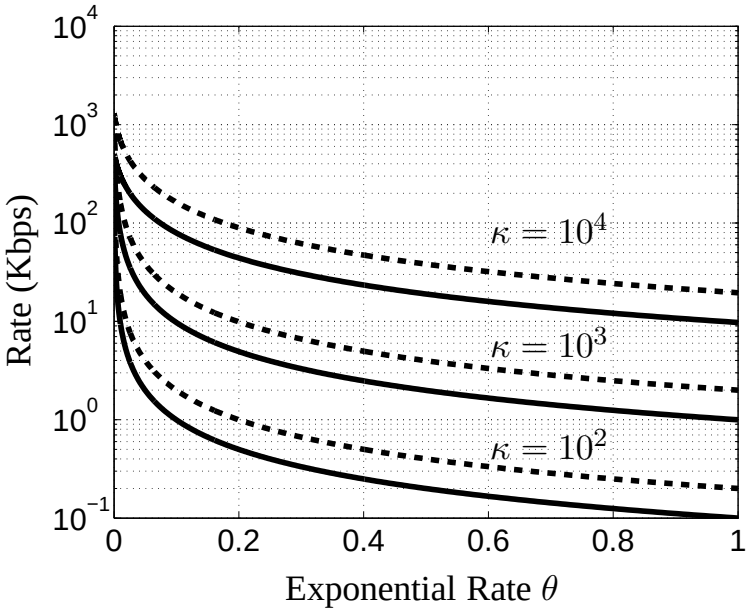


Fig. 7. Optimal coderate (dashed) and effective capacity (solid) as a function of exponential decay rate $\theta$ for various Markov decay parameters $\kappa \in \{10^2, 10^3, 10^4\}$.

This analysis provides a new and systematic way to select the coderate as a function of the channel profile and the service requirement of a specific system.

It is interesting to note that the maximum throughput and the corresponding coderate are independent of the Markov decay parameter $\kappa$,

$$\begin{aligned}\lim_{\theta\to 0}\alpha(\theta) &= \lim_{\theta\to 0}\frac{\theta R+\lambda+\mu-\sqrt{(\theta R+\lambda+\mu)^2-4\theta R\lambda}}{2\theta}\\ &= \lim_{\theta\to 0}\left[\frac{R}{2}-\frac{(\theta R+\lambda+\mu)R-2R\lambda}{2\sqrt{(\theta R+\lambda+\mu)^2-4\theta R\lambda}}\right]\\ &= \frac{R\lambda}{\lambda+\mu} = Re^{-\eta^2}.\end{aligned}$$

However, the effective capacity for $\theta > 0$ depends heavily on the statistical profile of the channel. Correlation impairs effective capacity. The higher the correlation coefficient, the lower the effective capacity. In other words, a throughput analysis of this system is not sufficient to provide an accurate assessment of supported rates under strict service constraints. This also implies that the common assumption that channel realizations are independent and identically distributed through time may lead to over-optimistic performance predictions on effective capacity.

### 2. Resource Requirement Analysis

The effective capacity shown in Fig. 7 decays rapidly as a function of $\theta$. It is therefore of interest to look at the reverse problem; for a given arrival rate $a$, we wish to characterize the amount of physical resources necessary to meet a prescribed service constraint $\theta > 0$. First, we note that a necessary condition for a solution to exist is the stability criterion $a < Re^{-\eta^2}$. However, this condition may not be sufficient. Looking at (3.13), we see that a solution exists if and only if we can find a power $P$

and a bandwidth $W$ such that

$$\theta a^2 - (\theta R + \kappa)a + R\kappa e^{-\eta^2} = 0.$$

This equation can be rearranged as

$$\eta^2 = \frac{N_0 W}{P}\left(2^{\frac{R}{W}} - 1\right) = -\log\left(\frac{\theta R a + \kappa a - \theta a^2}{R\kappa}\right).$$

Since $\eta^2 > 0$, the following inequality must apply

$$0 < \theta R a + \kappa a - \theta a^2 < R\kappa.$$

A necessary and sufficient condition for a solution to exist is $\theta < \kappa/a$. We emphasize that, even with an unlimited power and spectral bandwidth budget, only a finite arrival rate can be supported for a service constraint $\theta > 0$. Furthermore, this bound is independent of the actual coderate $R$ used in the system. This fact is in sharp contrast with Shannon capacity, which goes to infinity as power and spectral bandwidth grow unbounded. This limitation is partly due to the fact that, in the system under study, the transmitter has no knowledge of the channel gain. Thus, it cannot transmit at the (error-free) instantaneous channel capacity. Without channel state information, the best decay rate $\theta$ is limited by the ratio of $\kappa$ to the arrival rate $a$. In the limit where the power and spectral bandwidth become very large, the queueing behavior of the system is increasingly dominated by the holding time of its OFF state. The queue is drained almost instantaneously when the channel is ON, while it rises linearly when the channel is OFF. The probability of the queue exceeding a threshold is then dominated by the duration of an OFF period, which is exponentially distributed.

Fig. 8 shows the target power $P$ as a function of the service constraint $\theta$ for an arrival rate $a = 14.4$ Kbps and the parameters of Table I. Note that power $P$ can be

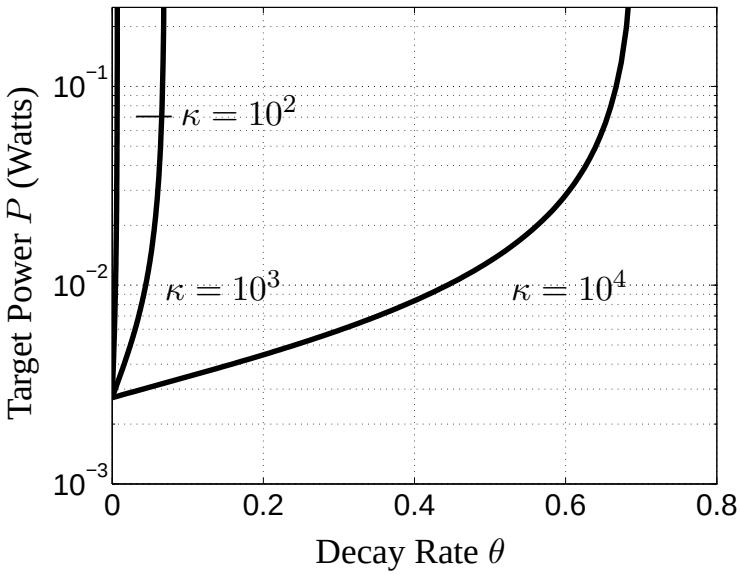


Fig. 8. Signal power as a function of exponential decay rate $\theta$ for various Markov decay parameters $\kappa \in \{10^2, 10^3, 10^4\}$.

obtained in closed form as

$$P = -\frac{N_0 W \left(2^{\frac{R}{W}} - 1\right)}{\left(\log\left(\theta R a + \kappa a - \theta a^2\right) - \log\left(R\kappa\right)\right)}.$$

As expected, the required power goes to infinity for a finite $\theta$. The set of supported exponential decay rates is intimately connected to the Markov decay factor $\kappa$. Not only does correlation decrease effective capacity, it also limits the rates and qualities of service that can be sustained on a given wireless channel. Similar findings can be obtained for the spectral bandwidth requirement as a function of arrival rate $a$ and service constraint $\theta$. Fig. 9 shows minimum spectral bandwidth as a function of decay rate $\theta$ for an arrival rate of $a = 14.4$ Kbps. Again, the amount of resource required goes to infinity for a finite $\theta$.

The speech process of an interlocutor involved in an English conversation can be

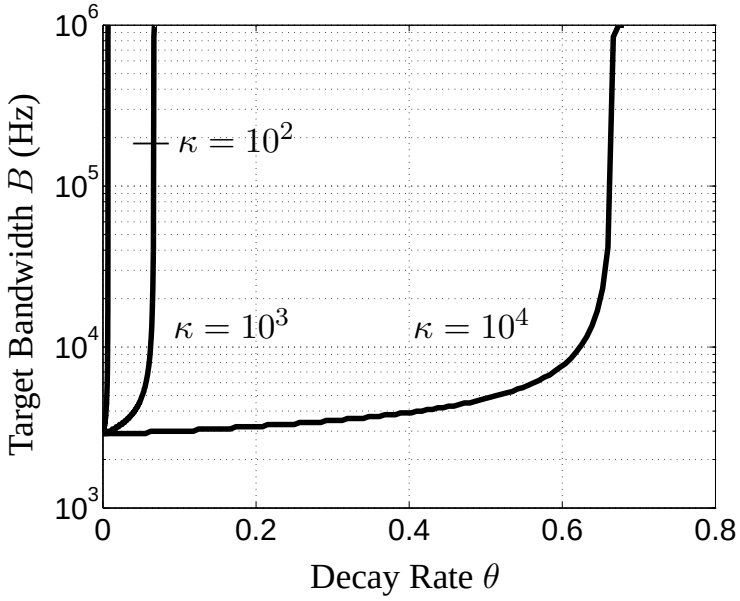


Fig. 9. Spectral bandwidth as a function of exponential decay rate $\theta$ for various Markov decay parameters $\kappa \in \{10^2, 10^3, 10^4\}$.

modeled as an on-off information source [46]. Exponentially distributed talk spurts with a mean duration of $\mu_a^{-1} \approx 352$ ms are followed by silent periods with mean $\lambda_a^{-1} \approx 650$ ms. Using advanced signal processing techniques, active speech can be compressed to a rate of 14.4 Kbps. The average throughput of an encoded speech process is therefore 5.06 Kbps. However, for a delay constraint of 20 ms or an approximately equivalent service constraint $\theta = 0.5$, the effective bandwidth of speech is essentially equal to 14.4 Kbps. The very strict delay constraint imposed on speech traffic forces the effective bandwidth to be nearly equal to its peak rate, which is much higher than the average throughput. As seen on Fig. 8 & 9, voice traffic cannot be successfully transmitted over highly correlated channels without sophisticated power control. This partly explains why power control is critical to cellular telephony [47, 48, 49].

## D. Numerical Analysis

In Section A, the Gilbert-Elliott channel model is introduced as a first-order approximation to the auto-correlation of a Rayleigh fading channel. This simplified channel model permits the derivation in close form of many important quantities, including the probability of buffer overflow and the effective capacity. Recall that the Gilbert-Elliott model is based on two assumptions. First, the state of the Gilbert-Elliott channel identifies whether the instantaneous realization of the underlying Rayleigh channel lies above or below a prescribed threshold. Second, the stochastic process representing the time evolution of this quantized channel is accurately modeled as a two-state, continuous-time Markov chain.

While it is mathematically convenient to assume that the quantized channel possesses the Markov property, a more common approach is to assume that the channel itself is Markov (not the quantized version). Furthermore, we note that it is straightforward to construct a Rayleigh fading channel that possesses the Markov property. In particular, consider the Ornstein-Uhlenbeck equation

$$dX_t = -\kappa X_t dt + \sigma dB_t$$

where $\kappa$, $\sigma$ are real constants and $B_t$ is a one-dimensional Brownian motion. The solution to this stochastic differential equation is called the Ornstein-Uhlenbeck process. This solution has the Markov property and it is given by [37, 50]

$$X_t = X_0 e^{-\kappa t} + \int_0^t e^{-\kappa(t-s)} \sigma dB_s.$$

The variance of this process at time $t$ can be computed explicitly as

$$\begin{aligned}
&\mathrm{E}\left[(\mathrm{X_t}-\mathrm{E}[\mathrm{X_t}])^2\right] \\
&= \mathrm{E}\left[\left(\mathrm{X_0 e^{-\kappa t}} + \int_0^{\mathrm{t}} \mathrm{e}^{-\kappa(\mathrm{t-s})}\sigma \mathrm{dB_s} - \mathrm{E}[\mathrm{X_0}]\mathrm{e}^{-\kappa \mathrm{t}}\right)^2\right] \\
&= \mathrm{E}\left[(\mathrm{X_0}-\mathrm{E}[\mathrm{X_0}])^2\right]\mathrm{e}^{-2\kappa \mathrm{t}} + \mathrm{E}\left[\left(\int_0^{\mathrm{t}} \mathrm{e}^{-\kappa(\mathrm{t-s})}\sigma \mathrm{dB_s}\right)^2\right] \\
&= \mathrm{E}\left[(\mathrm{X_0}-\mathrm{E}[\mathrm{X_0}])^2\right]\mathrm{e}^{-2\kappa \mathrm{t}} + \mathrm{E}\left[\int_0^{\mathrm{t}} \mathrm{e}^{-2\kappa(\mathrm{t-s})}\sigma^2 \mathrm{ds}\right] \\
&= \mathrm{E}\left[(\mathrm{X_0}-\mathrm{E}[\mathrm{X_0}])^2\right]\mathrm{e}^{-2\kappa \mathrm{t}} + \frac{\sigma^2}{2\kappa}\left(1-\mathrm{e}^{-2\kappa \mathrm{t}}\right).
\end{aligned}$$

If $X_0 \sim \mathcal{N}\left(0, \frac{1}{2}\right)$ and $\sigma^2 = \kappa$, then $X_t \sim \mathcal{N}\left(0, \frac{1}{2}\right)$ for all $t \geq 0$. A Raleigh fading channel that possesses the Markov property can therefore be obtained by assigning independent stationary Ornstein-Uhlenbeck processes to the in-phase and quadrature component of the channel. The first order statistic of the corresponding $h(t)$ is a zero-mean, proper complex Gaussian process as desired. The caveat in this approach is that the quantized version of the channel becomes a hidden Markov process. This precludes the application of various results and techniques including the Chapman-Kolmogorov equation of Section B and the large-deviation principle for Markov fluid processes. These limitations and the vast literature on Markov-modulated processes explain our early adoption of the Gilbert-Elliott model in Section A.

In this section, we use numerical simulations to assess the validity of the Gilbert-Elliott channel model in approximating the behavior of a Markov Raleigh fading channel. Since most of our results are based on the equilibrium distributions of queues, we compare the analytic probability of buffer overflow for the Gilbert-Elliott channel with the empirical distribution of the queue length for the Markov Raleigh fading channel. Recall that the probability of buffer overflow for the Gilbert-Elliott

channel is given by

$$\Pr\{\mathrm{L} > \mathrm{x}\} = \frac{\mathrm{R}}{\mathrm{R}-\mathrm{a}}\left(1-\mathrm{e}^{-\eta^2}\right)\exp\left(\frac{\kappa\mathrm{a}-\mathrm{R}\kappa\mathrm{e}^{-\eta^2}}{\mathrm{a}(\mathrm{R}-\mathrm{a})}\mathrm{x}\right).$$

Fig. 10 shows $\Pr\{\mathrm{L} > \mathrm{x}\}$ for the Gilbert-Elliott channel along with the empirically measured probabilities of buffer overflow for the Markov Raleigh fading model. As

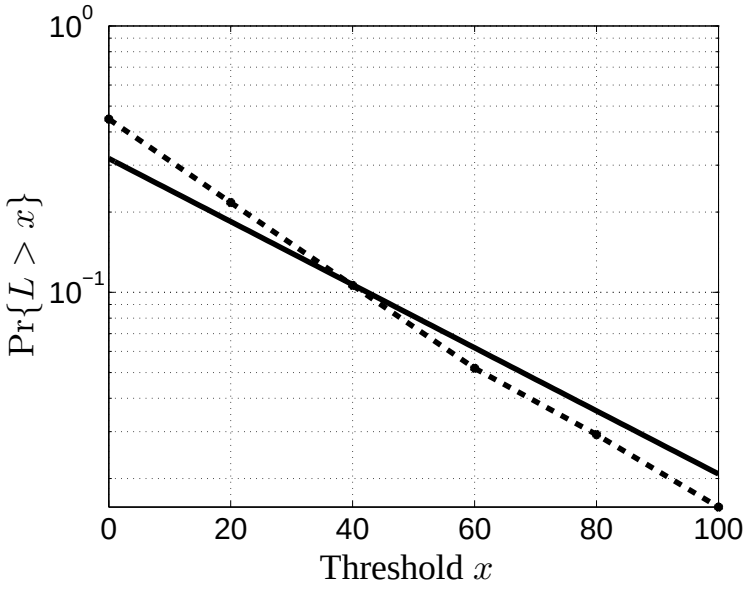


Fig. 10. Comparison of $\Pr\{\mathrm{L} > \mathrm{x}\}$ for the two-state Markov channel (solid) along with the empirically measure probabilities of buffer overflow for the Markov Raleigh fading model (dashed) for decay parameters $\kappa = 10^3$.

seen on the graph, there is a noticeable difference between the two systems. Nevertheless, the exponential decay rate associated with the Gilbert-Elliott channel seems to provide an upper bound for the decay rate of the Orstein-Uhlenbeck system. This is encouraging as the Gilbert-Elliott model appears to provide a conservative measure of effective capacity. Note also that the probability of buffer overflow will converge in distribution as $\kappa \uparrow \infty$. Thus, the Gilbert-Elliott model is an accurate representation

of the Orstein-Uhlenbeck system when the channel varies rapidly.

## E. Discussion

We considered the allocation of system resources in the context of wireless communications under quality of service constraints. The service metric is defined in Chapter II and it is used to conduct a performance analysis of the Gilbert-Elliott system as a function of physical resources. The effective capacity is found to decay rapidly as a function of service constraint $\theta > 0$. System resource should be allocated according to the service constraint. For example, a more stringent constraint on $\theta$ lowers the optimal coderate $R$ of the wireless system. An arrival rate $a$ can only be sustained if $\theta < \kappa/a$. This is somewhat surprising. With an unlimited spectral bandwidth and power budget, only a finite rate can be supported for a stringent service constraint. In addition, this bound is independent of actual coderate $R$ used in the system. This fact is in stark contrast with throughput, which goes to infinity as power and spectral bandwidth grow unbounded. It underlines the fundamental difference between Shannon capacity and delay-sensitive communications. Such insights are crucial for developing efficient wireless communication networks.

The numerical analysis section suggests that alternative Markov models for the underlying wireless channel should be explored. For instance, a finite-state Markov model can be used to represent the channel itself, rather than modeling the ability of the decoder to recover data reliably. The performance evaluation method presented in this work provides, nonetheless, an elegant framework to quantify the amount of physical resources necessary to support rate $a$ under service constraint $\theta$. Alternatively, this framework can be used in conjunction with the effective capacity to characterize the maximal arrival rate $a$ subject to specific resource and service constraints.

# CHAPTER IV

## WIRELESS USER-COOPERATION NETWORKS

The concept of user cooperation was first introduced by Sendonaris, Erkip, and Aazhang [24]. In a companion paper, they discuss implementation issues and provide a performance analysis for practical systems [51]. Virtual antenna arrays formed through user cooperation can be employed to increase the diversity or the spatial multiplexing gain of a communication system in a manner similar to a standard MIMO system. Research in this area mainly falls into three categories: designing coding techniques to improve the diversity of a system, deriving transmission strategies that increase the multiplexing gain of a system, or studying communication schemes that meet the optimal diversity-multiplexing trade-off curve [52]. In [53], transmission protocols for cooperative communications are classified into different approaches. The performance of each protocol is analyzed in terms of outage probability. This work has encouraged coding theorists to develop efficient error-control codes for user cooperation. From a coding theory perspective the work of Laneman et al. [54] employs repetition coding across time. Once a first user has transmitted its data, a second user simply retransmits the data it has received to the base station. More sophisticated coding schemes including *coded cooperation* [55] and *space-time cooperation scheme* [53] have been proposed.

User-cooperation techniques have also been employed to enlarge the achievable rate regions of wireless communication networks. One particular example where user-cooperation can be used is for wireless networks where several relay nodes are incorporated in the network to help improve the transmission rates of certain users. The literature on user-cooperation concepts applied to relay networks is rich. Comprehensive discussions on the subject are provided by Kramer et al. [56], Høst-Madsen [57], and

Khojastepour et al. [58]. Recent results by Azarian, El Gamal and Schniter [59] show that the diversity-multiplexing trade-off methodology can be extended to cooperative networks. Furthermore, the decode-and-forward transmission strategy achieves optimal performance.

Most of the existing work on user-cooperation focuses on improving peer-to-peer link quality while considering other users as relays. In this chapter, we explore the joint benefits that user-cooperation may offer to all the users present in a system. Specifically, we characterize the achievable rate region of a simple user-cooperation scheme as a function of statistical service requirements. The cooperative scheme is shown to significantly enlarge the achievable rate region of the service constrained communication system, provided that the quality of the wireless link between cooperating users is better than the individual connections from the users to their intended destination. Furthermore, the gain of the user-cooperation network increases as the service requirement becomes more and more stringent.

The remainder of the chapter is organized as follows. Section A presents the system model we adopt, along with a precise problem formulation. The proposed user-cooperation scheme is also introduced. Section B contains a derivation of the equilibrium queue-length distribution for the underlying communication system. This distribution is used to compute the statistical performance measure $\theta$ associated with the system introduced in Chapter II. This metric allows us to characterize the achievable rate-region of the cooperative scheme under study for an arbitrary service constraint $\theta_0$ in Section C. Generalizations of the system model considered in this chapter are compared and contrasted in Section D.

## A. System Model

Consider a wireless communication system where two users collaborate to transmit their respective data to a common destination, as shown in Fig. 11. The system is

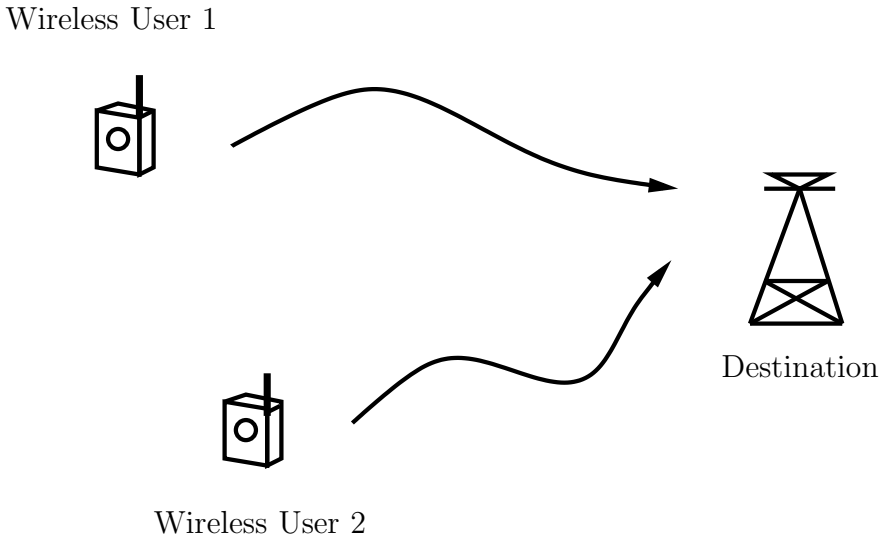


Fig. 11. Abstract model for a cooperative system with two users.

assumed to operate in a frequency division multiplexing (FDM) mode. Each wireless user is subject to a mean power constraint and a finite spectral bandwidth allocation. A large buffer is available at every transmitter where outgoing packets are stored before being sent to their destination. Furthermore, we assume that the system must satisfy a global service constraint specified by $\theta_0$. That is, the asymptotic decay-rates of buffer occupancy, $\theta_1$ for user 1 and $\theta_2$ for user 2, must satisfy $\min(\theta_1, \theta_2) \geq \theta_0$. Finally, we assume that channel state information is not available at the transmitters, although the channel statistics are. In practice, it is often costly for a transmitter to acquire accurate channel state information. This explains why we focus on the situation where channel state information is available only at the receiver, not at the transmitter.

Let $a_i(t)$ denote the instantaneous arrival rate of user $i$ at time $t$. Many real-time traffic sources such as voice, instant messaging, and wireless sensors can be accurately represented by on-off sources [43]. As such, we model $a_i(t)$ using a two-state Markov-modulated fluid process. We remark that a constant source can be viewed as a limiting case of an on-off source where the off-time approaches zero. For an on-off model, the instantaneous arrival rate of user $i$ is $a_i > 0$ when the source is *on*, and zero otherwise. The arrival drift matrix for wireless user $i$ can then be written as

$$D_{ai} = \begin{bmatrix} 0 & 0 \\ 0 & a_i \end{bmatrix}.$$

We denote the mean off-time of this user by $\lambda_{ai}^{-1}$; and its mean on-time, by $\mu_{ai}^{-1}$. The generator matrices for the underlying continuous-time Markov chain of the arrival processes can then be expressed as

$$Q_{ai} = \begin{bmatrix} -\lambda_{ai} & \lambda_{ai} \\ \mu_{ai} & -\mu_{ai} \end{bmatrix}, \quad i = 1, 2.$$

In the situation where users do not cooperate, each wireless device transmits its data independently based on its allocated bandwidth and power budget. The connection of each user can therefore be modeled as a single-server queue, where the arrival process represents the data produced by the user and the service process is determined by the information received at the destination. As in the case presented in Chapter III, we assume that the receiver have the ability to acknowledge reception of the transmitted packets and the wireless system can be modeled as a single server queue. We emphasize, again, that the links between the users and their destination are orthogonal in a frequency-division multiplexing system. In this case, a two-user system can effectively be decomposed into two independent point-to-point systems.

More specifically, the maximum arrival rate that each user can support under the service constraint $\theta_0$ can be obtained separately.

In a typical wireless environment, channel conditions vary with location and time. As such, the maximum throughput of two different users may be vastly asymmetric. For real-time traffic subject to stringent service constraints, this imbalance can be even larger. The goal of this chapter is to design a system where cooperation among users enables them to share system resources equitably. This is accomplished by designing a communication strategy that enlarges their collective achievable rate-region under various service requirements. An expanded rate-region creates the flexibility necessary to share system resources fairly among users.

To take advantage of their mutual wireless links, the two users must first exchange data. In the proposed user-cooperation scheme, we allow each user to apply part of its own power and bandwidth to the exchange of information with its counterpart, as shown in Fig. 12. We represent the fraction of physical resources employed by user $i$

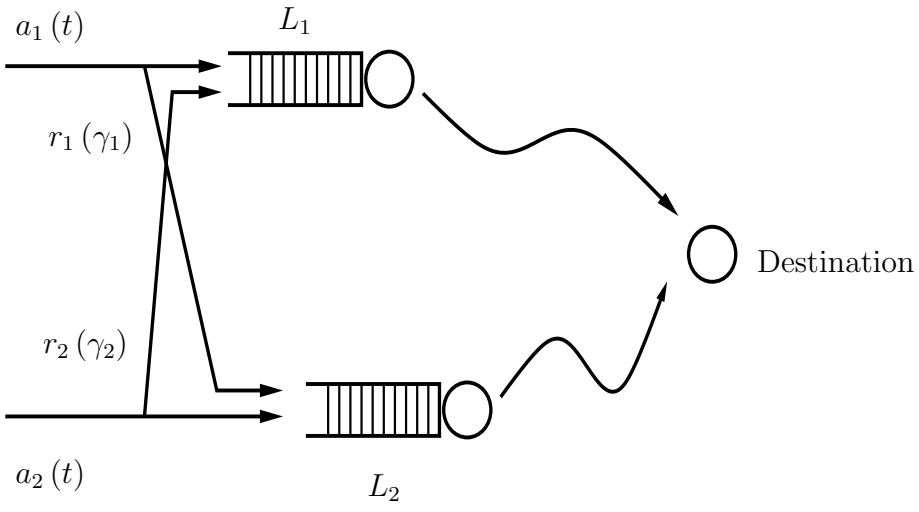


Fig. 12. User-cooperation scheme with two users.

to maintain communication with its peer by $\gamma_i$, and we let the capacity of the newly

created inter-user channel be denoted by $r_i(\gamma_i)$. We consider the specific scenario where the inter-user links are symmetric additive white Gaussian noise (AWGN) channels with constant gains. Thus, when generating traffic, user 1 sends data at rate $r_1(\gamma_1)$ to user 2, and stores the remaining data in its own buffer. User 2 follows a similar procedure, sending part of its data to user 1 and storing excess data locally whenever active. Based on the respective values of $\gamma_1$ and $\gamma_2$, we can characterize the achievable rate-region for the cooperative system of Fig. 12 under an arbitrary service requirement $\theta_0$. The union of these rate-regions over all admissible pairs $(\gamma_1, \gamma_2) \in [0, 1]^2$ yields an achievable rate-region for the proposed user-cooperation scheme. We denote this region by $\mathcal{R}(\theta_0)$, and point out that it is a function of the service requirement $\theta_0$. As $\theta_0 \to 0$, this achievable region converges to the stability region of the system, which is characterized by its throughput optimal boundary.

## B. Queueing Performance Analysis

To relate the effects of the physical layer to the performance of the network, we must first understand the queueing dynamics that govern the system. We start by considering a single server fluid queue with independent arrival and service processes. We assume that these two processes are coupled by a buffer of infinite length, and that they are modulated by finite-state continuous-time Markov chains with irreducible generator matrices $Q_a \in \mathbb{R}^{M\times M}$ and $Q_s \in \mathbb{R}^{N\times N}$, respectively. We let $a_m$ represent the instantaneous arrival rate when the arrival process is in state $m$, and we write $p_a(t; m)$ to denote the probability of occurrence of this event at any given time $t \geq 0$. Similarly, the offered service has instantaneous rate $R_n$ when the underlying process is in state $n$, and $p_s(t; n)$ represents the probability of being in this state at time $t$. We can write the arrival and the service drift matrices as $D_a = \text{diag}(a_1, \ldots, a_M)$

and $D_s = \text{diag}(R_1, \ldots, R_N)$, respectively. In vector form, the arrival and service probability distributions become

$$p_a(t) = (p_a(t;1), p_a(t;2), \ldots, p_a(t;M))$$
$$p_s(t) = (p_s(t;1), p_s(t;2), \ldots, p_s(t;N)) .$$

Using these definitions, we can write the evolution of the probability vectors in a compact fashion,

$$\frac{d}{dt}p_a(t) = p_a(t)Q_a$$
$$\frac{d}{dt}p_s(t) = p_s(t)Q_s.$$

We use the vectors $w_a$ and $w_s$ to represent the steady-state distributions of the arrival and service processes, with $w_aQ_a = w_sQ_s = 0$.

For the combined arrival and service process, let

$$X_t \in \{(m,n) : 1 \leq m \leq M, 1 \leq n \leq N\}$$

be the situation where the arrival is in state $m$ and the offered service is in state $n$ at time $t$. The probability of this event is simply equal to $p(t;m,n) = p_a(t;m)p_s(t;n)$. We employ $p(t)$ to denote the vector consisting of the elements $\{p(t;m,n)\}$ in lexicographic order. It follows that $p(t) = p_a(t) \otimes p_s(t)$, where $\otimes$ is the Kronecker product [60, 41]. The joint probability vector $p(t)$ satisfies

$$\frac{d}{dt}p(t) = p(t)Q,$$

where $Q$ is the generator matrix of the joint process. This matrix can be written as

$$Q = Q_a \otimes I_N + I_M \otimes Q_s \tag{4.1}$$

where $I_K$ is a $K \times K$ identity matrix. The matrix $Q$ is recurrent and irreducible, and $w = w_a \otimes w_s$ is the steady-state distribution for the aggregate process. The net drift matrix $D$ of the joint process is

$$D = D_a \otimes I_N - I_M \otimes D_s. \tag{4.2}$$

Let $L_t$ represent the queue-length of the user at time $t$. The evolution of $L_t$ in time can be expressed as [61, 62]

$$\frac{d}{dt}L_t = (a_m - R_n)\ \mathbf{1}_{\{L_t>0\}} + (a_m - R_n)^+ \mathbf{1}_{\{L_t=0\}}, \tag{4.3}$$

which is a stochastic differential equation on the Markov process $(L_t, X_t)$. Define the event probability

$$F(x, m, n, t) = \Pr\{X_t = (m, n), L_t \leq x\},$$

and let $F(x, t)$ be the lexicographic arrangement of $\{F(x, m, n, t)\}$. Using this notation, we can write the Chapman-Kolmogorov forward equation in matrix form as [61, 13, 41, 62]

$$\frac{\partial}{\partial t}F + \frac{\partial}{\partial x}FD = FQ.$$

The mean arrival rate and mean service rate are given by $\bar{a} = \langle w_a D_a, \mathbf{1}\rangle$ and $\bar{R} = \langle w_s D_s, \mathbf{1}\rangle$, respectively. If the system is stable (i.e., $\bar{a} < \bar{R}$), then the underlying Markov process is positive recurrent [61]. As such, there exists a steady-state distribution for the aggregate process $(L_t, X_t)$ [38]. Let $\pi(x, m, n)$ denote the steady-state queue-length distribution of the buffer, with

$$\frac{\partial}{\partial x}\pi(x)\, D = \pi(x)\, Q.$$

Since $\pi(x)$ is a bounded solution, it has spectral representation

$$\pi(x) = w - \sum_{l=1}^{k} \alpha_l \phi_l e^{z_l x}, \tag{4.4}$$

where $\{(\phi_l, z_l) : \mathrm{Re}(z_l) \leq 0\}$ are $k$ eigenvector/eigenvalue pairs that satisfy the eigenvalue problem

$$z\phi D = \phi Q. \tag{4.5}$$

We emphasize that $\alpha_l = 0$ for any $l$ such that $\mathrm{Re}\{z_l\} > 0$ because the system is stable [41]. Thus, we only need to consider eigenvalues with negative real parts. Note that the system is subject to the boundary conditions $\pi(0, m, n) = 0$ whenever $a_m - R_n > 0$. If $a_m \neq R_n$ for all $m$ and $n$, then there are exactly $k$ such boundary conditions and the steady-state distribution is unique [41].

Solving the eigenvalue problem of (4.5) for the whole system can be somewhat involved. However, taking advantage of the special structure of $D$ and $Q$, we can decompose the original system and reduce the complexity of the problem.

**Lemma 4** *For any eigenvector/eigenvalue pair $(\phi_l, z_l)$ that satisfies $z_l \phi_l D = \phi_l Q$, there exist $\phi_{a_l}$, $\phi_{s_l}$, and $\nu \in \mathbb{C}$ such that*

$$z_l \phi_{a_l} (D_a - \nu I_M) = \phi_{a_l} Q_a, \tag{4.6}$$

$$z_l \phi_{s_l} (\nu I_N - D_s) = \phi_{s_l} Q_s. \tag{4.7}$$

*Proof*: For $z_l = 0$, the result is trivial. For any $\nu$, the vector $\phi = \phi_{a_l} \otimes \phi_{s_l} = w_a \otimes w_s$ satisfies (4.6) and (4.7). Assume $z_l \neq 0$, then $z_l \phi_l D = \phi_l Q$ is equivalent to

$$\phi_l \left( D - \frac{Q}{z_l} \right) = 0. \tag{4.8}$$

Substituting (4.1) and (4.2) into (4.8), we obtain

$$\phi_l\left( (D_a \otimes I_N - I_M \otimes D_s) - \left( \frac{Q_a}{z_l} \otimes I_N + I_M \otimes \frac{Q_s}{z_l} \right) \right) = 0,$$

which can be rewritten as

$$\phi_l \left( \left( D_a - \frac{Q_a}{z_l} \right) \otimes I_N + I_M \otimes \left( -D_s - \frac{Q_s}{z_l} \right) \right) = 0.$$

The above equation shows that zero is an eigenvalue of the matrix

$$\left( D_a - \frac{Q_a}{z_l} \right) \otimes I_N + I_M \otimes \left( -D_s - \frac{Q_s}{z_l} \right).$$

According to [60, page 268], zero is an eigenvalue of the above matrix if and only if there exists $\nu \in \mathbb{C}$ such that

$$\begin{aligned} \nu &\in \sigma \left( D_a - \frac{Q_a}{z_l} \right) \\ -\nu &\in \sigma \left( -D_s - \frac{Q_s}{z_l} \right) \end{aligned} \tag{4.9}$$

where $\sigma(A)$ denotes the *spectrum* of matrix $A$. Expression (4.9) is equivalent to stating that there exist $\phi_{a_l}$ and $\phi_{s_l}$ such that

$$\begin{aligned} \phi_{a_l} \left( D_a - \frac{Q_a}{z_l} - \nu I_M \right) &= 0, \\ \phi_{s_l} \left( \nu I_N - D_s - \frac{Q_s}{z_l} \right) &= 0. \qquad \square \end{aligned}$$

Lemma 4 allows us to solve (4.5) by decomposing the original system into two subsystems: the arrival subsystem of (4.6) that features a Markov-modulated arrival process and a constant service rate $\nu$; and the subsystem described in (4.7) with a constant arrival rate $\nu$ and a Markov-modulated fluid service process. A similar decomposition argument can be found in [41] where (4.6) and (4.7) are shown to be sufficient conditions for $z_l$ to be a solution to (4.5). Lemma 4 provides both necessary

and sufficient conditions for this decomposition to exist.

Based on the equilibrium queue-length distribution of the buffer (4.4), a number of performance metrics can be computed as in Chapter III. A simple and important one is the probability of buffer overflow, which can be expressed as

$$\Pr\{L > x\} = 1 - \langle \pi(x), \mathbf{1} \rangle.$$

As shown in the previous chapter, the probability of buffer overflow is related to the delay violation probability of a communication system and hence can be treated as a service requirement for delay-sensitive applications over wireless networks. In practice, buffers are often large and their decay rates of buffer overflow probabilities are determined primarily by the large deviation principle governing each buffer occupancy. From this perspective, the service constraint $\theta$ for the Markov-modulated fluid process becomes

$$\begin{aligned} \theta &= -\lim_{x\to\infty} \frac{\log \Pr\{L > x\}}{x} = -\lim_{x\to\infty} \frac{\log\left(1 - \langle \pi(x), \mathbf{1}\rangle\right)}{x} \\ &= -\lim_{x\to\infty} \frac{\log\left(\sum_{l=1}^{k} \alpha_l \langle \phi_l, \mathbf{1}\rangle e^{z_l x}\right)}{x} = -\max_{l\in\{1,\ldots,k\}} \mathrm{Re}\{z_l\}. \end{aligned} \tag{4.10}$$

In other words, the service requirement $\theta$ of the system is the absolute value of the largest negative real eigenvalue satisfying (4.5).

Let the absolute values of the maximum negative eigenvalues for the aggregate system and its individual components be denoted by

$$\begin{aligned} \theta &= -\max\{\mathrm{Re}\{z\} < 0 : \det(zD - Q) = 0\} \\ \theta_a(\nu) &= -\max\{\mathrm{Re}\{z\} < 0 : \det(zD_a - z\nu I_M - Q_a) = 0\} \\ \theta_s(\nu) &= -\max\{\mathrm{Re}\{z\} < 0 : \det(z\nu I_N - zD_s - Q_s) = 0\}. \end{aligned}$$

It is well-known [63, 64, 65] that for an irreducible generator matrix $Q_a$ and a real

positive diagonal matrix $D_a$, $\theta_a(\nu)$ is continuous and monotonically increasing from zero to infinity as $\nu$ ranges from the mean rate to the peak rate, i.e. $\nu \in [\bar{a}, \max_m a_m]$. Similarly, $\theta_s(\nu)$ is continuous and monotonically decreasing from infinity to zero for $\nu \in \left[\min_n R_n, \bar{R}\right]$. If $\max_m a_m > \min_n R_n$ and $\bar{a} < \bar{R}$ then there exists a $\nu^* \in [\bar{a}, \max_m a_m]$ such that $\theta_a(\nu^*) = \theta_s(\nu^*)$, as illustrated in Fig. 13. On the other hand, if $\max_m a_m \leq \min_n R_n$, the buffer is always empty and hence $\theta = \infty$. The following theorem asserts that, for a stable system, the large deviation principle associated with the joint system is identical to that governing the two subsystems with parameter $\nu^*$.

**Theorem 1** *Let $Q_a$ and $Q_s$ be irreducible, recurrent generator matrices, and let $D_a$ and $D_s$ be non-negative diagonal matrices. If the system is stable (i.e., $\bar{a} < \bar{R}$), then there exists a $\nu^* \in \left[\bar{a}, \bar{R}\right]$ such that*

$$\theta = \theta_a(\nu^*) = \theta_s(\nu^*).$$

*Proof*: Denote the value where these two functions meet by $\theta^* = \theta_a(\nu^*) = \theta_s(\nu^*)$. Clearly, $\nu^* \in \left[\bar{a}, \bar{R}\right]$. We need to show that $\theta^* = \theta$. Assume not, then $\theta < \theta^*$ by the minimality of $\theta$. In addition, lemma 4 implies that there exists a $\nu_0 \in \mathbb{C}$ such that $z_0$ is an eigenvalue of both decoupled systems and $\theta = \mathrm{Re}\{z_0\}$. It follows from the minimality of $\theta_a(\nu)$ and $\theta_s(\nu)$ that $\theta_a(\nu_0) \leq \theta < \theta_a(\nu^*)$ and $\theta_s(\nu_0) \leq \theta < \theta_s(\nu^*)$. From the monotonicity of $\theta_a(\nu)$, we conclude that $\nu_0 < \nu^*$. However, from the monotonicity of $\theta_s(\nu)$, we get $\nu_0 > \nu^*$. This is a contradiction. We then conclude that $\theta = \theta^*$. □

From theorem 1, we find that once $\nu^*$ is determined, the service metric $\theta$ of the system can be obtained by analyzing the behavior of the two independent subsystems. Define

$$\beta(\theta) = \theta_a^{-1}(\theta)$$
$$\alpha(\theta) = \theta_s^{-1}(\theta).$$

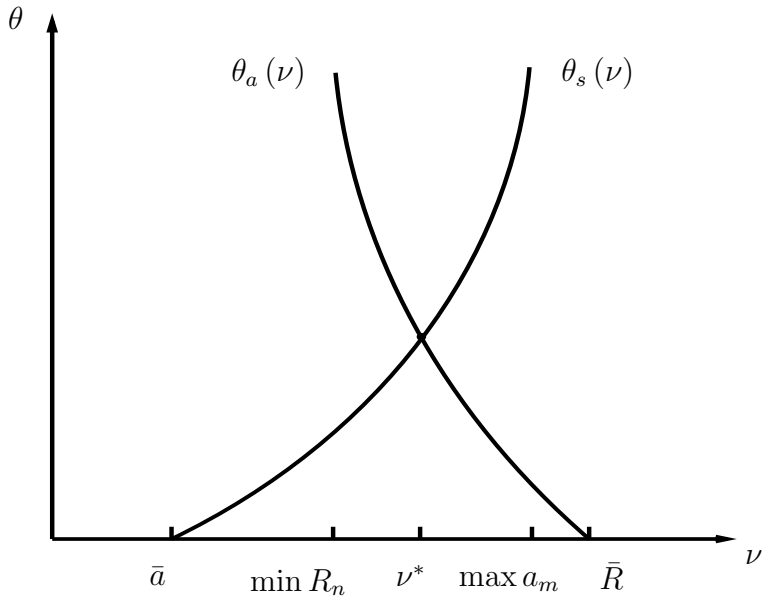


Fig. 13. $\theta_a(\nu)$ and $\theta_s(\nu)$ as a function of $\nu$.

For a specific service constraint $\theta^*$, $\beta(\theta^*)$ is the effective bandwidth of the arrival process [17], and $\alpha(\theta^*)$ is the effective capacity of the service process [19]. Under the conditions of Theorem 1, the parameter $\theta^*$ is the unique solution to the equation

$$\nu^* = \beta(\theta^*) = \alpha(\theta^*). \tag{4.11}$$

Note that in (4.11), $\nu^*$ is the effective bandwidth of the arrival process and the effective capacity of the service process under the service constraint $\theta^*$. A service constraint $\theta_0$ for the aggregate system is said to be achievable if and only if $\theta^* \geq \theta_0$. Since $\beta(\theta)$ is monotonically increasing in $\theta$ and $\alpha(\theta)$ is monotonically decreasing, the service constraint $\theta_0$ can be fulfilled if and only if

$$\beta(\theta_0) \leq \alpha(\theta_0). \tag{4.12}$$

C. Achievable Rate-Regions for a Two-User System

In this section, we characterize the achievable rate-region of the user-cooperation system depicted in Fig. 12 when operating under service constraint $\theta_0 > 0$. Because the system employs frequency-division multiplexing, the wireless links between the users and their common destination can be modeled as independent Gilbert-Elliott channels. Assume that both wireless channels have the same expected power gain. The generator matrix $Q_{si}$ corresponding to the modulating Markov process of user $i$ is given by

$$Q_{si} = \begin{bmatrix} -\kappa_i e^{-\eta_i^2} & \kappa_i e^{-\eta_i^2} \\ \kappa_i - \kappa_i e^{-\eta_i^2} & -\kappa_i + \kappa_i e^{-\eta_i^2} \end{bmatrix};$$

the drift matrix $D_{si}$ is equal to

$$D_{si} = \begin{bmatrix} 0 & 0 \\ 0 & R_i \end{bmatrix}$$

where $\kappa_i$ denotes the exponential decay rate of the channel of user $i$, and $R_i$ and $\eta_i$ are respectively the selected coderate and decoding threshold of that same user, as defined in (3.4). When the two users do not cooperate, user $i$ sets up a wireless connection to the destination using its own physical resources, power $P_i$ and spectral bandwidth allocation $W_i$. The effective bandwidth of source $i$, as described in Section A can then be expressed as [18]

$$\beta_i(\theta_0, a_i) = \frac{\theta_0 a_i - \lambda_{ai} - \mu_{ai} + \sqrt{(\theta_0 a_i - \lambda_{ai} - \mu_{ai})^2 + 4\theta_0 a_i \lambda_{ai}}}{2\theta_0}, \tag{4.13}$$

where $a_i$ denotes the peak rate of the underlying on-off source. Similarly, the effective capacity of the wireless channel of user $i$ is [23]

$$\alpha_i(\theta_0, W_i, P_i) = \max_{R_i} \left\{ \frac{\theta_0 R_i + \kappa_i - \sqrt{(\theta_0 R_i + \kappa_i)^2 - 4\theta_0 R_i \kappa_i e^{-\eta_i^2}}}{2\theta_0} \right\}.$$

Recall that the value of $\eta_i$ depends implicitly on $P_i$, $W_i$, and $R_i$, as seen in (3.4). According to (4.12), the peak rate pair $(a_1, a_2)$ is achievable under the QoS parameter $\theta_0$ if and only if the effective bandwidth of the traffic generated by user $i$ is less than or equal to the effective capacity of the corresponding wireless channel, i.e.,

$$\beta_i(\theta_0, a_i) \leq \alpha_i(\theta_0, W_i, P_i) \tag{4.14}$$

for $i = 1, 2$. Since the wireless channels are orthogonal, using (4.13) and (4.14) we can solve for the maximum supported arrival peak-rate $a_i^*$ for user $i$ subject to the service constraint $\theta_0$. The achievable rate-region of the non-cooperative FDM system is in the form of a rectangle limited by the maximum supported peak-rates of the two links under the parameter $\theta_0$,

$$a_i \leq a_i^* = \alpha_i(\theta_0, W_i, P_i) \left( 1 + \frac{\mu_{ai}}{\theta_0 \alpha_i(\theta_0, W_i, P_i) + \lambda_{ai}} \right). \tag{4.15}$$

Now, consider the situation where the two wireless users cooperate by taking advantage of the AWGN inter-user channels. We assume that user $i$ assigns a fraction $\gamma_i$ of its power and bandwidth to the exchange of information with its counterpart. If the expected gain of the inter-user channel is $G$, then its Shannon capacity is given by

$$r_i(\gamma_i) = \gamma_i W_i \log \left( 1 + \frac{G P_i}{N_0 W_i} \right).$$

The power and bandwidth remaining for the uplink connection between user $i$ and the destination become $(1 - \gamma_i)P_i$ and $(1 - \gamma_i)W_i$ respectively. The effective capacity

for the resulting wireless channel can be expressed as

$$\nu_i(\gamma_i) = \alpha_i(\theta_0, (1-\gamma_i)W_i, (1-\gamma_i)P_i).$$

It is clear from the system model described in Section A that the inter-user traffic originating from user $i$ is an on-off process with peak-rate $r_i(\gamma_i)$. This traffic is modulated by the same two-state Markov chain that modulates the original source. Therefore, the effective bandwidth of this traffic can be expressed as $\beta_i(\theta_0, r_i(\gamma_i))$. Similarly, the portion of the traffic generated by user $i$ which is stored locally and sent directly to the destination is an on-off process with peak-rate $a_i - r_i(\gamma_i)$. The effective bandwidth of this local traffic then becomes $\beta_i(\theta_0, a_i - r_i(\gamma_i))$.

Independence of the traffic generated by the two users and the additivity property of the effective bandwidth for independent sources [13] imply that the total effective bandwidth of the input process to buffer $i$ is the sum of the effective bandwidths of the local traffic and the inter-user traffic coming from its counterpart. Equation (4.12) states that a service constraint $\theta_0$ is achievable if and only if the total effective bandwidth of the incoming traffic is smaller than the effective capacity of the offered service. In the present case, this condition yields two inequalities

$$\begin{aligned} \beta_1(\theta_0, a_1 - r(\gamma_1)) + \beta_2(\theta_0, r(\gamma_2)) &\leq \nu_1(\gamma_1) \\ \beta_2(\theta_0, a_2 - r(\gamma_2)) + \beta_1(\theta_0, r(\gamma_1)) &\leq \nu_2(\gamma_2). \end{aligned} \tag{4.16}$$

Since $\beta_1(\theta_0, a_1 - r(\gamma_1))$ and $\beta_2(\theta_0, a_2 - r(\gamma_2))$ are both non-negative, the values of the parameter pair $(\gamma_1, \gamma_2)$ are further constrained by

$$\beta_2(\theta_0, r(\gamma_2)) \leq \nu_1(\gamma_1)$$

$$\beta_1(\theta_0, r(\gamma_1)) \leq \nu_2(\gamma_2).$$

Let $\mathcal{C}$ denote the set of pairs of the form $(\gamma_1, \gamma_2)$ for which the above inequalities hold.

For any $(\gamma_1, \gamma_2) \in \mathcal{C}$, the achievable rate-region of the cooperative system, which we denote by $\mathcal{R}(\theta_0, \gamma_1, \gamma_2)$, is found to be

$$a_1 \leq r_1(\gamma_1) + \big(\nu_1(\gamma_1) - \beta_2\big(\theta_0, r_2(\gamma_2)\big)\big) \times \left(1 + \frac{\mu_{a1}}{\theta_0\left(\nu_1(\gamma_1) - \beta_2\big(\theta_0, r_2(\gamma_2)\big)\right) + \lambda_{a1}}\right)$$
$$a_2 \leq r_2(\gamma_2) + \big(\nu_2(\gamma_2) - \beta_1\big(\theta_0, r_1(\gamma_1)\big)\big) \times \left(1 + \frac{\mu_{a2}}{\theta_0\left(\nu_2(\gamma_2) - \beta_1\big(\theta_0, r_1(\gamma_1)\big)\right) + \lambda_{a2}}\right). \tag{4.17}$$

The achievable rate-region of the user-cooperation scheme under service constraint $\theta_0$ is then given by

$$\mathcal{R}(\theta_0) = \bigcup_{(\gamma_1,\gamma_2)\in\mathcal{C}} \mathcal{R}(\theta_0, \gamma_1, \gamma_2).$$

Note that the achievable rate-region of the non-cooperative system is $\mathcal{R}(\theta_0, 0, 0)$. It is therefore a subset of $\mathcal{R}(\theta_0)$. The boundary of $\mathcal{R}(\theta_0)$ can be obtained by maximizing $a_2$ over $(\gamma_1, \gamma_2)$ while keeping $a_1$ fixed in (4.17). The solution of the boundary problem can be obtained by standard optimization techniques such as the Lagrange multiplier method.

The comparison between the achievable rate-regions is illustrated though an example. Numerical values of the parameters for the wireless channels and the arrival processes used in this example appear in Table II. The two wireless channels are assumed to have the same expected power gains. However, the channel of user 2 changes faster than that of user 1, with $\kappa_2 > \kappa_1$. Suppose that the gain of the AWGN inter-user channel is one ($G = 1$). The achievable rate-region of the system under the cooperative scheme is compared to that of the non-cooperative scheme in Fig. 14 and Fig. 15, where the numerical values for $\theta$ are equal to 0.001 and 0.01 respectively.

From these figures, we see that the achievable rate-region of the cooperative system is strictly larger than the region of the traditional FDM system. We note that,

Table II. System parameters for user-cooperation networks.

| | |
|---|---|
| $N_0 = 10^{-6}$ W/Hz | Noise power spectral density |
| $W_1 = W_2 = 11$ MHz | Bandwidth |
| $P_1 = P_2 = 100$ mW | Received power |
| $\kappa_1 = 10^2$ sec$^{-1}$ | Decay parameter of channel 1 |
| $\kappa_2 = 10^3$ sec$^{-1}$ | Decay parameter of channel 2 |
| $\lambda_{a1}^{-1} = \lambda_{a2}^{-1} = 650$ ms | Average silent period |
| $\mu_{a1}^{-1} = \mu_{a2}^{-1} = 352$ ms | Average talk burst |

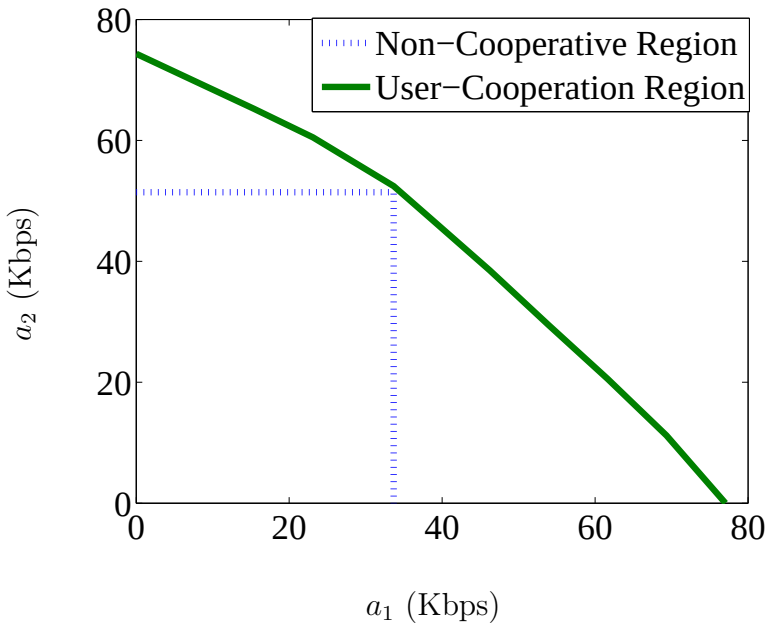


Fig. 14. Comparison of the achievable rate-regions when $\theta = 0.001$.

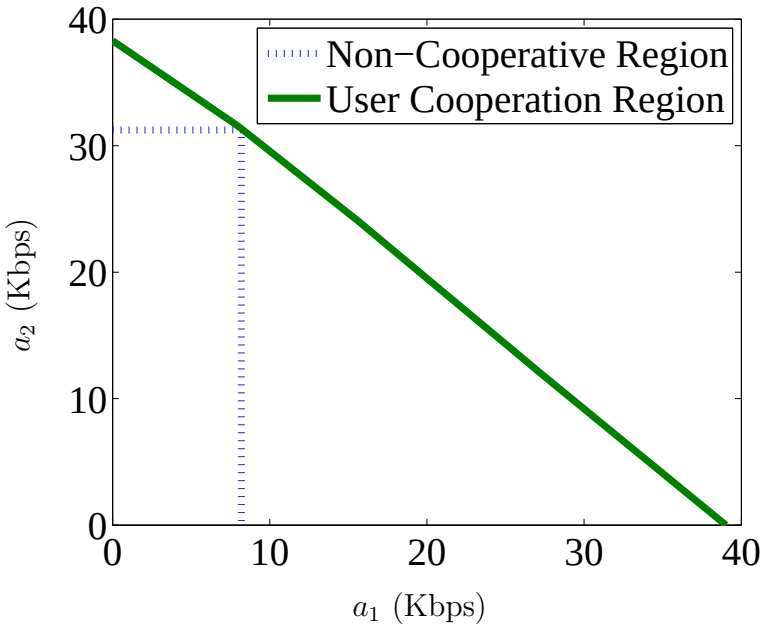


Fig. 15. Comparison of the achievable rate-regions when $\theta = 0.01$.

even though the expected channel gains of the two wireless channels are the same, there is a large imbalance between the maximum supported peak-rates of the two users under service constraint $\theta_0$. This can be explained by the fact that the channel memory of user 2 decays faster than the channel memory of user 1, resulting in a higher order of time-diversity for user 2 [66]. Furthermore, the asymmetry between the maximum achievable rates of the two users increases as the service constraint becomes more stringent. Both figures suggest that, under strict service requirements, user-cooperation provides an efficient means to share radio resources fairly among users.

In the idealized scenario where $G \to \infty$, the users can exchange an arbitrary amount of information at no extra cost in terms of power and bandwidth. Let $r_i$ be the rate at which user $i$ sends information to its counterpart through the inter-user channel when its source is *on*. The effective bandwidth of the inter-user traffic is

$\beta_i(\theta_0, r_i)$, while the effective bandwidth of the excess traffic stored in the local buffer becomes $\beta_i(\theta_0, a_i - r_i)$. From (4.12), we know that the rate-pair $(a_1, a_2)$ is achievable under service constraint $\theta_0$ provided that

$$\begin{aligned} \beta_1(\theta_0, a_1 - r_1) + \beta_2(\theta_0, r_2) &\leq \alpha_1(\theta_0, W_1) \\ \beta_2(\theta_0, a_2 - r_2) + \beta_1(\theta_0, r_1) &\leq \alpha_2(\theta_0, W_2). \end{aligned} \tag{4.18}$$

We note that the effective bandwidth is a concave function, with strict inequality over a non-trivial set of values [13]. There are therefore situations where peak-rates $a_1$ and $a_2$ can be supported through user-cooperation, but not by a traditional FDM system.

These results are easier to understand through an example. For the system parameters listed in Table II, but with $G \to \infty$, the achievable rate-region of the cooperative system is plotted along with that of the non-cooperative system in Fig. 16 for $\theta = 0.001$, and in Fig. 17 for $\theta = 0.01$.

As shown in the figures, user-cooperation provides a significant statistical gain over non-cooperative system in terms of achievable rates. This gain becomes larger as the service constraint becomes more stringent. We can infer from Fig. 16 that the sum peak-rate, $a_1 + a_2$, increases when the two users are cooperating through a perfect inter-user channel, as discussed above. The achievable rate-region of the user-cooperation scheme gets larger as the quality of the inter-user channel improves. Yet this gain is less significant when the service constraint $\theta_0$ becomes large. This can be explained by the fact that, as $\theta_0$ increases, the effective capacity of each variable channel decreases dramatically [23]. Because the throughput of the AWGN inter-user channel does not vary with $\theta_0$, this latter channel behaves more like an idealized channel to the users as the service constraint becomes increasingly stringent. Thus, user-cooperation seems to be beneficial as long as the channel gain of the AWGN

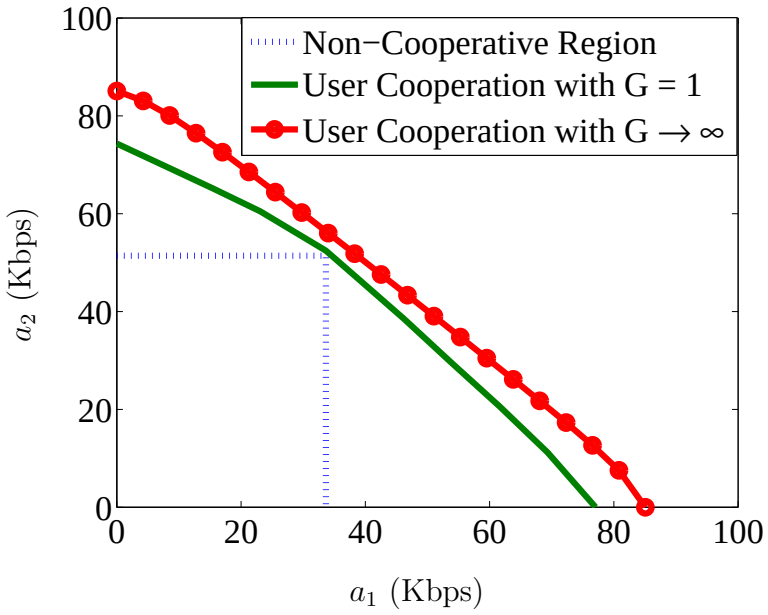


Fig. 16. Comparison of the achievable rate-regions when $\theta_0 = 0.001$.

inter-user channel is adequate.

## D. Alternative System Models

So far we have shown that, under various service constraints, the achievable rate-region of the cooperative strategy is significantly larger than that of the non-cooperative FDM system. The FDM model for the non-cooperative system is employed to circumvent mathematical difficulties that arise from inter-user interference. Moreover, to keep our abstract model simple, we assume that the inter-user traffic is transmitted instantaneously to the other users. That is, there is no buffer associated with the inter-user channel. A valid criticism of our model is that the FDM assumption may unfairly penalize the performance of the non-cooperative system, as compared to its performance when successive interference cancellation is used at the destination. Another observation regarding our model is the fact that having a queue for the inter-

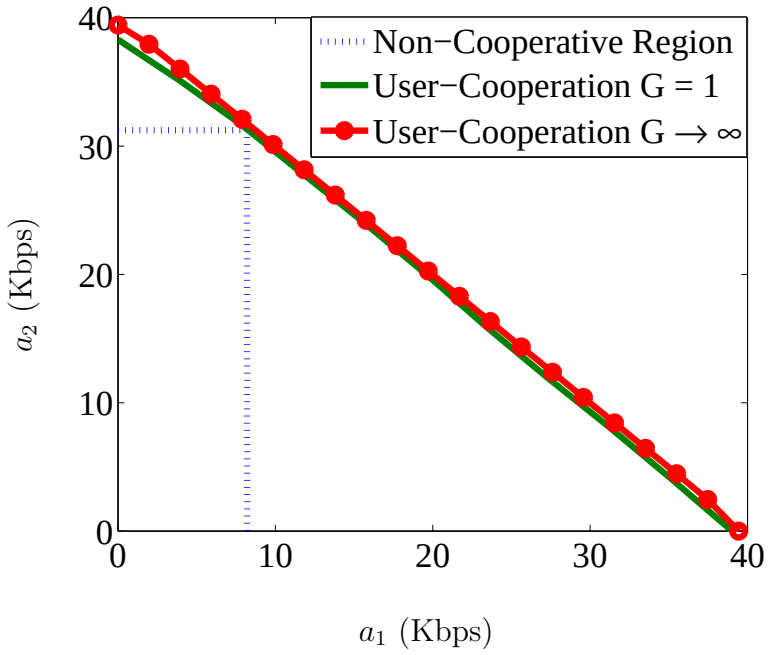


Fig. 17. Comparison of the achievable rate-regions when $\theta_0 = 0.01$.

user channel may improve the performance of the cooperative system. In this section, we consider these more elaborate systems and discuss their impacts on performance analysis. In particular, we argue that the intuition gained from the simpler model holds for these more intricate models as well.

### 1. Successive Interference Cancellation

From [1], we know that the maximum achievable rate-region for the multi-access channel encompasses the region achieved by a FDM system. This greater flexibility is obtained by using successive interference cancellation at the receiver. For any fading realization $(h_1, h_2)$, the achievable rate-region of a multi-access channel is the

polyhedron bounded by the inequalities

$$\begin{aligned} C_1 &\leq W \log_2 \left(1 + \frac{|h_1|^2 P_1}{N_0 W}\right) \\ C_2 &\leq W \log_2 \left(1 + \frac{|h_2|^2 P_2}{N_0 W}\right) \\ C_1 + C_2 &\leq W \log_2 \left(1 + \frac{|h_1|^2 P_1 + |h_2|^2 P_2}{N_0 W}\right). \end{aligned} \tag{4.19}$$

Here, $P_i$ is the mean received power of user $i$ and $W$ is the total spectral bandwidth available to the two users. We note that this region is upper-bounded by the rate-region of a FDM system with twice the spectral bandwidth ($2W$). In particular, consider a FDM system with allocation $W_1 = W_2 = W$, the achievable rate-region of this alternate system is specified by

$$\begin{aligned} C_1 &\leq W \log_2 \left(1 + \frac{|h_1|^2 P_1}{N_0 W}\right) \\ C_2 &\leq W \log_2 \left(1 + \frac{|h_2|^2 P_2}{N_0 W}\right). \end{aligned} \tag{4.20}$$

Clearly, the region defined by (4.19) is a subset of (4.20). Thus, we can upper-bound the rate-region of a non-cooperative multiple-access system that uses successive interference cancellation by that of a FDM system that has double the spectral bandwidth of the original system.

Assume the total system bandwidth is $W = 11$ MHz. We compare the achievable rate-region of the user-cooperation system to the region corresponding to a FDM system with twice the bandwidth in Fig. 18. The latter region is an absolute upper-bound for the region of a non-cooperative system that supports successive interference cancellation. The result suggests that user-cooperation may offer significant gains in performance over a non-cooperative system that uses successive interference cancellation. This behavior is explained, partly, by the fact that additional spectral bandwidth offers diminishing returns in terms of effective capacity. The effective capacity

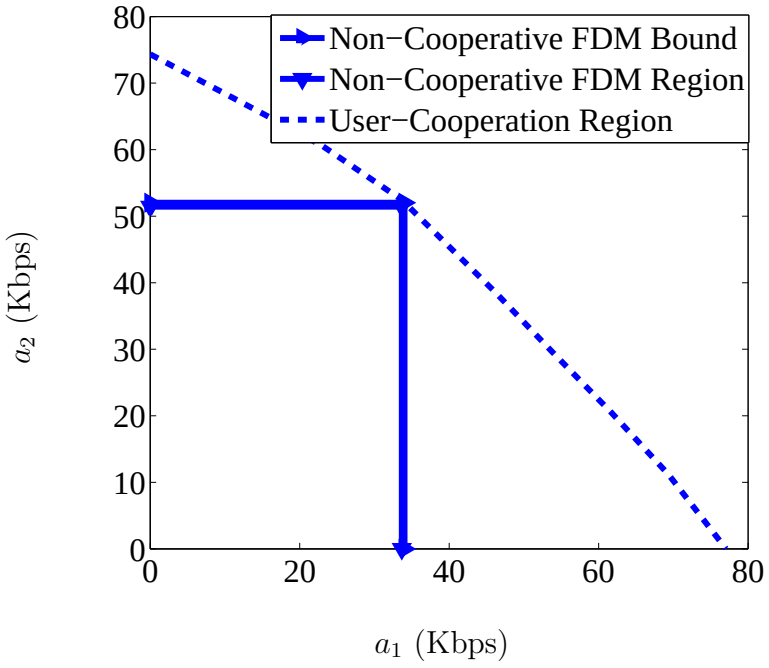


Fig. 18. Comparison of the rate-regions for $\theta_0 = 0.01$.

of a service constrained system appears to level off even before the system enters its information theoretic wideband regime [23].

For the system parameters listed in Table II, the effective capacities $\alpha_i(\theta_0, W, P)$ of the two wireless channels are plotted as a function of spectral bandwidth $W$ in Fig. 19. We can see from the figure that the effective capacities of the two channels level off rapidly once $W$ is large enough. This explains why doubling the spectral bandwidth of a FDM system does not necessarily improve the effective capacity by much. This limitation is also partly due to the underlying assumption that channel state information is not available at the transmitters. Incidentally, users cannot transmit at the (error-free) instantaneous Shannon capacity, and therefore they do not benefit from the additional degrees of freedom associated with a larger spectral bandwidth. When the available spectrum is large enough, the queueing behavior

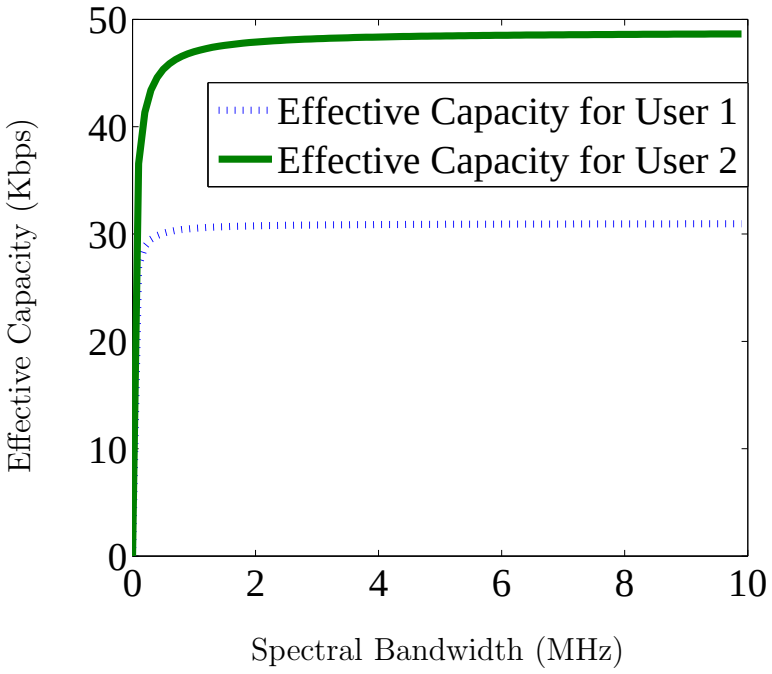


Fig. 19. Effective capacity for $\theta_0 = 0.01$.

of the system is dominated by the holding time of the service *off* state, which is independent of the channel bandwidth [23].

### 2. Cooperation with Inter-User Buffers

A straightforward generalization of the user-cooperation scheme proposed in Section A is to add buffers for the inter-user traffic of both transmitters. In this case, the inter-user traffic can be buffered locally and, as such, data can be sent to the other user even when the source is in its *off* state. This more flexible set-up can only improve system performance and thereby enlarge the achievable rate-region of the user-cooperation system. To characterize the achievable rate-region of this communication scheme, we need to derive the effective bandwidth of the departure process of the inter-user traffic. Since the gain of the inter-user channel is constant, the effective

bandwidth of the inter-user traffic can be analytically characterized [17]. However, we elect not to compute the exact achievable rate-region for the problem at hand when buffers are used for the inter-user channels. Rather, we provide tight upper and lower bounds for the periphery of the achievable rate-region.

The user-cooperation system without inter-user buffers can be thought of as a special case of the cooperative system described above. It corresponds to the situation where the inter-user buffers remain empty at all times. In this sense, the achievable rate-region of the user-cooperation system without inter-user buffers serves as a lower-bound for the achievable rate-region of the user-cooperation system with inter-user buffers. On the other hand, the achievable rate-region of the idealized user-cooperation system ($G \to \infty$) serves as an upper-bound for the user-cooperation system with inter-user buffers. Indeed, as $G$ approaches infinity, the constant service rates of the inter-user channels become increasingly large. This insures that these buffers remain empty. The boundary of the achievable rate-region for the buffered user-cooperation system must lie between the dotted line and the solid line in Fig. 16 and in Fig. 17. Since the gap between the upper and lower bounds is quite narrow, the gains associated with using inter-user buffers for the system under study must be somewhat marginal. The tedious analysis of the more elaborate buffered scheme provides little additional insight about the possible benefits of user-cooperation in wireless systems, it is therefore not included in this thesis.

## E. Discussion

In this chapter, we proposed a simple user-cooperation scheme that works under the assumption that channel state information is only available at the receivers, not at the transmitters. A Markov model was introduced to capture the unreliable nature of

the wireless environment. For a fixed coderate, the overall performance of the wireless channel is modeled as a two-state Gilbert-Elliott model.

The achievable rate-region of the proposed user-cooperation scheme is characterized and it is compared to the region of a non-cooperative system. Numerical results suggest that cooperation yields a large gain over traditional systems. Furthermore, the gain increases as the service constraint imposed on the system becomes more stringent. User-cooperation can therefore provide wireless users with the flexibility to better share system resources. Our queueing analysis also hints at the fact that overall performance depends heavily on the time correlation of the underlying physical channel. In that sense, effective capacity is much more sensitive to higher-order statistics than, say, ergodic capacity or outage capacity. It is therefore imperative to use channel models that are amenable to analysis while providing an accurate representation of reality.

# CHAPTER V

## MULTI-ANTENNA GAUSSIAN SYSTEMS

Using multiple antennas has been shown to improve the performance of wireless communication systems in various circumstances [30]. For instance, a multi-antenna configuration can be used to increase the diversity [29] or the spatial multiplexing [27] of a point-to-point wireless connection. Several research initiatives are focusing on designing practical multiple-input multiple-output systems that improve the throughput of wireless links. Yet, very little work has been done to quantify the impact of a multi-antenna configuration on delay-sensitive communication over wireless systems. This is an important research topic that demands attention. In this chapter, we seek to identify the potential benefits of a MIMO configuration on the effective capacity of a wireless system [67].

To insure that packets received at the destination can be decoded, we assume that channel state information is available both at the transmitter and at the receiver. It is clear that, when CSI is available at the transmitter, the optimal power control policy takes as input the channel state and the buffer occupancy level. However, this design strategy leads to a very difficult optimization problem because of the time-dynamics of the system [10]. Herein, we assume that power is fixed at the transmitter. This allows us to focus on the benefits of using multiple antennas at the transmitter and/or receiver. Expressions for the effective capacities of a single-antenna system, vector Gaussian channels, and MIMO Gaussian systems are found under a Rayleigh block fading model. The effective capacity of the single-antenna system is compared to those of the vector Gaussian systems in the low signal-to-noise ratio (SNR) regime. Our results suggest that there is a substantial gain in using multiple antennas at the transmitter or receiver for delay-sensitive communication.

At low SNR, just as there is a power gain associated with using multiple receive antennas in terms of ergodic capacity [30], there is a statistical gain associated with using multiple transmit antennas in terms of effective capacity. For the MIMO case, asymptotic upper and lower bounds for the effective capacity are derived. The lower bound indicates that the effective capacity of a MIMO system scales linearly with the minimum number of transmit or receive antennas. An approximation for the effective capacity of the MIMO system is obtained in the low SNR regime when the number of transmit and/or receive antennas is large. Again, the effective capacity expression indicates that in the low SNR regime, a multi-antenna system offers a statistical gain as well as a power gain over a single-antenna system. This suggests that multi-antenna systems are especially suitable for service constrained communication.

The remainder of this chapter is organized as follows. Explicit formulas for the effective capacities of single-antenna and multi-antenna systems are derived in Section A. To gain better design intuition, system performance is characterized in the low SNR regime. In Section B, we analyze the asymptotic behavior of the effective capacity for MIMO systems as the number of transmit and/or receive antennas grows large. An upper bound and a lower bound for the effective capacity are obtained. Both suggest that the effective capacity of a MIMO system scales linearly with the minimum number of transmit or receive antennas. For systems with a large but finite number of antennas, we obtain approximations for the effective capacities of the corresponding multi-antenna configurations. Performance is again analyzed in the low SNR regime. Section C contains conclusions and final remarks.

## A. Effective Capacity of Vector Gaussian Channels

In this section, we study the effective capacities of single-antenna and multi-antenna systems under Rayleigh block fading. Note that the general expression and the properties of the effective capacity of block fading channels are explained in Chapter II. The single-user multi-antenna wireless system of interest is illustrated in Fig. 20. Suppose that the wireless user has a mean power constraint $P$ and a total spectral

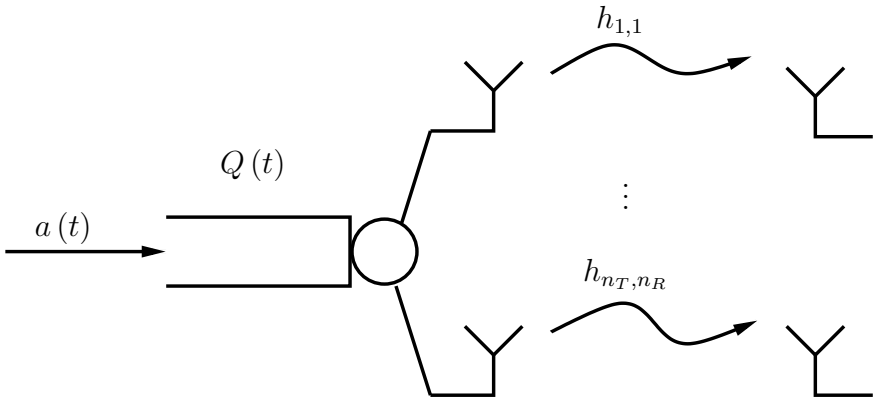


Fig. 20. A wireless queueing system model.

bandwidth allocation $W$. A large buffer is available to the wireless user, where the outgoing packets are stored before being transmitted to their destination. Throughout, we assume that CSI is known both at the transmitter and at the receiver. We also assume that the transmitter sends uncorrelated circularly symmetric zero-mean complex Gaussian signals of equal power across all the transmit antennas [27].

In this model, the multi-antenna system has $n_T$ transmit antennas and $n_R$ receive antennas, and the channel is Rayleigh block fading. Let $\mathbf{x}$ denote the $n_T \times 1$ vector of transmitted symbols, and $\mathbf{y}$ be the $n_R \times 1$ vector of received signals related by

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{n},$$

where $\mathbf{H}$ is an $n_R \times n_T$ complex matrix, and $\mathbf{n} \sim \mathcal{CN}(0, N_0\mathbf{I}_{n_R})$ is an $n_R \times 1$ vector of additive white Gaussian noise. The matrix $\mathbf{H}$ is called the channel matrix for this multi-antenna system. The element $h_{i,j}$ of this matrix denotes the channel gain from transmit antenna $j$ to receive antenna $i$. We assume that the $\{h_{i,j}\}$'s are i.i.d. zero-mean complex Gaussian random variables with unit variance.

### 1. Single-Antenna System

In the situation where the service process is governed by a single-input single-output (SISO) channel, the channel matrix $\mathbf{H}$ reduces to a scalar $h$ for each block. Under the assumption that CSI is known at the transmitter, the channel capacity $r$ during each block can be expressed as [1]

$$r = WT \log\left(1 + \frac{|h|^2 P}{N_0 W}\right) \qquad \text{nats per second,} \tag{5.1}$$

where $N_0/2$ denotes the power spectral density of the noise process. This maximum achievable rate during each block can be realized through Gaussian signaling with mean power $P$. We note that the assumption about CSI at the transmitter can be relaxed if the receiver has the ability to acknowledge reception of the data to ensure that the erroneous data is retransmitted. In this case, the maximum achievable rate $r$ in (5.1) can still be achieved using hybrid ARQ [68] or rateless codes [69, 70, 71]. The variable $r$ can be used to represent the instantaneous service rate of the corresponding wireless system. The moment generating function of the service process $E\left[e^{-\theta r}\right]$ can be obtained as follows,

$$\begin{aligned} E\left[e^{-\theta r}\right] &= \int_0^\infty e^{-\theta WT \log\left(1+\frac{|h|^2 P}{N_0 W}\right)} 2|h| e^{-|h|^2} d|h| = \int_0^\infty \left(1 + \frac{|h|^2 P}{N_0 W}\right)^{-\theta WT} e^{-|h|^2} d|h|^2 \\ &= e^{\frac{N_0 W}{P}} \left(\frac{P}{N_0 W}\right)^{-\theta WT} \Gamma\left(1 - \theta WT, \frac{N_0 W}{P}\right), \end{aligned}$$

where $\Gamma(z,x)$ is the *upper incomplete gamma function* given by

$$\Gamma(z,x) = \int_x^\infty t^{z-1}e^{-t}dt.$$

According to (2.3), the Gärtner-Ellis limit of the service process and the effective capacity for the SISO system can be obtained as

$$\begin{aligned}\Lambda(-\theta) &= \frac{1}{T}\left(\log\left(\Gamma\left(1-\theta WT, \frac{N_0W}{P}\right)\right) + \frac{N_0W}{P}\right) - \theta W \log\left(\frac{P}{N_0W}\right) \\ \alpha(\theta) &= W\log\left(\frac{P}{N_0W}\right) - \frac{1}{\theta T}\left(\log\left(\Gamma\left(1-\theta WT, \frac{N_0W}{P}\right)\right) + \frac{N_0W}{P}\right).\end{aligned} \tag{5.2}$$

Even though the expression for the effective capacity in (5.2) can be evaluated numerically, not much intuition can be drawn from looking at these equations alone. In the low SNR regime, the nature of the effective capacity for the SISO system can be seen more clearly. At low SNRs, the approximation

$$\log\left(1+\frac{|h|^2P}{N_0W}\right) \approx \frac{|h|^2P}{N_0W} \tag{5.3}$$

holds and $E\left[e^{-\theta r}\right]$ can be approximated by

$$E\left[e^{-\theta r}\right] \approx \int_0^\infty e^{-\frac{\theta T|h|^2P}{N_0}} e^{-|h|^2} d|h|^2 = \left(1+\frac{\theta TP}{N_0}\right)^{-1}.$$

Accordingly, in the power-limited regime, the effective capacity of the corresponding SISO system can be expressed as

$$\alpha(\theta) = \frac{1}{\theta T}\log\left(1+\frac{\theta TP}{N_0}\right). \tag{5.4}$$

For example, when $P = 10$ mW, $W = 1$ MHz, $T = 5$ ms, and $N_0 = 10^{-6}$ W/Hz, the effective capacity and the low SNR approximation of the SISO system appear in Fig. 21. As shown in Chapter II, the effective capacity is a monotonically decreasing function of $\theta$. When the mean received power $P$ is small, the system is operating in the

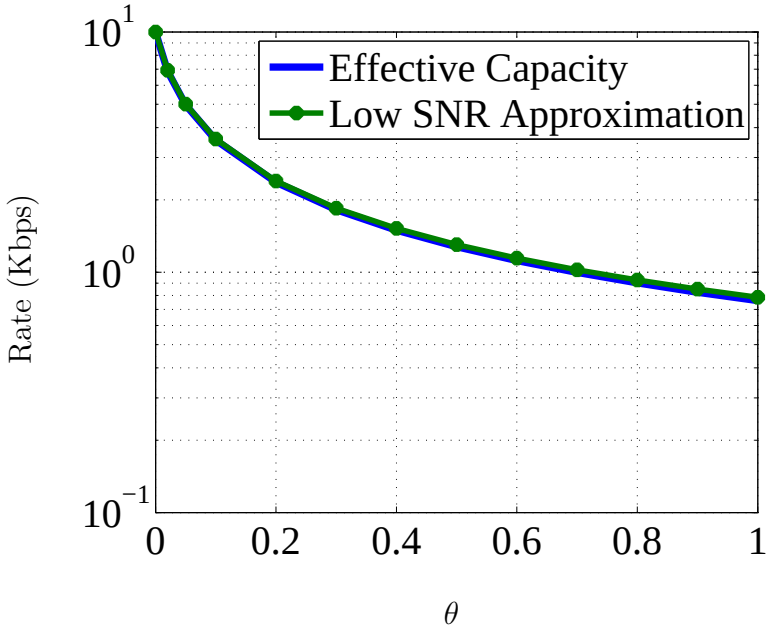


Fig. 21. Effective capacity and the low SNR approximation for SISO systems.

power-limited regime and the first-order approximation is very accurate. To further our understanding of the SISO system, we analyze (5.4) in both the small service exponent regime (large delay asymptotic) and the large service exponent regime (small delay asymptotic). It is easy to show that the effective capacity evaluated at $\theta \downarrow 0$ converges to the ergodic capacity of the wireless system in the low SNR regime,

$$\begin{aligned}\lim_{\theta\downarrow 0}\alpha(\theta) &= \lim_{\theta\downarrow 0}\frac{1}{\theta T}\log\left(1+\frac{\theta TP}{N_0}\right) = \lim_{\theta\downarrow 0}\frac{TP}{T(N_0+\theta TP)} \\ &= \frac{P}{N_0} = E\left[\frac{|h|^2P}{N_0}\right] \approx E\left[W\log\left(1+\frac{|h|^2P}{N_0W}\right)\right].\end{aligned}$$

This result coincides with the information theoretic statement that a wireless system can support a positive rate when the service has no delay requirement [72]. In the large service exponent regime, the effective capacity of the SISO system is given by

$$\lim_{\theta\uparrow\infty}\alpha(\theta) = \lim_{\theta\uparrow\infty}\frac{1}{\theta T}\log\left(1+\frac{\theta TP}{N_0}\right) = \lim_{\theta\uparrow\infty}\frac{TP}{T(N_0+\theta TP)} = 0.$$

Thus, when the communication system cannot tolerate any delay, the effective capacity of the SISO system converges to zero. That is, for any arrival rate $\epsilon > 0$, there exits a non-negligible probability that $\epsilon$ exceeds the instantaneous system throughput $r$ in (5.1). The zero-outage capacity of this SISO channel is zero. These findings are not surprising in the light of Lemma 3. The effective capacity is upper bounded by the ergodic capacity of the wireless system and lower bounded by the minimum instantaneous system throughput which is zero for Rayleigh single-antenna system.

Define the first-order derivative of the effective capacity as the decay function of the wireless system,

$$\alpha'(\theta) = \frac{\theta T^2 P - (N_0 T + \theta T^2 P) \log\left(1 + \frac{\theta T P}{N_0}\right)}{\theta^2 T^2 (N_0 + \theta T P)}.$$

The magnitude of this function indicates the decay speed of the effective capacity as the service exponent $\theta$ increases. As illustrated in Fig. 21, the decay speed of the effective capacity increases as $\theta$ approaches zero. Let

$$\rho \triangleq -\lim_{\theta \downarrow 0} \alpha'(\theta),$$

denote the initial decay-rate of the effective capacity for the corresponding wireless system. The initial decay-rate for the SISO system is equal to

$$\rho_{\text{SISO}} = \frac{TP^2}{2N_0^2}. \tag{5.5}$$

The usefulness of this quantity will become obvious when we start comparing various systems.

## 2. Multi-Antenna Systems

In this section, we consider the more general situation where the service process of a wireless system is governed by a multi-antenna channel. The effective capacity for the MIMO channel is obtained based on the moment generating function of its mutual information. The performance gains of the vector Gaussian systems over the single-antenna configuration are evaluated in the low SNR regime. Our results suggest that, at low SNR, a multiple receive antenna configuration provides a power gain of $n_R$ and a multiple transmit antenna configuration provides a statistical gain of $n_T$ over a single-antenna system.

For a fixed channel matrix $\mathbf{H}$, the mutual information between $\mathbf{x}$ and $\mathbf{y}$ during a block is given by [27]

$$\begin{aligned} I\left(\mathbf{H}, \mathbf{K}_x\right) &= WT \log\det\left(I_{n_T} + \frac{1}{N_0 W}\mathbf{K}_x\mathbf{H}^*\mathbf{H}\right) \\ &= WT \log\det\left(I_{n_R} + \frac{1}{N_0 W}\mathbf{H}\mathbf{K}_x\mathbf{H}^*\right), \end{aligned}$$

where $\mathbf{H}^*$ denotes the conjugate transpose of the channel matrix $\mathbf{H}$, and $\mathbf{K}_x$ is the correlation matrix of the transmitted symbols $\mathbf{K}_x = E\left[\mathbf{x}\mathbf{x}^*\right]$. Since the transmitter sends uncorrelated circularly symmetric zero mean complex Gaussian signals of equal power across all the transmit antennas, we have $\mathbf{K}_x = \left(P/n_T\right)\mathbf{I}_{n_T}$. Accordingly, the mutual information between $\mathbf{x}$ and $\mathbf{y}$, which is also the channel capacity of the corresponding MIMO system during each block, is equal to

$$\begin{aligned} I\left(\mathbf{H}, \left(P/n_T\right)\mathbf{I}_{n_T}\right) &= WT \log\det\left(\mathbf{I}_{n_T} + \frac{\mathbf{H}^*\mathbf{H}P}{N_0 W n_T}\right) \\ &= WT \log\det\left(\mathbf{I}_{n_R} + \frac{\mathbf{H}\mathbf{H}^*P}{N_0 W n_T}\right). \end{aligned}$$

As in the single-antenna case, we use $r = I\left(\mathbf{H}, \left(P/n_T\right)\mathbf{I}_{n_T}\right)$ to represent the realized service rate of the wireless system during each block.

Let

$$\mathbf{W} = \begin{cases} \mathbf{H}^*\mathbf{H} & n_T \leq n_R \\ \mathbf{H}\mathbf{H}^* & n_T > n_R, \end{cases}$$

$n = \max\{n_T, n_R\}$, and $m = \min\{n_T, n_R\}$. Then $\mathbf{W}$ is an $m \times m$ random non-negative definite matrix called the Wishart matrix [73]. Using a weak version of the matrix determinant lemma, $\det(\mathbf{I}_{n_T} + \mathbf{AC}) = \det(\mathbf{I}_{n_R} + \mathbf{CA})$ where $\mathbf{A}$ and $\mathbf{C}$ are $n_T \times n_R$ and $n_R \times n_T$ matrices respectively, it is easy to show that

$$r = WT \log\det\left(\mathbf{I}_m + \frac{\mathbf{W}P}{N_0 W n_T}\right) = WT \sum_{i=1}^{m} \log\left(1 + \frac{\lambda_i P}{N_0 W n_T}\right), \tag{5.6}$$

where $\lambda_i$ is the $i$th unordered eigenvalue of $\mathbf{W}$. The unordered eigenvalues of the Wishart matrix $\mathbf{W}$ have the joint probability density function [73]

$$p_\lambda(\lambda_1, \ldots, \lambda_m) = (m! K_{m,n})^{-1} e^{-\sum_i \lambda_i} \prod_i \lambda_i^{n-m} \prod_{i<j} (\lambda_i - \lambda_j)^2,$$

where $K_{m,n}$ is a normalizing factor. Based on the joint density function of the eigenvalues, the moment generating function $E\left[e^{-\theta r}\right]$ of the MIMO service process can be directly computed as shown in [74, Theorem 1]. Let $d = n - m$, then

$$E\left[e^{-\theta r}\right] = \mathbf{B}^{-1} \det\left[\mathbf{G}(\theta)\right]$$

where $\mathbf{B} = \prod_{i=1}^{k} \Gamma(d+i)$, and $\mathbf{G}(\theta)$ is an $m \times m$ Hankel matrix whose $(i,j)$th entry is defined by

$$g_{i,j} = \int_0^\infty \left(1 + \frac{P\lambda}{N_0 W n_T}\right)^{-\theta WT} \lambda^{i+j+d} e^{-\lambda} d\lambda, \quad i,j = 0, \ldots, m-1.$$

The effective capacity for the MIMO system can therefore be expressed as

$$\alpha(\theta) = -\frac{1}{\theta T} \log\left(\mathbf{B}^{-1} \det\left[\mathbf{G}(\theta)\right]\right). \tag{5.7}$$

The effective capacity expressed in (5.7) is somewhat intricate. Not much insight can

be gained from it alone. When considering the low SNR regime where the analog of (5.3) applies, the performance gains associated with the vector Gaussian systems can be seen more clearly.

Consider the situation where the wireless system has one transmit antenna and $n_R$ receive antennas. The channel matrix of the corresponding single-input multiple-output (SIMO) system $\mathbf{H} = [h_1, \ldots, h_{n_R}]^T$ becomes an $n_R \times 1$ vector of i.i.d. complex Gaussian random variables. During each block, the realized system throughput of the corresponding SIMO channel can be expressed as

$$r = WT \log\left(1 + \frac{\sum_{k=1}^{n_R} |h_k|^2 P}{N_0 W}\right) \qquad \text{nats per second.}$$

The moment generating function of the service process, $E\left[e^{-\theta r}\right]$, can be written as

$$E\left[e^{-\theta r}\right] = \int_{\mathbf{H}} e^{-\theta WT \log\left(1+\frac{\sum_{k=1}^{n_R} |h_k|^2 P}{N_0 W}\right)} f_{\mathbf{H}}(\mathbf{H})\, d\mathbf{H}.$$

Using the low SNR approximation of (5.3), and the fact that the components of the channel vector $\mathbf{H}$ are independent, $E\left[e^{-\theta r}\right]$ can be expressed as

$$E\left[e^{-\theta r}\right] = \left[\int_0^\infty e^{-\frac{\theta T |h|^2 P}{N_0}} e^{-|h|^2} d|h|^2\right]^{n_R} = \left(1 + \frac{\theta TP}{N_0}\right)^{-n_R}.$$

Accordingly, the effective capacity of the SIMO system in the low SNR regime is found to be

$$\alpha(\theta) = \frac{n_R}{\theta T} \log\left(1 + \frac{\theta TP}{N_0}\right). \tag{5.8}$$

Comparing (5.8) with (5.4), we find that the use of multiple receive antenna results in a power gain of $n_R$. As in the single-antenna case, the behavior of the effective capacity for a SIMO system can be seen more clearly in the asymptotic service exponent

regimes. Evaluating the effective capacity as $\theta \downarrow 0$ in the low SNR regime gives us

$$\lim_{\theta \downarrow 0} \alpha(\theta) = \lim_{\theta \downarrow 0} \frac{n_R TP}{T(N_0 + \theta TP)} = \frac{n_R P}{N_0} \approx \sum_{i=1}^{n_R} E\left[W \log\left(1 + \frac{|h_i|^2 P}{N_0 W}\right)\right],$$

which is the ergodic capacity of the SIMO channel in the low SNR regime [30]. As in the single-antenna case, in the large service exponent regime, the effective capacity of the SIMO channel goes to zero. This coincides with the fact that the minimum instantaneous service rate of a SIMO Rayleigh channel is zero. Furthermore, the decay function for the effective capacity of the SIMO system can be expressed as

$$\alpha'(\theta) = \frac{n_R \theta TP - (n_R N_0 + n_R \theta TP) \log\left(1 + \frac{\theta TP}{N_0}\right)}{\theta^2 T(N_0 + \theta TP)}. \tag{5.9}$$

Since the effective capacity of the SIMO system is $n_R$ times that of the SISO system, it is not surprising to find out that the initial decay-rate of the SIMO system is also $n_R$ times that of the SISO system,

$$\rho_{\text{SIMO}} = \frac{n_R TP^2}{2N_0^2}. \tag{5.10}$$

Again, the multiple receive antenna configuration provides a power gain of $n_R$.

For a SIMO system with the same parameters as in Fig. 21, the exact and approximated effective capacities of the SISO system and SIMO systems are shown in Fig. 22. Just as in the single-antenna case, the approximation of the effective capacity for the SIMO system in the low SNR regime is quite accurate. Furthermore, the effective capacity of the SIMO system scales linearly with the number of receive antennas for all service exponent $\theta_0$. This result can be viewed as an extension of the statement that the ergodic capacity of the SIMO system scales linearly with $n_R$ in the low SNR regime [30].

In the situation where the service process is governed by a multiple-input single-

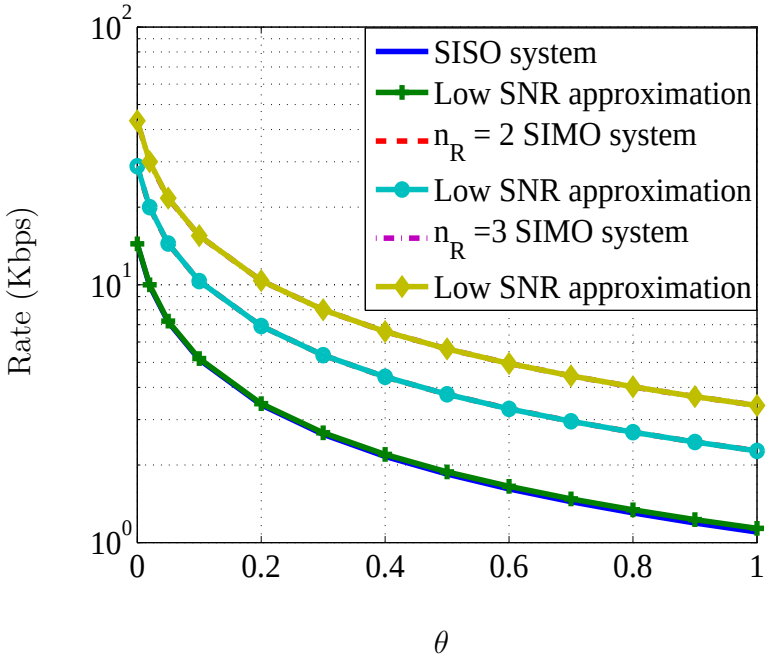


Fig. 22. Effective capacity for SIMO systems.

output (MISO) channel ($n_T$ transmit antennas and one receive antenna), the channel matrix $\mathbf{H} = [h_1, \ldots, h_{n_T}]$ becomes a $1 \times n_T$ vector of i.i.d. complex Gaussian random variables. The realized system throughput of the MISO channel during each block can be expressed as

$$r = WT \log\left(1 + \frac{\sum_{k=1}^{n_T} |h_k|^2 P}{N_0 W n_T}\right) \qquad \text{nats per second.}$$

The moment generating function of the service process can be expressed as

$$E\left[e^{-\theta r}\right] = \int_{\mathbf{H}} e^{-\theta WT \log\left(1+\frac{\sum_{k=1}^{n_T} |h_k|^2 P}{N_0 W n_T}\right)} f_{\mathbf{H}}(\mathbf{H})\, d\mathbf{H}.$$

Using the low SNR approximation of (5.3) and the independence assumption between

the components of $\mathbf{H}$, $E\left[e^{-\theta r}\right]$ can be further simplified to

$$E\left[e^{-\theta r}\right] = \left[\int_0^\infty e^{-\frac{\theta T|h|^2 P}{N_0 n_T}} e^{-|h|^2} d|h|^2\right]^{n_T} = \left(1 + \frac{\theta T P}{N_0 n_T}\right)^{-n_T}.$$

The effective capacity of the service process for the MISO system can then be expressed as

$$\alpha(\theta) = \frac{n_T}{\theta T} \log\left(1 + \frac{\theta T P}{N_0 n_T}\right). \tag{5.11}$$

The gain of using multiple transmit antennas over the single-antenna case is not entirely obvious by comparing (5.11) and (5.4). As before, we compare the performance of the effective capacity in the two asymptotic delay regimes. We know that in the low SNR regime, a multiple transmit antenna configuration is not beneficial without dynamic power allocation among the transmit antennas [30]. This statement follows from the fact that, without power allocation, having multiple transmit antennas does not increase the ergodic capacity at low SNRs. This can also be shown by evaluating the effective capacity as $\theta \downarrow 0$,

$$\lim_{\theta \downarrow 0} \alpha(\theta) = \lim_{\theta \downarrow 0} \frac{n_T T P}{T(N_0 n_T + \theta T P)} = \frac{P}{N_0} \approx E\left[W \log\left(1 + \frac{\sum_{i=1}^{n_T} |h_i|^2 P}{N_0 W n_T}\right)\right].$$

However, by analyzing the effective capacity of the MISO system, it can be shown that there is a statistical gain of $n_T$ associated with a multiple transmit antenna configuration. This statistical gain can be seen more clearly from the decay function of the effective capacity. The decay function of the MISO system is given by

$$\alpha'(\theta) = \frac{n_T \theta T P - (N_0 n_T^2 + n_T \theta T P) \log\left(1 + \frac{\theta T P}{N_0 n_T}\right)}{\theta^2 T (N_0 n_T + \theta T P)}, \tag{5.12}$$

and the initial decay-rate of the corresponding effective capacity becomes

$$\rho_{\text{MISO}} = \frac{T P^2}{2 n_T N_0^2}. \tag{5.13}$$

Comparing (5.13) and (5.5), we find that the decay-rate of the effective capacity for the MISO system is $1/n_T$ that of the single-antenna system. This implies that having multiple transmit antennas reduces the decay of the effective capacity as a function of $\theta$. This is especially beneficial to delay sensitive traffic. Since both the SIMO system and the SISO system have the same ergodic capacity in the low SNR regime, we define the statistical gain of the MISO system over the single-antenna system by

$$g = \frac{\rho_{\text{SISO}}}{\rho_{\text{MISO}}} = n_T. \tag{5.14}$$

The multiple transmit antenna configuration results in a statistical gain of $n_T$, but no gain in terms of ergodic capacity. It is also interesting to note that as $n_T \uparrow \infty$, $\rho_{\text{MISO}} \downarrow 0$ which means that as the number of transmit antenna becomes large the effective capacity does not decay at all when the service exponent $\theta$ is small. The results are easier to understand through a simple example. For the same parameters as in the SIMO case, the exact and approximated effective capacities for the SISO and MISO systems are shown in Fig. 23. The ergodic capacities of the $2\times 1$ system and the $3 \times 1$ system are the same as that of the single-antenna system. This result coincides with the information theoretic prediction that multiple transmit antennas cannot improve the ergodic capacity of a wireless system in the low SNR regime. However, when considering the effective capacities of the corresponding systems, using multiple transmit antennas induces a statistical gain that prevents the effective capacity from decaying rapidly as a function of service constraint $\theta$. Furthermore, as the service requirement $\theta$ increases, the gains in terms of the effective capacity for the MISO systems over the SISO system become larger. In fact, in the large antenna array regime, we will show that as the number of transmit antennas grows unbounded, the effective capacity of the corresponding MISO system becomes a constant over all values of $\theta$.

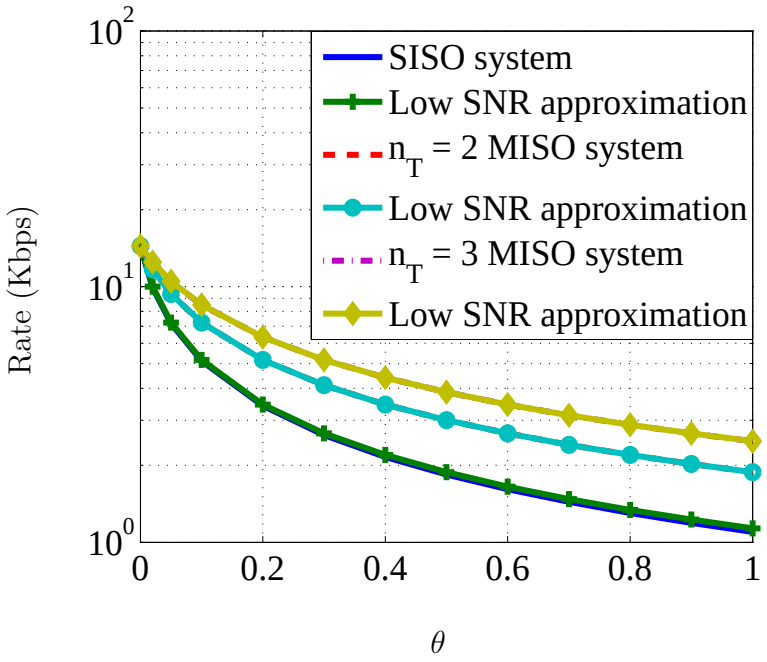


Fig. 23. Effective capacity for MISO systems.

For the general case of MIMO systems, the effective capacity in the low SNR regime can be illustrated in Fig. 24. That is, the multiple transmit antennas will provide a statistical gain of $n_T$ which prevent the effective capacity from decaying too fast and the multiple receive antennas will bring a power gain of $n_R$ that scales the whole curve up.

The results obtained in this section suggest that a multi-antenna configuration is especially beneficial for delay-sensitive applications over sensor networks using a two-tier architecture [75]. The timely processing and dissemination of information over wireless sensor networks is a key aspect of their future success. This is important for delay-sensitive applications such as detection and estimation. A two-tier architecture for wireless sensor networks contains sensor nodes which are simple devices equipped with one antenna, and cluster heads which are more powerful units endowed with

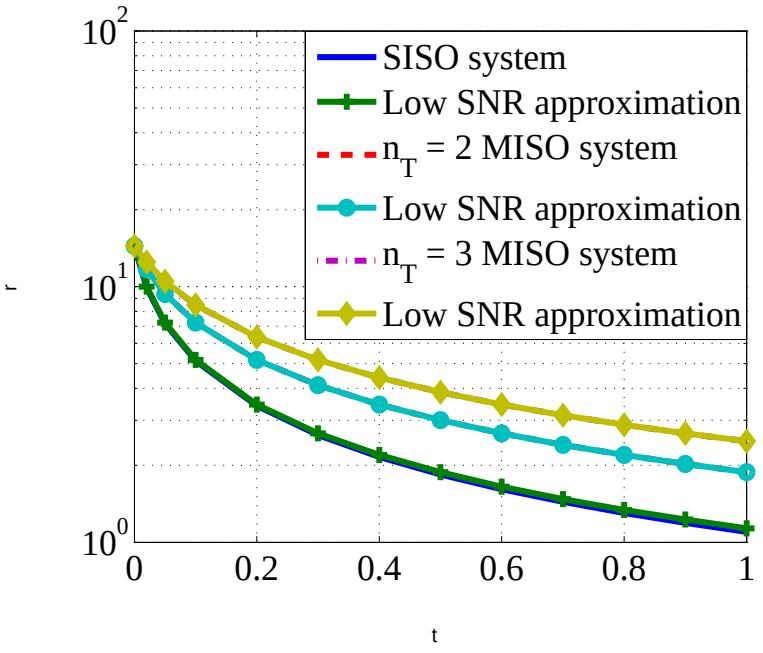


Fig. 24. Effective capacity for MIMO systems.

multiple antennas each. The cluster heads collect information from the sensor nodes, process data locally, and then relay pertinent information across the network. The fact that transmit antenna can bring a statistical gain of $n_T$ and the receive antenna can bring a power gain of $n_R$ provides support for the two-tier architecture and for cluster heads having multiple antennas. This is a direct application of the effective capacity analysis of multi-antenna systems.

## B. Asymptotic Analysis of MIMO Systems

In this section, we analyze the behavior of the effective capacity for MIMO systems in the large antenna-array regime. Various asymptotic analysis of multi-antenna systems are possible: (1) large $n_R$ and fixed $n_T$, (2) large $n_T$ and fixed $n_R$, and (3) increasingly large $n_T$ and $n_R$, while keeping $\beta = n_R/n_T$ constant. As described in Section A, the

realized system throughput of a MIMO channel during each block can be expressed as

$$r = WT \log\det\left(\mathbf{I}_m + \frac{\mathbf{W}P}{N_0 W n_T}\right),$$

where $\mathbf{W}$ is the Wishart matrix defined in Section 2. Moreover, the moment generating function of the service process $E\left[e^{-\theta r}\right]$ can be expressed as

$$E\left[e^{-\theta r}\right] = E\left[e^{-\theta WT \log\det\left(\mathbf{I}_m + \frac{\mathbf{W}P}{N_0 W n_T}\right)}\right] = E\left[e^{-\theta WT \sum_{i=1}^{m} \log\left(1+\frac{\lambda_i P}{N_0 W n_T}\right)}\right], \qquad (5.15)$$

where the joint density of the ordered eigenvalues is known to be

$$p_{\lambda,\text{ordered}}\left(\lambda_1,\ldots,\lambda_m\right) = K_{m,n}^{-1} e^{-\sum_i \lambda_i} \prod_i \lambda_i^{n-m} \prod_{i<j} \left(\lambda_i - \lambda_j\right)^2, \qquad \lambda_1 \geq \ldots \geq \lambda_m \geq 0.$$

The moment generating function of $E\left[e^{-\theta r}\right]$ can be computed numerically from the joint density function of the eigenvalues. To gain better insight, we proceed to derive upper and lower bounds that can be used to characterize the behavior of the effective capacity in the large antenna array regime. Since the eigenvalues of the Wishart matrix can be ordered as $\lambda_1 \geq \ldots \geq \lambda_m \geq 0$, it is clear that $E\left[e^{-\theta r}\right]$ is upper-bounded by

$$E\left[e^{-\theta r}\right] \leq E\left[e^{-m\theta WT \log\left(1+\frac{\lambda_m P}{N_0 W n_T}\right)}\right],$$

where $\lambda_m$ is the minimal eigenvalue of the Wishart matrix $\mathbf{W}$. The effective capacity $\alpha(\theta)$ of the MIMO channel can therefore be lower-bounded by

$$-\frac{1}{\theta T} \log\left(E\left[\left(1+\frac{\lambda_m P}{N_0 W n_T}\right)^{-m\theta WT}\right]\right) \leq \alpha(\theta). \qquad (5.16)$$

According to Lemma 3, the effective capacity $\alpha(\theta)$ of a MIMO system is upper-bounded by its ergodic capacity. Using Jensen's inequality, we obtain

$$\begin{aligned}\alpha(\theta) &= -\frac{1}{\theta T}\log E\left[e^{-\theta r}\right] = -\frac{1}{\theta T}\log E\left[e^{-\theta WT\log\det\left(\mathbf{I}_m+\frac{\mathbf{W}P}{N_0Wn_T}\right)}\right] \\ &\leq WE\left[\log\det\left(\mathbf{I}_m+\frac{\mathbf{W}P}{N_0Wn_T}\right)\right] = WE\left[\sum_{i=1}^{m}\log\left(1+\frac{\lambda_i P}{N_0Wn_T}\right)\right] \\ &= \sum_{i=1}^{m} WE\left[\log\left(1+\frac{\lambda_i P}{N_0Wn_T}\right)\right] = mWE\left[\log\left(1+\frac{\lambda P}{N_0Wn_T}\right)\right],\end{aligned}$$

where $\lambda$ is an unordered eigenvalue of the matrix $\mathbf{W}$ with probability density function

$$p_\lambda(\lambda) = \int\cdots\int p_\lambda(\lambda_1,\ldots,\lambda_m)\,d\lambda_2\cdots d\lambda_m.$$

Taking (5.16) into consideration, $\alpha(\theta)$ is bounded by

$$-\frac{1}{\theta T}\log\left(E\left[\left(1+\frac{\lambda_m P}{N_0Wn_T}\right)^{-m\theta WT}\right]\right) \leq \alpha(\theta) \leq mWE\left[\log\left(1+\frac{\lambda P}{N_0Wn_T}\right)\right].$$

### 1. Large $n_R$ and Fixed $n_T$

In the case where $n_R$ is asymptotically large and $n_T$ is fixed, we have $n = n_R$ and $m = n_T$. By a random matrix result due to Marčenko and Pastur [76], the empirical distribution of the eigenvalues of the normalized $m \times m$ Wishart matrix $\mathbf{W}/n$ converges in distribution to a point mass at $\lambda = 1$ as $n \to \infty$. Note that

$$\frac{\mathbf{W}}{n_T} = \frac{\mathbf{W}}{n_R}\times\frac{n_R}{n_T} = \frac{\mathbf{W}}{n}\times\frac{n}{m}.$$

The effective capacity of the MIMO system can then be obtained from (5.15),

$$\alpha(\theta) \doteq n_T W\log\left(1+\frac{n_R P}{N_0Wn_T}\right) \qquad \text{as } n_R\to\infty \tag{5.17}$$

where the above notation means that the ratio between the two sides of the equation tends to one as $n_R$ increases. For a fixed number of transmit antennas, when the

number of receive antenna goes to infinity, the effective capacity does not decay as $\theta$ increases. Furthermore, the effective capacity $\alpha(\theta)$ converges to the ergodic capacity of this MIMO system for all $\theta$. This phenomenon is termed *channel-hardening*. It plays an important role for delay-sensitive communication. Consider the situation where $n = n_R$ is large but finite. It can be shown that the distribution of the realized throughput during each block is well approximated by [77, Theorem 1]

$$r \sim \mathcal{N}\left(n_T W T \log\left(1 + \frac{n_R P}{N_0 W n_T}\right), \frac{n_T W^2 T^2}{n_R}\right).$$

The Gaussianity of the realized throughput for the MIMO system can be justified using Lyapunov's central limit theorem. The realized throughput in (5.6) is expressed as a sum of random variables that are correlated. When the number of transmit antennas and/or receive antennas becomes large, the throughput converges to a normal distribution. Accordingly, the effective capacity can be approximated by

$$\alpha(\theta) \approx n_T W \log\left(1 + \frac{n_R P}{N_0 W n_T}\right) - \frac{n_T \theta W^2 T}{2 n_R}. \tag{5.18}$$

In particular, when the number of receive antennas $n_R$ is finite, the effective capacity decays with the service constraint $\theta$ and is strictly less than the ergodic capacity of the same MIMO system. We note that for positive random variables, convergence in distribution implies convergence in effective capacity. This fact provides reasonable ground for (5.17) and for the approximations of (5.18), (5.19), and (5.23). This convergence property is discussed in greater detail in Appendix C.

### 2. Large $n_T$ and Fixed $n_R$

In this second scenario, $n_T$ grows asymptotically large and $n_R$ is fixed. We then have $n = n_T$ and $m = n_R$. To obtain convergence results, we use arguments similar to our previous ones. The $m$ eigenvalues of matrix $\mathbf{W}/n_T$ converge to a point mass at

$\lambda = 1$. The effective capacity of the MIMO system can then be obtained as

$$\lim_{n_T \to \infty} \alpha(\theta) = n_R W \log\left(1 + \frac{P}{N_0 W}\right).$$

As $n_T$ increases, the effective capacity converges to the ergodic capacity of the MIMO channel for all values of $\theta$. When $n_T$ is large but finite, the distribution of the realized throughput during each block is well approximated by [77, Theorem 2]

$$r \sim \mathcal{N}\left(n_R W T \log\left(1 + \frac{P}{N_0 W}\right), \frac{n_R W^2 T^2 P^2}{n_T \left(N_0 W + P\right)^2}\right).$$

The effective capacity can then be approximated by

$$\alpha(\theta) \approx n_R W \log\left(1 + \frac{P}{N_0 W}\right) - \frac{n_R \theta W^2 T P^2}{2 n_T \left(N_0 W + P\right)^2}. \tag{5.19}$$

We can evaluate the effective capacity for this MIMO systems in the low SNR regime. Its decay-rate can be computed from the approximated effective capacity of (5.19) as

$$\rho = \frac{n_R T P^2}{2 n_T N_0^2}.$$

As pointed out in [30], the ergodic capacity of the MIMO system in this setting does not change with the number of transmit antennas. In other words, the effective capacities of the corresponding MIMO systems converge to a unique value as $\theta \downarrow 0$, independently of the number of transmit antennas. However, a statistical gain of $n_T$ is achieved by the multiple transmit antenna systems over the single-antenna system. As the number of transmit antennas becomes large, the statistical gain of the multiple transmit antenna system increases and the effective capacity converges to the ergodic capacity for all $\theta > 0$.

### 3. Asymptotically Large $n_T$ and $n_R$

We now address the third case where the number of transmit antennas $n_T$ and the number of receive antennas $n_R$ increase with their ratio $\beta = n_R/n_T$ kept constant. In this situation, the eigenvalues of the normalized Wishart matrix $\mathbf{W}/n_T$ do not necessarily converge to deterministic quantities. However, these eigenvalues converge weakly to a known distribution.

Suppose that $n_T < n_R$, then $n = n_R$, $m = n_T$, and $\beta = n_R/n_T > 1$. The unordered eigenvalues of the normalized Wishart matrix $\mathbf{W}/n_T$ converge in distribution to [78],

$$p_\lambda(\lambda) \to \begin{cases} \frac{1}{\pi}\sqrt{\frac{\beta}{\lambda} - \frac{1}{4}\left(1 + \frac{\beta-1}{\lambda}\right)^2} & \left(\sqrt{\beta}-1\right)^2 \le \lambda \le \left(\sqrt{\beta}+1\right)^2 \\ 0 & \text{otherwise.} \end{cases}$$

The ergodic capacity which is an upper-bound for the effective capacity of the MIMO system can be computed as

$$\begin{aligned} &mWE\left[\log\left(1 + \frac{\lambda P}{N_0 W n_T}\right)\right] \\ &\quad = \frac{mW}{\pi}\int_{\left(\sqrt{\beta}-1\right)^2}^{\left(\sqrt{\beta}+1\right)^2} \log\left(1 + \frac{\lambda P}{N_0 W}\right)\sqrt{\frac{\beta}{\lambda} - \frac{1}{4}\left(1 + \frac{\beta-1}{\lambda}\right)^2}\, d\lambda. \end{aligned}$$

Let

$$F\left(\beta, \frac{P}{N_0 W}\right) = \frac{1}{\pi}\int_{\left(\sqrt{\beta}-1\right)^2}^{\left(\sqrt{\beta}+1\right)^2} \log\left(1 + \frac{\lambda P}{N_0 W}\right)\sqrt{\frac{\beta}{\lambda} - \frac{1}{4}\left(1 + \frac{\beta-1}{\lambda}\right)^2}\, d\lambda.$$

Using this notation, we can write this asymptotic upper bound for the effective capacity $\alpha(\theta)$ as

$$mWF\left(\beta, \frac{P}{N_0 W}\right).$$

On the other hand, the minimal eigenvalue of the normalized Wishart matrix $\mathbf{W}/n_T$

converges almost surely to $\left(\sqrt{\beta}-1\right)^2$ [79]. That is,

$$\lim_{m\to\infty}\lambda_m\left(\frac{\mathbf{W}}{n_T}\right)=\left(\sqrt{\beta}-1\right)^2 \qquad \text{almost surely.}$$

Hence an asymptotic lower-bound for the effective capacity of the MIMO system is

$$-\frac{1}{\theta T}\log\left(E\left[\left(1+\frac{\lambda_m P}{N_0 W n_T}\right)^{-m\theta WT}\right]\right)=mW\log\left(1+\frac{\left(\sqrt{\beta}-1\right)^2 P}{N_0 W}\right).$$

Combining these two results, we get an asymptotic interval for the effective capacity $\alpha(\theta)$,

$$mW\log\left(1+\frac{\left(\sqrt{\beta}-1\right)^2 P}{N_0 W}\right)\le\alpha(\theta)\le mWF\left(\beta,\frac{P}{N_0 W}\right). \tag{5.20}$$

It is interesting to note that unlike the SISO case, the effective capacity of the MIMO channel in the large antenna regime is bounded away from zero. It scales linearly with $m$, which is the minimum number of transmit and receive antennas.

When $n_T=n_R$, the unordered eigenvalues of $\mathbf{W}/n_T$ converge in distribution to

$$p_\lambda(\lambda)\to\begin{cases}\frac{1}{\pi}\sqrt{\frac{1}{\lambda}-\frac{1}{4}} & 0\le\lambda\le 4\\ 0 & \text{otherwise.}\end{cases}$$

An asymptotic upper-bound for the effective capacity of this MIMO system can be computed as

$$\begin{aligned}mWE\left[\log\left(1+\frac{\lambda P}{N_0 W n_T}\right)\right]&=\frac{mW}{\pi}\int_0^4\log\left(1+\frac{\lambda P}{N_0 W}\right)\sqrt{\frac{1}{\lambda}-\frac{1}{4}}d\lambda\\&=mWF\left(1,\frac{P}{N_0 W}\right).\end{aligned}$$

On the other hand, the realized throughput $r$ of the corresponding MIMO system during each block can be lower-bounded by [28]

$$\sum_{k=1}^{m}WT\log\left(1+\frac{P}{N_0 W m}\chi^2_{2k}\right)<r,$$

where $\{\chi^2_{2k} : k = 1, \ldots, m\}$ are independent chi-square random variables with the given degrees of freedom as subscripts. The lower-bound above should be interpreted in a probabilistic sense. Two random variables $A$ and $B$ satisfy the inequality $A < B$ if $\Pr\{B < \tau\} < \Pr\{A < \tau\}$ for any $\tau$. As $m$ becomes large, it is shown in [28] that the scaled lower-bound converges to a constant

$$\frac{1}{m}\sum_{k=1}^{m} WT \log\left(1 + \frac{P}{N_0 W m}\chi^2_{2k}\right) \to \int_0^1 \log\left(1 + \frac{xP}{N_0 W}\right) dx \qquad \text{as } m \to \infty.$$

Consequently,

$$\frac{1}{m}\sum_{k=1}^{m} WT \log\left(1 + \frac{P}{N_0 W m}\chi^2_{2k}\right) \to \left(1 + \frac{N_0 W}{P}\right)\log\left(1 + \frac{P}{N_0 W}\right) - 1 \qquad \text{as } m \to \infty.$$

Using Lemma 3, we conclude that the effective capacity of the MIMO system can be lower-bounded by

$$\alpha(\theta) \geq m\left(\left(1 + \frac{N_0 W}{P}\right)\log\left(1 + \frac{P}{N_0 W}\right) - 1 - \epsilon\right),$$

where $\epsilon > 0$. Since $\epsilon$ is arbitrary, $\alpha(\theta)$ must lie in the interval

$$m\left(\left(1 + \frac{N_0 W}{P}\right)\log\left(1 + \frac{P}{N_0 W}\right) - 1\right) \leq \alpha(\theta) \leq mWF\left(1, \frac{P}{N_0 W}\right) \qquad (5.21)$$

for all large enough systems. Hence, the effective capacity is bounded away from zero for all values of $\theta$ and scales linearly with $m$.

Finally, suppose that $n_T > n_R$, then $n = n_T$, $m = n_R$ and $\beta = n_R/n_T < 1$. The unordered eigenvalues of the normalized Wishart matrix $\mathbf{W}/n_R$ converge in distribution to [78]

$$p_\lambda(\lambda) \to \begin{cases} \frac{1}{\pi}\sqrt{\frac{1}{\lambda\beta} - \frac{1}{4}\left(1 + \frac{1-\beta}{\lambda\beta}\right)^2} & \frac{\left(1-\sqrt{\beta}\right)^2}{\beta} \leq \lambda \leq \frac{\left(1+\sqrt{\beta}\right)^2}{\beta} \\ 0 & \text{otherwise.} \end{cases}$$

Accordingly, the upper-bound for the effective capacity of the MIMO system can be

computed as

$$mWE\left[\log\left(1+\frac{\lambda\beta P}{N_0W}\right)\right] = \frac{mW}{\pi}\int_{\frac{(1-\sqrt{\beta})^2}{\beta}}^{\frac{(1+\sqrt{\beta})^2}{\beta}} \log\left(1+\frac{\lambda\beta P}{N_0W}\right)\sqrt{\frac{1}{\lambda\beta}-\frac{1}{4}\left(1+\frac{1-\beta}{\lambda\beta}\right)^2}\,d\lambda$$
$$= mWF\left(\frac{1}{\beta},\frac{\beta P}{N_0W}\right).$$

The minimal eigenvalue of the normalized Wishart matrix $\mathbf{W}/n_R$ can be shown to equal [79]

$$\lim_{m\to\infty}\lambda_m\left(\frac{\mathbf{W}}{n_R}\right) = \frac{\left(1-\sqrt{\beta}\right)^2}{\beta} \qquad \text{almost surely.}$$

Thus, a lower bound for the effective capacity of the MIMO system is

$$-\frac{1}{\theta T}\log\left(E\left[\left(1+\frac{\lambda_m\beta P}{N_0W}\right)^{-m\theta WT}\right]\right) = mW\log\left(1+\frac{\left(1-\sqrt{\beta}\right)^2 P}{N_0W}\right).$$

The effective capacity of this system can then be asymptotically bounded by

$$mW\log\left(1+\frac{\left(\sqrt{\beta}-1\right)^2 P}{N_0W}\right) \le \alpha(\theta) \le mWF\left(\frac{1}{\beta},\frac{\beta P}{N_0W}\right). \tag{5.22}$$

Both the upper and lower bounds produce a linear growth in the effective capacity at any SNR.

The performance gains of the MIMO system studied in the current section can be seen most clearly in the low SNR regime. When the number of transmit antennas $n_T$ and the number of receive antennas $n_R$ are large but finite, the realized throughput $r$ of each block can be approximated by a Gaussian random variable [77, Theorem 3]

$$r \sim \mathcal{N}\left(\frac{n_RTP}{N_0},\frac{n_RT^2P^2}{n_TN_0^2}\right).$$

The effective capacity of the MIMO system becomes

$$\alpha(\theta) \approx \frac{n_RP}{N_0} - \frac{n_R\theta TP^2}{2n_TN_0^2}. \tag{5.23}$$

From the approximated effective capacity of (5.23), it is clear that in the low SNR

regime an $n_T \times n_R$ MIMO system yields a power gain of $n_R$ over a $n_T \times 1$ MIMO system and a statistical gain of $n_T$ over an $1 \times n_R$ MIMO system. In the situation where $n_T$ and $n_R$ increase jointly with $\beta = n_R/n_T$ kept constant, and considering the low SNR regime, the effective capacity of the corresponding MIMO system can be approximated by

$$\alpha(\theta) \approx n_R \left( \frac{P}{N_0} - \frac{\theta \beta T P^2}{2 n_R N_0^2} \right).$$

As $n_T$ becomes large, an $n_T \times n_R$ system will achieve a statistical gain of $\beta = n_T/n_R$ over an $n_R \times n_R$ MIMO system.

## C. Discussion

In this chapter, we studied the interplay between the physical layer infrastructure and the queueing behavior of a wireless communication system. More specifically, we studied the impacts of a multi-antenna configuration on the perceived service quality at the link layer. As described in Chapter II, the quality of service is defined in terms of the LDP governing the probability of buffer overflow. It is closely related to the probability of delay violation. Based on this framework, we characterized the effective capacities of single-antenna systems as well as multi-antenna systems. There is a substantial gain in using a multi-antenna configuration over a single-antenna. In the low SNR regime, we have shown that having multiple transmit antennas can provide a statistical gain of $n_T$, while multiple receive antennas bring a power gain of $n_R$. The power gain of receive antennas is not surprising, however, the statistical gain of the transmit antennas offers some new insights on the multi-antenna systems. The statistical gain is a result of *channel-hardening* which cannot be visible from an ergodic capacity analysis of the multi-antenna systems alone. This suggests that a multi-antenna configuration is especially beneficial to delay-sensitive traffic.

# CHAPTER VI

## SUMMARY AND FUTURE WORKS

This dissertation seeks to take a cross-disciplinary approach to analyze delay-sensitive communication over wireless systems. The problem formulation in this dissertation addresses issues that lie at the boundary between the physical layer and higher network layers. It is often tempting to adopt a layered framework where each layer serves as a "black box" abstraction to higher layers. A system can then be implemented by designing each layer separately. However, in the context of wireless networks with delay-sensitive traffic, such a layered architecture does not offer a complete picture. For instance, using multiple transmit antennas in the low SNR regime yields a statistical gain that cannot be observed by looking solely at the throughput of a wireless connection. Accordingly, we proposed an integrated framework to analyze delay-sensitive applications over wireless systems. More specifically, the large deviation principle governing buffer occupancy serves as the foundation for our performance evaluation methodology. The analysis presented in this dissertation offers a new perspective to improve our understanding of delay-sensitive applications over wireless systems. The insights obtained from this work include the following guidelines.

- Delay constraints significantly influence how to allocate system resources: as service requirements become more stringent, a communication system should use a lower coderate to increase the probability of the channel being *ON*.

- Channel correlation is found to have a major impact on system performance as it impairs system performance. The higher the correlation coefficient is, the lower the effective capacity becomes.

- When channel state information is not available at the transmitter, a system may only be able to support a finite arrival rate, even with unlimited amount of physical resources.

- User cooperation is beneficial for delay-sensitive applications. The gain increases as the service constraint becomes more stringent.

- Multi-antenna systems can bring a statistical gain of $n_T$ and a power gain of $n_R$ over single antenna systems.

The results obtained in the framework of this dissertation lead to new research questions on the topic of delay-sensitive communication over wireless systems. Avenues of future research in the area include the following topics.

- **Joint Analysis of Queueing Delay and Coding Delay:** As mentioned in Chapter II, the overall transmission delay is the sum of the two distinct components. For a given finite-state channel and service requirement $\theta$, we can derive expressions for the optimum block length with random codes under various traffic and channel profiles. Intuitively, if the block length is large, the block error probability will be small while the queueing delay will be large since we will have to wait until the whole block is received to decode the sent information. Expressing the LDP of the queue as a function of code rate and block error probability will help us characterize the closed form solution to an optimal block length selection problem.

- **Delay-Aware Resource Allocation:** Much of the work on resource allocation for wireless systems focuses on the situation where the channel gain of every user is either constant or subject to slow fading. However, these assumptions may not hold in practice. Part of the benefits of the proposed methodology is

to have the ability to include a stochastic dynamic channel model as part of the problem definition. This strategy will allow us to take full advantage of our understanding of wireless channels rather than relying chiefly on their first-order statistics. Queueing models can thereafter be used for the optimal allocation of radio resources for correlated communication channels under various service constraints. This work can be seen as the extension of the resource allocation of the single-antenna system described in Chapter III. Extensions to multi-antenna systems in the framework described in Chapter V are also possible.

- **Delay and Distortion Trade-off:** It is clear that there is a trade-off between delay and distortion for end-to-end communication systems. However, in most of the current literature, delay is captured as the average delay of the communication system obtaining through Little's Law. Our work in [67] shows that the delay-violation probability of a communication system can be upper-bounded through the LDP governing buffer occupancy. Accordingly, we will be able to obtain a more meaningful delay outage and distortion tradeoff curve based on the new methodology proposed in Chapter II.

# REFERENCES

[1] T. Cover and J. Thomas, *Elements of Information Theory*. New York: Wiley-Interscience, 1991.

[2] R. Cruz, "A calculus for network delay, I: network elements in isolation," *IEEE Transactions on Information Theory*, vol. 37, no. 1, pp. 114 –131, January 1991.

[3] R. Cruz, "A calculus for network delay, II: network analysis," *IEEE Transactions on Information Theory*, vol. 37, no. 1, pp. 132 –141, January 1991.

[4] N. Malcolm and W. Zhao, "Hard real-time communication in multiple-access networks," *Real Time Systems*, vol. 8, no. 1, pp. 35 – 77, January 1995.

[5] S. Wang, R. Nathuji, R. Bettati, and W. Zhao, "Providing statistical delay guarantees in wireless networks," in *24th International Conference on Distributed Computing Systems*. IEEE, Tokyo, Japan, March 2004.

[6] W. Zhao and J. A. Stankovic, "Performance analysis of fcfs and improved fcfs scheduling algorithms for dynamic real-time computer systems," in *Real Time Systems Symposium*. IEEE, Santa Monica, California, USA, December 1989.

[7] S. Wang, D. Xuan, R. Bettati, and W. Zhao, "Providing absolute differentiated services for real-time applications in static-priority scheduling networks," *IEEE/ACM Transactions on Networking*, vol. 12, no. 2, pp. 326 – 339, April 2004.

[8] A. Ephremides and B. Hajek, "Information theory and communication networks: an unconsummated union," *IEEE Transactions on Information Theory*, vol. 44, no. 6, pp. 2416 – 2434, October 1998.

[9] M. Neely and E. Modiano, "Capacity and delay tradeoffs for ad-hoc mobile networks," *IEEE Transactions on Information Theory*, vol. 51, no. 6, pp. 1917 – 1937, June 2005.

[10] R. Berry and R. Gallagar, "Communication over fading channels with delay constraints," *IEEE Transactions on Information Theory*, vol. 48, no. 5, pp. 1135 – 1149, May 2002.

[11] X. Lin, N. B. Shroff, and R. Srikant, "A tutorial on cross-layer optimization in wireless networks," *IEEE Journal on Selected Areas in Communications*, vol. 24, no. 8, pp. 1452 – 1463, August 2006.

[12] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*, Stochastic Modelling and Applied Probability, 2nd ed. New York: Springer, 1998.

[13] A. Elwalid and D. Mitra, "Effective bandwidth of general markovian traffic sources and admission control of high speed networks," *IEEE/ACM Transactions on Networking*, vol. 1, no. 3, pp. 329–343, June 1993.

[14] C.-S. Chang, "Stability, queue length, and delay of deterministic and stochastic queueing networks," *IEEE Transactions on Automatic Control*, vol. 39, no. 5, pp. 913–931, May 1994.

[15] C.-S. Chang and J. Thomas, "Effective bandwidth in high-speed digital networks," *IEEE Journal on Selected Areas in Communications*, vol. 13, no. 6, pp. 1091–1100, August 1995.

[16] F. P. Kelly, "Effective bandwidths at multi-type queues," *Queueing Systems*, vol. 9, pp. 5–15, 1991.

[17] C.-S. Chang, *Performance Guarantees in Communication Networks*. New York: Springer-Verlag, 2000.

[18] F. P. Kelly, S. Zachary, and I. B. Ziedins, *Stochastic Networks: theory and applications*, Royal Statistical Society Lecture Notes. Cambridge: Oxford University Press, 1996.

[19] Dapeng Wu and Rhoit Negi, "Effective capacity: a wireless link model for support of quality of service," *IEEE Transactions on Wireless Communications*, vol. 2, no. 4, pp. 630–643, July 2003.

[20] D. Wu and R. Negi, "Downlink scheduling in a cellular network for quality-of-service assrance," *IEEE Transactions on Vehicular Technology*, vol. 53, no. 5, pp. 1547–1557, September 2004.

[21] D. Wu and R. Negi, "Utilizing multiuser diversity for efficient support of quality of service over a fading channel," *IEEE Transactions on Vehicular Technology*, vol. 54, no. 3, pp. 1198–1206, May 2005.

[22] L. Liu, P. Parag, J. Tang, W. Chen, and J.-F. Chamberland, "Resource allocation and quality of service evaluation for wireless communication systems using fluid models," in *Forty-Fourth Annual Allerton Conference*, Monticello, Illinois, USA, September 2006.

[23] L. Liu, P. Parag, J. Tang, W. Chen, and J.-F. Chamberland, "Resource allocation and quality of service evaluation for wireless communication systems using fluid models," *IEEE Transactions on Information Theory*, vol. 53, no. 5, pp. 1767 – 1777, May 2007.

[24] A. Sendonaris, E. Erkip, and B. Aazhang, "User cooperation diversity – part

I: system description," *IEEE Transactions on Communications*, vol. 51, no. 11, pp. 1927–1938, November 2003.

[25] L. Liu, J.-F. Chamberland, and S. Miller, "The uplink achievable rate region of a user cooperation scheme," in *Canadian Workshop on Information Theory*. IEEE, Montreal, Canada, June 2005, pp. 163–166.

[26] L. Liu, J.-F. Chamberland, and S. Miller, "User cooperation in the absence of phase information at the transmitters," *IEEE Transactions on Information Theory*, vol. 54, no. 3, pp. 1197 – 1206, March 2008.

[27] E. Telatar, "Capacity of multi-antenna gaussian channels," *European Transactions on Telecommunication*, vol. 10, no. 6, pp. 585–596, November 1999.

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

[29] V. Tarokh, H. Jafarkhani, and A. R. Calderbank, "Space-time block code from orthogonal designs," *IEEE Transactions on Information Theory*, vol. 45, no. 5, pp. 1456–1467, July 1999.

[30] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*. Cambridge: Cambridge University Press, 2005.

[31] R. Gallager, *Information Theory and Reliable Communication*. New York: Wiley, 1968.

[32] A. Viterbi, "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm," *IEEE Transactions on Information Theory*, vol. 13, no. 2, pp. 260 – 269, April 1967.

[33] A. Sahai, "Why block length and delay behave differently if feedback is present," To appear, the *IEEE Transactions on Information Theory*, 2008.

[34] D. Wu and R. Negi, "Effective capacity-based quality of service measures for wireless networks," *ACM Mobile Networks and Applications (MONET)*, vol. 11, no. 1, pp. 91 – 99, February 2006.

[35] T. Rappaport, *Wireless Communications: Principles and Practice*, 2nd ed. Upper Saddle River: Prentice Hall PTR, 2001.

[36] E. Biglieri, J. Proakis, and S. Shamai, "Fading channels: information-theoretic and communications aspects," *IEEE Transactions on Information Theory*, vol. 44, no. 6, pp. 2619 – 2692, October 1998.

[37] B. Oksendal, *Stochastic Differential Equations : An Introduction with Applications*. Universitext, 6th ed. New York: Springer, 2003.

[38] J. Norris, *Markov Chains*, Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge: Cambridge University Press, 1998.

[39] F. P. Kelly, *Reversibility and Stochastic Networks*. New York: Wiley, 1979.

[40] B. Hajek, *Analysis of computer networks*. Urbana-Champaign: University of Illinois Press, 2003.

[41] D. Mitra, "Stochastic theory of a fluid model of producers and consumers coupled by a buffer," *Advances in Applied Probability*, vol. 20, pp. 646–676, 1993.

[42] S. P. Meyn and R. L. Tweedie, *Markov Chains and Stochastic Stability*. New York: Springer-Verlag, 1996.

[43] C. Rago, P. Willett, and Y. Bar-Shalom, "Censoring sensors: a low-communication-rate scheme for distributed detection," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 32, no. 2, pp. 554–568, April 1996.

[44] G. Kesidis, J. Walrand, and C.-S. Chang, "Effective bandwidths for multiclass markov fluids and other atm sources," *IEEE/ACM Transactions on Networking*, vol. 1, no. 4, pp. 424–428, August 1993.

[45] M. M. Krunz and J. G. Kim, "Fluid analysis of delay and packet discard performance for QoS support in wireless networks," *IEEE Journal on Selected Areas in Communications*, vol. 19, no. 2, pp. 384 – 395, February 2001.

[46] H. Heffes and D. M. Lucantoni, "A Markov modulated characterization of packetized voice and data traffic and related statistical multiplexer performance," *IEEE Journal on Selected Areas in Communications*, vol. 4, no. 6, pp. 856 – 868, September 1986.

[47] R. D. Yates, "A framework for uplink power control in cellular radio systems," *IEEE Journal on Selected Areas in Communications*, vol. 13, no. 7, pp. 1341 – 1347, September 1995.

[48] J.-F. Chamberland and V. V. Veeravalli, "Decentralized dynamic power control for cellular CDMA systems," *IEEE Transactions on Wireless Communications*, vol. 2, no. 3, pp. 549–559, May 2003.

[49] N. Bambos and S. Kandukuri, "Power-controlled multiple access schemes for next-generation wireless packet networks," *IEEE Wireless Communications*, vol. 9, no. 3, pp. 58 – 64, June 2002.

[50] I. Karatzas and S. E. Shreve, *Brownian Motion and Stochastic Calculus*, 2nd ed. Graduate Texts in Mathematics. New York: Springer, 1997.

[51] A. Sendonaris, E. Erkip, and B. Aazhang, "User cooperation diversity – part II: implementation aspects and performance analysis," *IEEE Transactions on Communications*, vol. 51, no. 11, pp. 1939–1948, November 2003.

[52] L. Zheng and D. Tse, "Diversity and multiplexing: a fundamental tradeoff in multiple antenna channels," *IEEE Transactions on Information Theory*, vol. 49, no. 5, pp. 1073 – 1096, May 2003.

[53] J. Laneman and G. Wornell, "Distributed space-time-coded protocols for exploiting cooperative diversity in wireless networks," *IEEE Transactions on Information Theory*, vol. 49, no. 10, pp. 2415–2425, October 2003.

[54] J. Laneman, D. Tse, and G. Wornell, "Cooperative diversity in wireless networks: efficient protocols and outage behavior," *IEEE Transactions on Information Theory*, vol. 50, no. 12, pp. 3062–3080, December 2004.

[55] M. Janani, A. Hedayat, T. E. Hunter, and A. Nosratinia, "Coded cooperation in wireless communications: space-time transmission and iterative decoding," *IEEE Transactions on Signal Processing*, vol. 52, no. 2, pp. 362 – 371, February 2004.

[56] G. Kramer, M. Gastpar, and P. Gupta, "Cooperative strategies and capacity theorems for relay networks," *IEEE Transactions on Information Theory*, vol. 51, no. 9, pp. 3037–3063, September 2005.

[57] A. Høst Madsen, "Capacity bounds for cooperative diversity," *IEEE Transactions on Information Theory*, vol. 52, no. 4, pp. 1522–1544, April 2006.

[58] M. A. Khojastepour, A. Sabharwal, and B. Aazhang, "On capacity of gaussian 'cheap' relay channel," in *IEEE Global Communications Conference*, San Francisco, USA, 2003, vol. 3, pp. 1776 – 1780.

[59] K. Azarian, H. El Gamal, and P. Schniter, "On the achievable diversitymultiplexing tradeoff in half-duplex cooperative channels," *IEEE Transactions on Information Theory*, vol. 51, no. 12, pp. 4152 – 4172, December 2005.

[60] R. A. Horn and C. R. Johnson, *Topics in Matrix Analysis*. Cambridge: Cambridge University Press, 1994.

[61] R. J. Gibbens and P. J . Hunt, "Effective bandwidths for the multi-type uas channel," *Queueing Systems*, vol. 9, pp. 17–28, 1991.

[62] T. Stern and A. Elwalid, "Analysis of separable markov-modulated rate models for information-handling systems," *Advances in Applied Probability*, vol. 23, pp. 105–139, 1991.

[63] G. Debreu and I. Herstein, "Nonnegative square matrices," *Econometrica*, vol. 21, pp. 597–607, 1953.

[64] J. E. Cohen, "Random evolutions and the spectral radius of a nonnegative matrix," *Mathematical Proceedings of Cambridge Philosophical Society*, vol. 86, pp. 345–350, 1979.

[65] J. E. Cohen, "Convexity of the dominant eigenvalue of an essentially nonnegative matrix," *Proceedings of American Mathematical Society*, pp. 657–658, 1981.

[66] A. Goldsmith, *Wireless Communications*. Cambridge: Cambridge University Press, 2005.

[67] L. Liu and J.-F. Chamberland, "On the effective capacities of multiple-antenna gaussian channels," Submitted to the *IEEE Transactions on Information Theory*, September 2007.

[68] G. Caire and D. Tuninetti, "The throughput of hybrid-arq protocols for the gaussian collision channel," *IEEE Transactions on Information Theory*, vol. 47, no. 5, pp. 1971 – 1988, July 2001.

[69] M. Luby, "Lt codes," in *Forty-Third Annual IEEE Symposium on Foundations of Computer Science*, Vancouver, Canada, 2002, pp. 271 – 280.

[70] A. Shokrollahi, "Raptor codes," in *IEEE International Symposium on Information Theory*, Chicago, Illinois, USA, June 2004.

[71] R. Palanki and J. S. Yedidia, "Rateless codes on noisy channels," in *IEEE International Symposium on Information Theory*, Chicago, Illinois, USA, June 2004.

[72] E. Biglieri, J. Proakis, and S. Shamai, "Fading channels: information-theoretic and communications aspects," *IEEE Transactions on Information Theory*, vol. 44, no. 6, pp. 2619 – 2692, October 1998.

[73] A. Edelman, *Eigenvalues and Condition Numbers of Random Matrices*, Ph.D. dissertation, Department of Mathematics, Massachusetts Institute of Technology, Cambridge, MA, 1989.

[74] Z. Wang and G. B. Giannakis, "Outage mutual information of space-time mimo channels," *IEEE Transactions on Information Theory*, vol. 50, no. 4, pp. 657–662, April 2004.

[75] L. Liu, J.-F. Chamberland, and K. Qaraqe, "Enabling delay-sensitive applications over sensor networks using a two-tier architecture and multi-antenna cluster heads," in *IEEE Second International Workshop on Computational Advances in Multi-Sensor Adaptive Processing*, St. Thomas, U.S. Virgin Islands, USA, December 2007.

[76] V. A. Marčenko and L. A. Pastur, "Distribution of eigenvalues for some sets of random matrices," *Math USSR Sbornik*, , no. 1, pp. 457– 483, 1967.

[77] B. M. Hochwald, T. L. Marzetta, and V. Tarokh, "Multi-antenna channel-hardening and its implications for rate feedback and scheduling," *IEEE Transactions on Information Theory*, vol. 50, no. 9, pp. 1893– 1909, September 2004.

[78] J. W. Silverstein and Z. D. Bai, "On the empirical distribution of eigenvalues of a class of large dimensional random matrices," *Journal of Multivariate Analysis*, vol. 54, pp. 175 – 192, 1995.

[79] J. W. Silverstein, "Strong convergence of the empirical distribution of eigenvalues of large dimensional random matrices," *Journal of Multivariate Analysis*, vol. 55, pp. 331 – 339, 1995.

[80] W. Rudin, *Real and Complex Analysis*, 3rd ed. New York: McGraw-Hill Science Engineering, 1986.

# APPENDIX A

## DELAY VIOLATION PROBABILITY

Suppose that the joint process of the queue length and channel, $(Q(t), h(t))$, is stationary and ergodic. Then this process converges in distribution to a probability law $\mu$. Let $D(\infty)$ be a random variable whose distribution coincides with the delay experienced by packets at steady-state, and $D_{\max}$ be the delay constraint imposed on the traffic. Similarly, let $Q(\infty)$ be a random variable whose distribution is equal to the queue-length distribution of the buffer at steady-state, and $Q_{\max} = aD_{\max}$ be the delay-violation threshold for the queue. Note that this relationship holds because of the constant arrival rate in the buffer. The probability that the queue-length at steady-state exceeds $Q_{\max}$ is

$$\Pr\{Q(\infty) > Q_{\max}\} = \int_{\mathcal{R}^+ \times \mathcal{H}} \mathbf{1}_{\{q > Q_{\max}\}} d\mu(q, h),$$

where $\mathbf{1}_{\{\cdot\}}$ is the indicator function.

At time $t$, the delay $D(t)$ experienced by a packet that is about to leave the buffer is related to the queue length of the buffer $Q(t)$ through $Q(t) = aD(t)$. For a specific realization of the system, the empirical probability that a packet transmitted during time interval $[0, T]$ exceeds $D_{\max}$ is given by

$$\frac{\int_0^T \mathbf{1}_{\{Q(t) > Q_{\max}\}} v\,(t)\, dt}{\int_0^T v\,(t)\, dt},$$

where $v\,(t)$ is the instantaneous departure rate of the system at time $t$. Note that when the buffer is non-empty, $v(t)$ is equal to $r(h(t))$, the instantaneous service rate

of the wireless channel. Thus, the limiting delay-violation probability is equal to

$$\lim_{T\to\infty} \frac{\frac{1}{T}\int_0^T \mathbf{1}_{\{Q(t)>Q_{\max}\}} r\left(h(t)\right) dt}{\frac{1}{T}\int_0^T v\left(t\right) dt}$$
$$= \lim_{T\to\infty} \frac{\frac{1}{T}\int_0^T \mathbf{1}_{\{Q(t)>Q_{\max}\}} r\left(h(t)\right) dt}{\frac{1}{T}\left(aT + Q(0) - Q(T)\right)}.$$

The stability of the system implies that

$$\lim_{T\to\infty} \frac{1}{T}\left(aT + Q(0) - Q(T)\right) = a.$$

Since the joint process $(Q(t), h(t))$ is stationary and ergodic, we can compute the delay-violation probability using the limiting distribution $\mu$,

$$\Pr\left\{D(\infty) > D_{\max}\right\} = \frac{1}{a}\int_{\mathcal{R}^+\times\mathcal{H}} \mathbf{1}_{\{q>Q_{\max}\}} r\left(h\right) d\mu(q,h).$$

Accordingly, the delay-violation probability can be bounded as follows,

$$\begin{aligned}
\Pr\left\{D(\infty) > D_{\max}\right\} &= \frac{1}{a}\int_{\mathcal{R}^+\times\mathcal{H}} \mathbf{1}_{\{q>Q_{\max}\}} r\left(h\right) d\mu(q,h) \\
&\le \frac{1}{a}\sqrt{\int_{\mathcal{R}^+\times\mathcal{H}} \mathbf{1}^2_{\{q>Q_{\max}\}} d\mu(q,h)}\sqrt{\int_{\mathcal{R}^+\times\mathcal{H}} r^2(h) d\mu(q,h)} \\
&= \frac{1}{a}\sqrt{\int_{\mathcal{R}^+\times\mathcal{H}} \mathbf{1}_{\{q>Q_{\max}\}} d\mu(q,h)}\sqrt{\int_{\mathcal{R}^+\times\mathcal{H}} r^2(h) d\mu(q,h)} \\
&= \frac{1}{a}\sqrt{\Pr\left\{Q(\infty) > Q_{\max}\right\}}\sqrt{\int_{\mathcal{R}^+\times\mathcal{H}} r^2(h) d\mu(q,h)},
\end{aligned}$$

where the first inequality comes from the Cauchy-Schwartz inequality, and the second equality comes from the fact that $\mathbf{1}_A^2 = \mathbf{1}_A$. Let $\nu$ be the probability law for the marginal distribution of the channel in steady-state, we have

$$\int_{\mathcal{R}^+\times\mathcal{H}} r^2(h) d\mu(q,h) = \int_{\mathcal{H}} r^2(h) d\nu(h).$$

Therefore,

$$\Pr\{D((\infty) > D_{\max}\} \leq c\sqrt{\Pr\{Q(\infty) > Q_{\max}\}},$$

where $c = \frac{1}{a}\sqrt{\int_{\mathcal{H}} r^2(h) d\nu(h)}$ is a constant independent of the queue distribution.

# APPENDIX B

## EFFECTIVE CAPACITY OF SERVICE PROCESS

In this section, we use the Kolmogorov backward equation to derive a formula for the effective capacity of a Markov-modulated service process. We parallel an argument by Kesidis et al. [44], albeit in the context of effective capacity.

Consider the stationary fluid process introduced in Section B. Recall that a process is said to be Markov fluid if its time derivative is a function of a continuous-time, finite-state Markov chain. Let $S[0,t]$ be the amount of service offered to a user during the interval $[0,t]$, and suppose that $S[0,t]$ is a Markov fluid process. Let $u_t$ denote the state of the modulating Markov chain, taking value in $\{1,2,\ldots,M\}$. Using our previously established notation, $u_t$ has generator matrix $Q_s$ and invariant distribution $w$. When the modulating chain $u_t$ is in state $m$, we denote the offered service rate by $s_m$. We assume that $0 \leq s_m \leq s_{m+1} < \infty$ for all $m \in \{1,2,\ldots,M-1\}$. Given that the generator matrix $Q_s$ is irreducible and reversible, we can write the effective capacity of this channel as

$$\alpha(\theta) = \lim_{t\to\infty} -\frac{1}{\theta t} \log \mathrm{E}\left[\mathrm{e}^{-\theta \mathrm{S}[0,\mathrm{t}]}\right], \quad 0 < \theta < \infty.$$

We proceed to evaluate $\alpha(\theta)$ explicitly. Define the function

$$\psi_j(\theta,t) = \mathrm{E}_{\mathrm{j}}\left[\mathrm{e}^{-\theta \mathrm{S}[0,\mathrm{t}]}\right] = \mathrm{E}\left[\mathrm{e}^{-\theta \mathrm{S}[0,\mathrm{t}]}\middle|\mathrm{u}_0 = \mathrm{j}\right].$$

For positive $\epsilon \ll 1$, the transition matrix $P_t(\epsilon)$ can be written as

$$P_t(\epsilon) = e^{Q_s \epsilon} = I + \epsilon Q_s + o(\epsilon).$$

Using this notation, the standard backward equation becomes

$$\begin{aligned}\psi_j(\theta,t) &= \mathrm{E}\left[\mathrm{E}\left[\mathrm{e}^{-\theta \mathrm{S}[0,\mathrm{t}]}\middle|\mathrm{u}_\epsilon\right]\middle|\mathrm{u}_0=\mathrm{j}\right] \\ &= \sum_{i=1}^{M} e^{-\theta\epsilon s_j}\psi_i(\theta,t-\epsilon)e^{\epsilon Q_s}(j,i)+o(\epsilon) \\ &= \sum_{i=1}^{M}\left(1-\theta\epsilon s_j\right)\psi_i(\theta,t-\epsilon)\left(I+\epsilon Q_s\right)(j,i)+o(\epsilon).\end{aligned}$$

Rearrange the above equation, we get

$$\begin{aligned}&\frac{\psi_j(\theta,t)-\psi_j(\theta,t-\epsilon)}{\epsilon} \\ &= \psi_j(\theta,t-\epsilon)\left(Q_s(j,j)-\theta s_j\right) \\ &+ \sum_{i\neq j}(1-\theta\epsilon s_j)\psi_i(\theta,t-\epsilon)Q_s(j,i)+\frac{o(\epsilon)}{\epsilon}.\end{aligned} \tag{B.1}$$

As $\epsilon \to 0$, (B.1) becomes

$$\frac{\partial\psi_j(\theta,t)}{\partial t} = \psi_j(\theta,t)\left(Q_s\left(j,j\right)-\theta s_j\right)+\sum_{i\neq j}\psi_i(\theta,t)Q_s(j,i).$$

Defining the diagonal matrix $S = \mathrm{diag}\left(s_1,\ldots,s_M\right)$ and the vector

$$\Psi(\theta,t) = \left(\psi_1(\theta,t),\ldots,\psi_M(\theta,t)\right),$$

we can write the above equations in matrix form as

$$\frac{\partial\Psi(\theta,t)}{\partial t} = (Q_s-\theta S)\Psi(\theta,t). \tag{B.2}$$

This differential equation is subject to the boundary conditions $\Psi(\theta,0)=\mathbf{1}$. It follows that

$$\Psi(\theta,t) = \exp\left(\left(Q_s-\theta S\right)t\right)\mathbf{1}.$$

We can rewrite the effective capacity as

$$\begin{aligned}\alpha(\theta) &= \lim_{t\to\infty} -\frac{1}{\theta t}\log \mathrm{E}\left[\mathrm{e}^{-\theta \mathrm{S}[0,\mathrm{t}]}\right] \\ &= \lim_{t\to\infty} -\frac{1}{\theta t}\log\left(w\exp\left((Q_s-\theta S)t\right)\mathbf{1}\right).\end{aligned}$$

Using the Perron-Frobenius theorem [12], we obtain

$$\alpha(\theta) = -\frac{1}{\theta}\max_i \gamma_i,$$

where $\{\gamma_i\}$ are the eigenvalues of the matrix $Q_s - \theta S$.

Consider the Gilbert-Elliott channel model introduced in Section A. For this channel, the generator matrix $Q_s$ is given by

$$Q_s = \begin{bmatrix} -\lambda & \lambda \\ \mu & -\mu \end{bmatrix}.$$

The characteristic equation of the matrix $(Q_s - \theta S)$ is equal to

$$\begin{aligned}\det\left[\gamma I - (Q_s-\theta S)\right] &= \det\begin{bmatrix} \gamma+\lambda & -\lambda \\ -\mu & \gamma+\mu+\theta R \end{bmatrix} \\ &= \gamma^2 + (\theta R+\lambda+\mu)\gamma + \theta R\lambda.\end{aligned}$$

The maximum eigenvalue of the matrix $(Q_s - \theta S)$ is immediately found to be

$$\max_i \gamma_i = \frac{-(\theta R+\lambda+\mu)+\sqrt{(\theta R+\lambda+\mu)^2-4\theta R\lambda}}{2}.$$

Thus, the effective capacity of the Gilbert-Elliott channel model is

$$\begin{aligned}\alpha(\theta) &= \lim_{t\to\infty} -\frac{1}{\theta t}\log\left(w\exp\left(\begin{bmatrix} -\lambda & \lambda \\ \mu & -\mu-\theta R \end{bmatrix} t\right)\mathbf{1}\right) \\ &= \frac{(\theta R+\lambda+\mu)-\sqrt{(\theta R+\lambda+\mu)^2-4\theta R\lambda}}{2\theta}.\end{aligned} \tag{B.3}$$

Using the relations $\lambda = \kappa e^{-\eta^2}$ and $\mu = \kappa - \kappa e^{-\eta^2}$, the effective capacity function can be expressed as

$$\alpha(\theta) = \frac{\theta R + \kappa - \sqrt{(\theta R + \kappa)^2 - 4\theta R \kappa e^{-\eta^2}}}{2\theta},$$

which coincides with (3.14).

# APPENDIX C

## CONVERGENCE IN EFFECTIVE CAPACITY

In this section, we show that under suitable conditions convergence in distribution implies convergence in effective capacity. Let $r_1, r_2, \ldots$ be a sequence of positive random variables, each with cumulative distribution function $F_{r_i}$. Assume that this sequence converges in distribution to a positive random variable $r$ with cumulative distribution function $F_r$. We claim that, for $\theta > 0$, the equation $\lim_{i\to\infty} \alpha_{r_i}(\theta) = \alpha_r(\theta)$ holds.

First, consider the moment generating function of random variable $r_i$ with $\theta > 0$,

$$\begin{aligned} E\left[e^{-\theta r_i}\right] &= \int_0^\infty e^{-\theta x} dF_{r_i}(x) \\ &= \int_0^\infty \int_x^\infty \theta e^{-\theta y} dy dF_{r_i}(x) \\ &= \int_0^\infty \int_0^y dF_{r_i}(x) \theta e^{-\theta y} dy \\ &= \int_0^\infty F_{r_i}(y) \theta e^{-\theta y} dy. \end{aligned}$$

Interchanging the order of integration above can be justified using Fubini's Theorem [80]. Furthermore, we note that $|F_{r_i}(y)\theta e^{-\theta y}| \leq \theta e^{-\theta y}$ for all values of $i$ and $y$, with

$$\int_0^\infty \theta e^{-\theta y} dy = 1 < \infty.$$

Then, by Lebesgue's Dominated Convergence Theorem [80], we have

$$\begin{aligned}\lim_{i\to\infty} E\left[e^{-\theta r_i}\right] &= \lim_{i\to\infty} \int_0^\infty F_{r_i}(y)\theta e^{-\theta y} dy \\ &= \int_0^\infty \lim_{i\to\infty} \left(F_{r_i}(y)\theta e^{-\theta y}\right) dy \\ &= \int_0^\infty F_r(y)\theta e^{-\theta y} dy \\ &= E\left[e^{-\theta r}\right].\end{aligned}$$

Using the effective capacity of (2.3) and noting that log is a continuous function on $(0, \infty)$, we get

$$\begin{aligned}\lim_{i\to\infty} \alpha_{r_i}(\theta) &= -\lim_{i\to\infty} \frac{1}{\theta T} \log E\left[e^{-\theta r_i}\right] \\ &= -\frac{1}{\theta T} \log\left(\lim_{i\to\infty} E\left[e^{-\theta r_i}\right]\right) \\ &= -\frac{1}{\theta T} \log E\left[e^{-\theta r}\right] = \alpha_r(\theta).\end{aligned}$$

Thus, convergence in distribution implies convergence in effective capacity for all $\theta > 0$. This result provides ground for the approximate expressions of Sections 1 and 2.

# VITA

Lingjia Liu was born in Suzhou, China. His permanent address is: Caihong Yi Cun, Building 30, Room 307, Suzhou, Jiangsu, 215000, China. He received the B.S. degree in Electronic Engineering Department from Shanghai Jiao Tong University, Shanghai, China, in 2003. He started working towards the Ph.D. degree at the Department of Electrical and Computer Engineering, Texas A&M University, College Station in 2003. His general research interests lie in the areas of wireless communication systems, statistical signal processing, queueing theory, and information theory, with emphasis on delay-sensitive communication over wireless systems and networks.

Lingjia Liu won the title of Exceptional Student of Shanghai Jiao Tong University in both rounds of the selection from 2000 to 2002. He is a recipient of the Texas Telecommunications Engineering Consortium (TxTEC) Fellowship from the Department of Electrical and Computer Engineering at Texas A&M University in 2003 – 2004. Lingjia Liu serves as a Technical Program Committee (TPC) member of the *IEEE ChinaCom* 2008.

The typist for this dissertation was Lingjia Liu.