Impact of Mobility and Wireless Channel on the Performance of Wireless Networks

by

Majid Ghaderi

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Computer Science

Waterloo, Ontario, Canada, 2006

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

This thesis studies the impact of mobility and wireless channel characteristics, *i.e.*, variability and high bit-error-rate, on the performance of integrated voice and data wireless systems from network, transport protocol and application perspectives.

From the network perspective, we study the impact of user mobility on radio resource allocation. The goal is to design resource allocation mechanisms that provide seamless mobility for voice calls while being fair to data calls. In particular, we develop a distributed admission control for a general integrated voice and data wireless system. We model the number of active calls in a cell of the network as a Gaussian process with time-dependent mean and variance. The Gaussian model is updated periodically using the information obtained from neighboring cells about their load conditions. We show that the proposed scheme guarantees a prespecified dropping probability for voice calls while being fair to data calls. Furthermore, the scheme is stable, insensitive to user mobility process and robust to load variations.

From the transport protocol perspective, we study the impact of wireless channel variations and rate scheduling on the performance of elastic data traffic carried by TCP. We explore cross-layer optimization of the rate adaptation feature of cellular networks to optimize TCP throughput. We propose a TCP-aware scheduler that switches between two rates as a function of TCP sending rate. We develop a fluid model of the steady-state TCP behavior for such a system and derive analytical expressions for TCP throughput that explicitly account for rate variability as well as the dependency between the scheduler and TCP. The model is used to choose RF layer parameters that, in conjunction with the TCP-aware scheduler, improve long-term TCP throughput in wireless networks. A distinctive feature of our model is its ability to capture variability of round-trip-time, channel rate and packet error probability inherent to wireless communications.

From the application perspective, we study the performance of wireless messaging systems. Two popular wireless applications, the short messaging service and multimedia messaging service are considered. We develop a mathematical model to evaluate the performance of these systems taking into consideration the fact that each message tolerates only a limited amount of waiting time in the system. Using the model, closed-form expressions for critical performance parameters such as message loss, message delay and expiry probability are derived. Furthermore, a simple algorithm is presented to find the optimal temporary storage size that minimizes message delay for a given set of system parameters.

Acknowledgments

I would like to thank my supervisors Srinivasan Keshav and Raouf Boutaba. My special thanks go to Keshav for his support, encouragement and guidance during the course of this work. Keshav has been more than a supervisor to me; he is a friend and a mentor in my life. I am grateful to Keshav for recommending me to Jean Bolot for a summer internship at Sprint Advanced Technology Labs in California. Many thanks to Ashwin Sridharan and Hui Zang at Sprint Labs who spent countless hours discussing my ideas and helping me carry out my research. Part of this thesis was done during my internship at Sprint Labs where I was supervised by Ashwin and Jean. I am grateful to Don Towsley, Rene Cruz and François Baccelli for their insightful discussions and comments about my work. I would also like to thank members of my examining committee, Mostafa Ammar, Martin Karsten, Sherman Shen and Johnny Wong for their time and thoughtful comments.

My sincerest thanks go to Mina, my beloved wife, who contributed in countless ways to the successful completion of this work. I thank Mina for her support, confidence and patience during the hard times we have passed through.

Over the years, I have been fortunate to have the support of many people. I will be always grateful to my parents for their inspiration, encouragement and building on me the spirit of hard work.

I express my deep appreciation to all who have contributed to this work in any way. Thanks to all my friends for their advice and encouragement: Abbas, Amirreza, Bashir, Ehsan, Hadi, Hamidreza, Mahdi, Mahmoud and Yashar. Thanks to my office mate Basem Shihada and my colleagues at the University of Waterloo, Mumtaz Ahmad, Nabeel Ahmed, Reaz Ahmed, Andre Allavena, Mirza Beg, Joe Capka, Shimin Guo, David Hadaller, Anwar Haque, Brent Ishibashi, Darcy Kroeker, Wenli Liu, Earl Oliver, Sumar Ur Rahman, Aditeshwar Seth and Jin Xiao for all those presentations and discussions that helped me shape my ideas.

I would also like to thank Nortel Networks, NSERC, Sprint, the Government of Ontario and the University of Waterloo for their financial support during this work.

Dedication

*To Mina whose vivid presence has been crucial in decisive moments of my life,
and to Navid who brings joy and delight to my soul.*

Contents

List of Tables

List of Figures

Chapter 1

Introduction

In this chapter, we provide a brief overview of wireless cellular networks and the problem considered in this thesis. We discuss problems arising from user mobility and wireless channel variations and how this thesis intends to address those problems.

1.1 Mobile Communications

Starting in 1921 in the United States, police department experimental mobile radios began operating just above the present AM radio broadcast band. On June 17, 1946 in Saint Louis, AT&T and Southwestern Bell introduced the first American commercial mobile telephone service (typically in automobiles). A centrally located antenna, which was installed above Southwestern Bell's headquarters, paged mobiles and provided radio-telephone traffic on the downlink. As early as 1947, it was realized that small *cells* with frequency reuse could increase traffic capacity substantially and the basic *cellular* concept was developed. However, the technology did not exist. In the late 1960s and early 1970s, the cellular concept was used to improve system capacity and radio efficiency. With the development of digital technologies and microprocessors in the late 1980's and up to today, enormous interest emerged in digital cellular systems, which promised higher capacity and higher quality of service (QoS) at reduced costs.

Today, the cell phone is the most popular electronics device in the world and

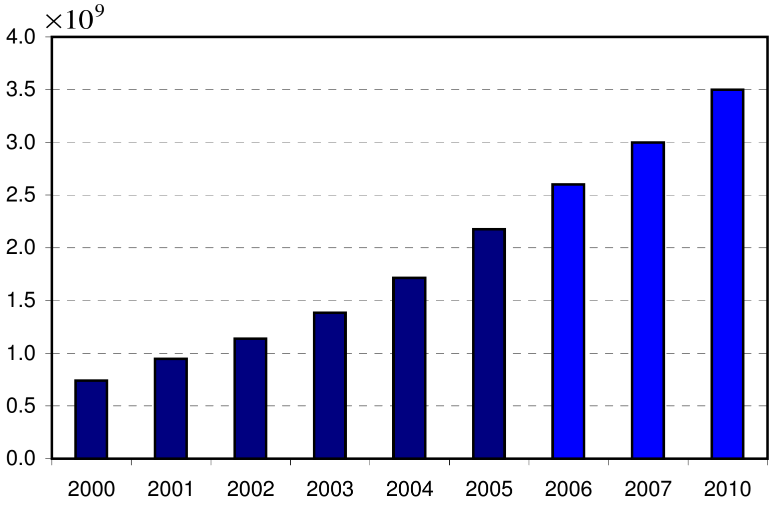


Figure 1.1: Cell phone is the most popular electronics device in the world.

considered the fourth window of content after television, the big screen and the personal computer. Figure 1.1 shows the number of cell phones in the world from year 2000 to 2010[1]. It is expected that by the end of 2007 there will be 3 billions cell phones in the world that will reach 3.5 billion by the end of 2010[2].

Historically, mobile cellular communications have undertaken four evolution stages or generations, which are shown in Table 1.1 taken from [1]. Analog cellular systems belong to the first generation where the major service provided is voice. Second generation cellular systems used digital technologies to provide better quality of service including voice and limited data with higher system capacity and lower cost. Third generation cellular networks offer multimedia transmission, global roaming across a homogeneous wireless network, and bit rates ranging from 384 Kbps to several Mbps. Migration to 3G is already ongoing worldwide. Meanwhile, researchers and vendors are expressing a growing interest in 4G wireless networks that support global roaming across heterogeneous wireless and mobile networks, for example, from a cellular network to a satellite-based network to a high-bandwidth wireless LAN [2–4].

Figure 1.2 shows a simplified architecture of a cellular network. The coverage

[1] Source: GSM Association, `www.gsmworld.com`

[2] Source: Pyramid Research via Yahoo! News, `www.pyr.com`

Property	1G	2G	2.5G&3G	4G
Starting Time	1985	1992	2002	2010-2012
Representative Standard	AMPS	GSM	IMT-2000	UWB
Radio Frequency (Hz)	400M-800M	800M-900M	1800M-2400M	2G-8G
Bandwidth(bps)	2.4K-3K	9.6K-14.4K	384K-2M	20M-100M
Multiple Access Technique	FDMA	TDMA, CDMA	WCDMA	OFDM
Switching Basis	Circuit	Circuit	Circuit,Packet	Packet
Cellular Coverage	Large area	Medium area	Small area	Mini area
Service Type	Voice	Voice, limited data	Voice, data, limited multimedia	Multimedia

Table 1.1: Evolution of mobile communication systems.

area is divided into small regions called cells. Each cell is equipped with a base station and a number of radio channels assigned according to the transmission power constraints and availability of spectrum. A channel can be a frequency, a time slot or a code sequence. Any mobile terminal residing in a cell can communicate through a radio link with the base station located in the cell, which is in turn connected to the core of the network through a base station controller. A group of base stations are controlled by a base station controller. The core network consists of circuit-switched and packet-switched domains. The former basically provides voice service over a circuit-based infrastructure which has evolved from analog technologies to more advanced digital technologies. The latter, on the other hand, is recently added to the infrastructure in order to provide packet-based data services. Both voice calls and data packets follow the same path until the base station controller. In the core, voice calls are routed to Public-Switched Telephone Network (PSTN) via a Mobile Switching Center, while data packets are forwarded through a gateway to the Internet.

1.2 Problem Definition

The impact of mobility and wireless channel characteristics, *i.e.*, variability and high bit-error-rate, on the performance of integrated voice and data wireless systems is the problem considered in this thesis. Typically, cellular networks employ power control mechanisms to combat channel fluctuations for voice calls, whereas,

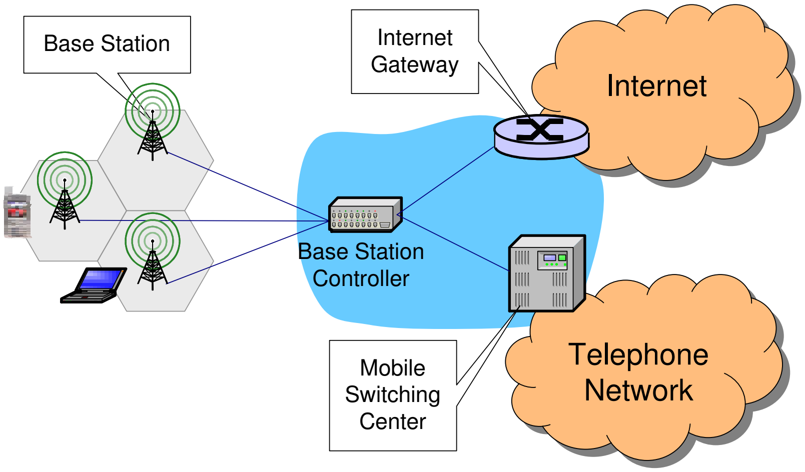


Figure 1.2: A cellular system.

for data calls[3], they take advantage of such fluctuations to improve the overall system performance [5–7]. Data calls employ retransmission mechanisms to hide connection disruptions from applications. As the result, voice calls are more sensitive to interruptions caused by mobility while data calls show more sensitivity to channel fluctuations. Hence, we study the impact of mobility[4] on voice and the impact of wireless channel behavior on data performance. One of the challenges in considering such integrated systems is that the already limited wireless bandwidth has to be shared among voice and data traffics in a fair and efficient manner. The following sub-sections elaborate on the problem.

1.2.1 Impact of User Mobility

User mobility is a distinguishing characteristic of wireless communications and, certainly, the primary factor contributing to their success. When a mobile terminal (mobile user) requests service, it may either be granted or denied service.

[3] For the purpose of this thesis, a *data call* (a flow) is defined as the sequence of packets pertaining to one instance of some application. A data call might correspond to a TCP connection established for the transfer of one element of a Web document or an entire page if this can be identified as a single entity. For the sake of simplicity, we suppose all data calls are elastic.

[4] We are interested in inter-cell mobility in this thesis.

This denial of service is known as call blocking, and its probability as *call blocking probability*. An active terminal[5] in a cellular network may move from one cell to another. The continuity of service to the mobile terminal in the new cell requires a successful *handoff* from the previous cell to the new cell. A handoff is successful if the required resources are available and allocated for the mobile terminal. The probability of a handoff failure is called *handoff failure probability*. During the life of a call, a mobile user may cross several cell boundaries and hence may require several successful handoffs. Failure to get a successful handoff at any cell in the path forces the network to discontinue service to the user. This is known as call dropping or forced termination of the call and the probability of such an event is known as *call dropping probability*.

In general, dropping a call in progress is considered to have a more negative impact from the user's perspective than blocking a newly requested call. However, data calls typically apply time-out and retransmission mechanisms and hence, are more sensitive to blocking than dropping. Consequently, at *call-level*, call dropping and call blocking are the most important quality of service parameters for voice and data calls respectively.

A simple solution for reducing the number of dropped voice calls and hence reducing the call dropping probability is to reserve a portion of available radio resources in each cell to be exclusively used by voice handoffs. Cellular systems have adopted this strategy by over-provisioning their radio resources. Typically, a cellular system is designed to accommodate traffic loads higher than its typical load. The excess capacity is then implicitly reserved for voice handoffs. Therefore, there is often sufficient resources to accommodate incoming handoffs. Indeed, this over-provisioning design, to some extent, works for traditional cellular systems supporting circuit-based voice calls.

However, as depicted in Figure 1.1, the number of wireless users is growing very fast and their desire for high-bandwidth multimedia services will push the existing systems to their limits. To increase the capacity of cellular networks, micro/pico cellular architectures are deployed in the current systems. Smaller cell size of these architectures leads to a higher handoff rate. Therefore, modern cellular systems

[5] An active terminal is a terminal with an active call.

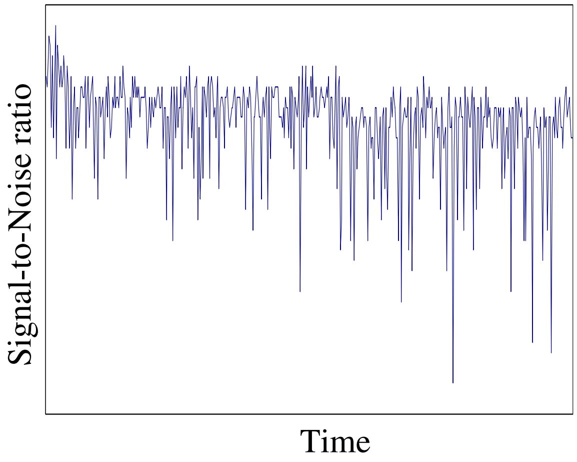


Figure 1.3: Wireless channel is variable.

suffer from problems arising from user mobility even more than traditional macro cellular systems. Furthermore, wireless bandwidth is scarce and expensive compared to abundant bandwidth of wired networks[6]. The cost of obtaining licensed spectrum is so high that there is simply no margin for under-utilization. A naive use of the over-provisioning approach will waste radio resources and result in high blocking probability for data call.

1.2.2 Impact of Wireless Channel

Wireless networks suffer from wireless channel fluctuations. Wireless channels have variable capacity and are subject to time-varying and location-dependent errors. Figure 1.3 illustrates the signal quality received at a mobile receiver communicating over a wireless channel. The signal quality fluctuates over time and sometimes there is basically no detectable signal at the receiver.

Recently, data services have been deployed in wireless networks and gained significant popularity among mobile subscribers. Existing wireless data services rely on traditional Internet protocols, *i.e.*, TCP/IP, to preserve compatibility with legacy applications and software technologies developed for wired networks. In particular, TCP is still the dominant transport protocol over both wired and wire-

[6] A single fiber can carry traffic at the speed of several Terabits/second.

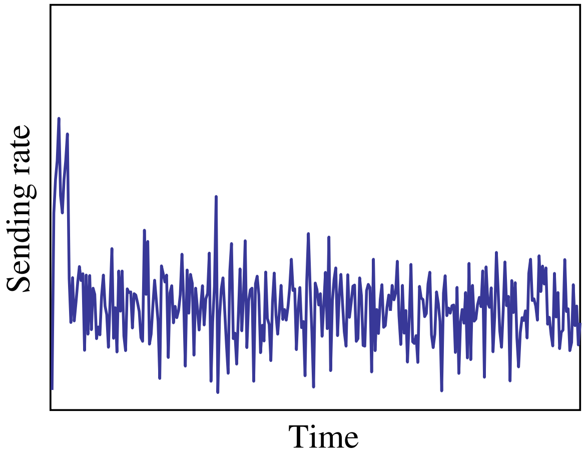


Figure 1.4: Data traffic is bursty.

less links despite being known to be unsuitable for wireless networks. There is a mismatch between TCP and wireless networks: TCP is not designed to work over wireless channels and wireless networks are not designed with TCP congestion control mechanism in mind. Cross-layer optimization is a promising approach to tackle the shortcomings of TCP. It is important to adapt lower layer protocols to TCP dynamics in order to improve TCP performance in wireless networks.

Modern cellular networks incorporate RF technology that allows them to dynamically vary the wireless channel rate in response to user demand and channel conditions. However, the set of data rates as well as the scheduler policy are typically chosen to optimize system throughput at radio link layer. Such a system configuration may not result in optimal throughput at TCP layer. To optimize system throughput at TCP layer, the set of data rates as well as the scheduler policy must be *aware* of TCP dynamics. This requires accurate models that capture TCP dynamics and their interactions with lower layer protocols.

Finally, resource management mechanisms, *e.g.*, admission control, need to know how much resources are required to support an application at some minimum throughput level. For example, such information is required for fair bandwidth sharing between voice and data traffic in integrated systems discussed in subsection 1.2.1. However, the variable nature of wireless channels and the burstiness of data traffic makes the computation of resource requirements of data applications

(Figure 1.4 shows the sending rate of a TCP flow in a wireless environment) a challenging problem. This further motivates the need for accurate models of TCP throughput in wireless networks.

1.3 Thesis Objectives

The primary goal of this thesis is to study the impact of mobility and wireless channel characteristics, *i.e.*, high bit-error-rate and random variability, on the performance of integrated wireless systems. The goal is to address the following questions. First, how should the scarce radio resources be allocated to mobile users? Any answer to this question effectively determines system capacity with respect to the number of voice and data users that can be simultaneously supported by the network. Second, how should user traffic be carried over wireless channels to improve user traffic performance? Understanding the interactions between transport protocol and radio link layer mechanisms is essential in answering this question. Finally, how is user experience at application level? No matter how sophisticated a wireless system is, what really matters for end users is the application-level performance of the system. To address these questions, wireless systems will be studied from three perspectives: (1) from network perspective to study performance issues related to resource and mobility management, (2) from transport perspective to study the behavior of TCP in wireless environments, and, (3) from application perspective to study the performance of wireless messaging systems. Below is a brief description of the issues addressed in this thesis.

Network Perspective:

Call dropping and call blocking are the most important issues from network perspective for voice and data calls respectively. To support seamless mobility of voice calls, handoff process is guaranteed to be successful without degradation of the service quality. Efficient radio resource management to support seamless mobility for voice calls and prevent starvation of data calls is the problem we intend to address. Although several system components are involved in resource management, *e.g.*, channel assignment, we will focus on *admission control* in this thesis. The goal is to design efficient admission control algorithms that maximize wireless bandwidth utilization subject to pre-specified constraints on call dropping and blocking probabilities for voice and data calls. Such algorithms provide seamless mobility for voice calls while being fair to data calls.

Transport Perspective:

TCP throughput is the most important issue from transport perspective. We explore cross-layer optimization of the rate adaptation feature of cellular networks to optimize TCP throughput. The goal of this study is to: (i) model TCP dynamics in wireless networks with respect to the behavior of wireless channel and downlink scheduler, and, (ii) develop new scheduling techniques that take into consideration TCP dynamics in order to maximize system throughput.

Application Perspective:

No matter how sophisticated a wireless system is, what really matters for end users is the application-level performance of the system. Although we study TCP throughput in wireless networks, there are other wireless applications such as mobile messaging (text messaging and picture messaging) that do not rely on TCP. It is estimated that a worldwide total of 1 trillion text messages were sent in 2005[7]. With the increasing size and volume of messages being transmitted, the fast and robust delivery of messages becomes a challenging problem. We study mobile messaging systems, as the most popular wireless applications, where message delay and loss probability are the relevant quality of service measures from the application perspective. Our goal is to: (i) develop a mathematical model for evaluating the performance of mobile messaging systems, and, (ii) identify system bottlenecks and propose techniques for proper dimensioning of those bottlenecks.

1.4 Thesis Organization

The rest of this thesis is organized as follows. Chapter 2 studies the impact of mobility from network perspective. A general wireless system integrating voice and

[7] Source: GSM Association, `www.gsmworld.com`

data is considered where radio resources are completely shared among voice and data traffic, *i.e.*, no hard partitioning. We develop a call admission control scheme that guarantees a pre-specified call dropping probability for voice calls while being fair to data calls (and hence, preventing data calls from starvation). The idea is to utilize information about traffic condition in neighboring cells in order to estimate future handoff load in a cell. A distinguishing feature of the proposed scheme is its adaptability to traffic load fluctuations.

Chapter 3 is dedicated to study the impact of wireless channel from transport perspective. There, we explore cross-layer optimization of the rate adaptation feature of cellular networks to optimize TCP throughput. We propose a two-state TCP-aware scheduler that switches between two rates as a function of the TCP sending rate. We develop a fluid model of the steady-state TCP behavior for such a system and derive analytical expressions for TCP throughput that explicitly account for rate variability as well as the dependency between the scheduler and TCP. The model is then used to choose RF layer parameters that, in conjunction with the TCP-aware scheduler, improve long-term TCP throughput in wireless networks.

In Chapter 4, we study the performance of wireless networks from application perspective. As two popular wireless applications, we study short messaging service (SMS) [8] and multimedia messaging service (MMS) [9]. We develop a mathematical model to evaluate the performance of such systems. Using the model, closed-form expressions for major performance parameters such as message loss and message delay are derived and used for dimensioning temporary storage at message centers.

Finally, a summary of the thesis and some future research directions are presented in Chapter 5.

Chapter 2

Network Perspective: Admission Control

This chapter addresses bandwidth allocation in an integrated voice/data broadband wireless network supporting seamless mobility. Specifically, we propose a new admission control scheme called EFGC, which is an extension of the well-known fractional guard channel scheme proposed for cellular networks supporting voice traffic. The main idea is to use two acceptance ratios, one for voice calls and the other for data calls in order to maintain fairness between voice and data traffic while guaranteeing a target handoff failure probability for voice calls. We describe two variations of the proposed scheme: EFGC-REST, a conservative approach which aims at preserving fairness by sacrificing the bandwidth utilization; and EFGC-UTIL, a greedy approach which achieves higher bandwidth utilization at the expense of increasing the handoff failure probability for voice calls. Extensive simulation results show that our schemes satisfy the hard constraints on handoff failure probability and fairness while maintaining a high bandwidth utilization.

2.1 Chapter Organization

The rest of this chapter is organized as follows. Section 2.2 is an introduction to the problem considered in this chapter. In Section 2.3, we briefly review existing research on admission control in cellular networks. Our system model, assumptions and notations are described in section 2.4. Section 2.5 is dedicated to the proposed admission control algorithm and presents the analysis of the proposed algorithm in detail. In section 2.6, we discuss the estimation of control parameters such as arrival rates, then we address the multiple handoffs problem and control interval length. Extensive simulation results and their analysis are presented in section 2.7. Finally, section 2.8 concludes this chapter.

2.2 Introduction

Emerging wireless technologies such as 3G and 4G will increase the cell capacity of wireless cellular networks to several Mbps [2]. With this expansion of wireless bandwidth, the next generations of mobile cellular networks are expected to support diverse applications such as voice, data and multimedia with varying quality of service (QoS) and bandwidth requirements [3]. Wireless links bandwidth is limited and is generally much smaller than that of wireline access links. Therefore, for integrated voice/data mobile networks it is necessary to develop mechanisms that can provide effective bandwidth management while satisfying the QoS requirements of both types of traffic.

At call-level, two important quality of service parameters are the call blocking probability (p_b) and the call dropping probability (p_d). Since dropping a call in progress has a more negative impact from the user perspective, handoff calls are given higher priority than new calls in accessing the wireless resources. This preferential treatment of handoffs increases the probability of blocking new calls and hence may degrade the bandwidth utilization. The most popular approach to prioritize handoff calls over new calls is by reserving a portion of available bandwidth in each cell to be used exclusively for handoffs. Based on this idea, a number of call admission control (CAC) schemes have been proposed which basically differ from each other in the way they calculate the reservation threshold [10–15].

Bandwidth allocation has been extensively studied in single-service (voice) wireless cellular networks. Hong and Rappaport [10] are the first who systematically analyzed the famous *guard channel* (GC) scheme, which is currently deployed in cellular networks supporting voice calls. Ramjee *et al.* [16] have formally defined and categorized the admission control problem in cellular networks. They showed that the guard channel scheme is optimal for minimizing a linear objective function of call blocking and dropping probabilities while the *fractional guard channel* scheme (FGC) [16] is optimal for minimizing call blocking probability subject to a hard constraint on call dropping probability. Instead of explicit bandwidth reservation as in GC, the FGC accepts new calls according to a randomization parameter called the *acceptance ratio*. One advantage of FGC over GC is that it distributes the new accepted calls evenly over time which leads to a more stable control [17].

Because of user mobility, it is impossible to describe the state of the system by using only local information, unless we assume that the network is uniform and approximate the overall state of the system by the state of a single cell in isolation. To include the global effect of mobility, *collaborative* or *distributed* admission control schemes have been proposed [11–15, 17, 18]. Information exchange among a cluster of neighboring cells is the approach adopted by all distributed schemes.

In particular, Naghshineh and Schwartz [11] proposed a collaborative admission control known as *distributed call admission control* (DCA). DCA periodically gathers information, namely the number of active calls, from the adjacent cells to make, in combination with the local information, the admission decision. It has been shown that DCA is not stable and violates the required dropping probability as the load increases [17]. Levin *et al.* [12] proposed a more sophisticated version of the original DCA based on the *shadow cluster* concept, which uses dynamic clusters for each user based on its mobility pattern instead of restricting itself (as DCA) to direct neighbors only. A practical limitation of the shadow cluster scheme in addition to its complexity and inherent overhead is that it requires a precise knowledge of the mobile's trajectory. Recently, Wu *et al.* [17] proposed a distributed scheme called SDCA based on the classical fractional guard channel scheme which can precisely achieve the target call dropping probability. A key feature of SDCA is the formulation of the time-dependent call dropping probability which can be computed by the diffusion approximation of the channel occupancy.

One of the challenges in considering multi-services systems is that the already limited bandwidth has to be shared among multiple traffics. Epstein and Schwartz [19] investigated complete sharing, complete partitioning and hybrid reservation schemes for two classes of traffic, namely narrow-band and wide-band traffic. In general, complete sharing strategy achieves the highest bandwidth utilization [19].

Fixed and movable boundary schemes for bandwidth allocation in wireless networks were studied by Wieselthier and Ephremides [20]. They concluded that movable boundary schemes can achieve a better utilization than fixed boundary schemes for voice and data integration. Since then, a number of papers have been published focusing on the performance of fixed and movable boundary schemes given different assumptions and network configurations [21–27].

In particular, Haung *et al.* [25] proposed a bandwidth allocation scheme for voice/data integration based on the idea of movable boundaries (MB). In their scheme, bandwidth is divided into two portions that can be dynamically adjusted to achieve the desired performance. However, they completely neglected the prioritization of handoff calls over new calls and treated the two identically. Yin *et al.* [26] proposed a *dual threshold reservation* (DTR) scheme, which extends the basic guard channel to use two thresholds, one for reserving channels for voice handoff, and the other for limiting the data traffic into the network in order to preserve the voice performance. An extended version of DTR which implements queueing for data calls (DTR-Q) was proposed in [27]. In general, queueing of new/handoff calls, can further improve the performance of call admission control [28]. The main limitation of DTR (DTR-Q) is that it is static, *i.e.*, the two reservation thresholds are fixed over time regardless of the state of the network. Interested readers are referred to [29] for a comparison between DTR and MB schemes.

This chapter introduces an *extended fractional guard channel call admission mechanism* (EFGC) for integrated voice and data mobile cellular networks. EFGC maximizes the wireless bandwidth utilization while satisfying a target call dropping probability and a relative voice/data service differentiation. The main idea is to use two acceptance ratios for voice and data according to the desired dropping probability of voice calls and the relative priority of voice calls over data calls. Similar to [22–27], we assume that call dropping is not an important issue for data calls and treat handoff and new data calls in the same way. We define the extended MINBLOCK [16] problem as follows:

> *for a given cell capacity, maximize the bandwidth utilization subject to a hard constraint on the voice call dropping probability and relative*

voice/data call blocking probability.

To the best of our knowledge, extending the basic fractional guard channel scheme to address the extended MINBLOCK problem is a novel work. We follow an approach similar to the admission control algorithm proposed by Wu *et al.* [17] to derive the acceptance ratios for voice and data calls. The main features of EFGC are as follows:

1. EFGC is dynamic, therefore, adapts to a wide range of system parameters and traffic conditions.

2. EFGC uses separate acceptance ratios for voice and data calls, therefore, it is very straightforward to enforce a relative or even strict service differentiation between voice and data traffic.

3. EFGC is distributed and takes into consideration the information from direct neighboring cells in making admission decisions.

4. The control mechanism is stochastic and periodical to reduce the overhead associated with distributed control schemes. EFGC determines the appropriate control parameters such as the control interval length in order to restrict the impact of the network to the direct neighbors only.

2.3 Admission Control in Cellular Networks

In this section, we provide a survey on CAC schemes proposed for cellular networks. Interested readers are referred to [30] for a comprehensive survey on the topic. Figure 2.1 depicts a classification of CAC schemes proposed for cellular networks. As mentioned before, we are interested in prioritized scheme in which handoff calls are given priority over new calls in access to network resources in order to reduce call dropping probability. We briefly discuss channel borrowing, call queueing and reservation schemes as handoff prioritization schemes. We then focus on reservation schemes as the most common prioritization techniques.

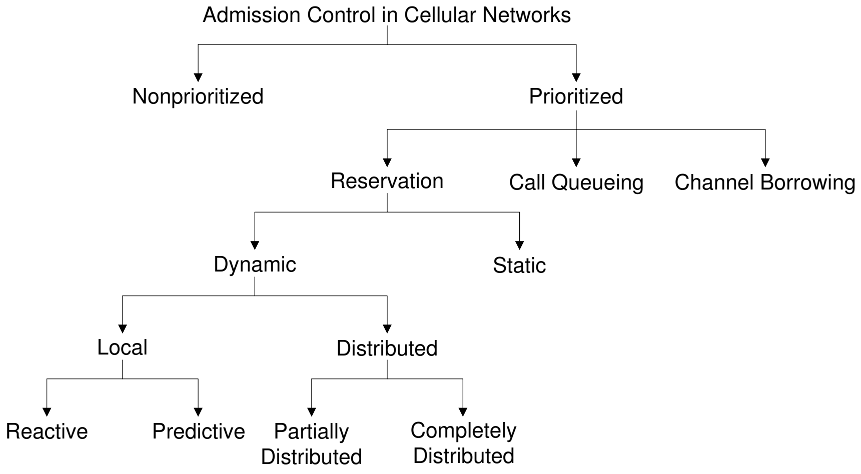


Figure 2.1: Stochastic call admission control schemes in cellular networks.

2.3.1 Channel Borrowing Schemes

In a channel borrowing scheme, a cell (an acceptor) that has used all its assigned channels can borrow free channels from its neighboring cells (donors) to accommodate handoffs [31–33]. A channel can be borrowed by a cell if the borrowed channel does not interfere with existing calls. When a channel is borrowed, several other cells are prohibited from using it. This is called channel locking and has a great impact on the performance of channel borrowing schemes [34]. The number of such cells depends on the cell layout and the initial channel allocation. For example, for a hexagonal planar layout with reuse distance of one cell, a borrowed channel is locked in three neighboring cells. Channel borrowing schemes differ in the way a free channel is selected from a donor cell to be borrowed by an acceptor cell. A complete survey on channel borrowing schemes is provided by Katzela and Naghshinehin [31].

2.3.2 Call Queueing Schemes

Queueing of handoff requests, when there is no channel available, can reduce the dropping probability at the expense of higher new call blocking. If the handoff attempt finds all the channels in the target cell occupied it can be queued. If any channel is released it is assigned to the next handoff waiting in the queue. Queueing

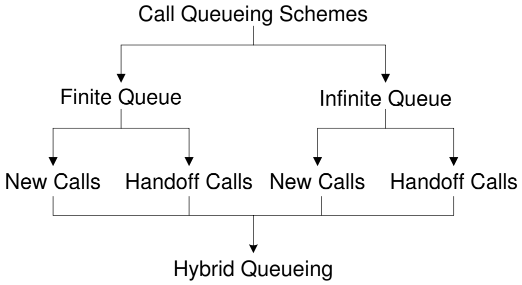


Figure 2.2: Call queueing schemes.

can be done for any combination of new and handoff calls. The queue itself can be finite [28] or infinite [10]. Although finite queue systems are more realistic, systems with infinite queue are more convenient for analysis. Figure 2.2 depicts a classification of call queueing schemes.

2.3.3 Reservation Schemes

The notion of guard channels was introduced in the mid 80s as a call admission control mechanism to give priority to handoff calls over new calls. In this policy, a set of channels called the guard channels are permanently reserved for handoff calls. Hong and Rappaport [10] showed that this scheme reduces handoff dropping probability significantly compared to the nonprioritized case. They found that p_d decreases by a significantly larger order of magnitude compared to the increase of p_b when more priority is given to handoff calls by increasing the number of handoff channels.

A critical parameter in this basic scheme is the optimal number of guard channels. In fact, there is a tradeoff between minimizing p_d and minimizing p_b. If the number of guard channels is conservatively chosen then admission control fails to satisfy the specified p_d. A static reservation typically results in poor resource utilization. To deal with this problem, several dynamic reservation schemes [11–14,35] were proposed in which the optimal number of guard channels is adjusted dynamically based on the observed traffic load and dropping rate in a control time window. If the observed dropping rate is above the guaranteed p_d then the number of reserved channels is increased. On the other hand, if the current dropping rate is far below the target p_d then the number of reserved channels is decreased. The next section investigates dynamic reservation schemes.

2.3.4 Dynamic Reservation Schemes

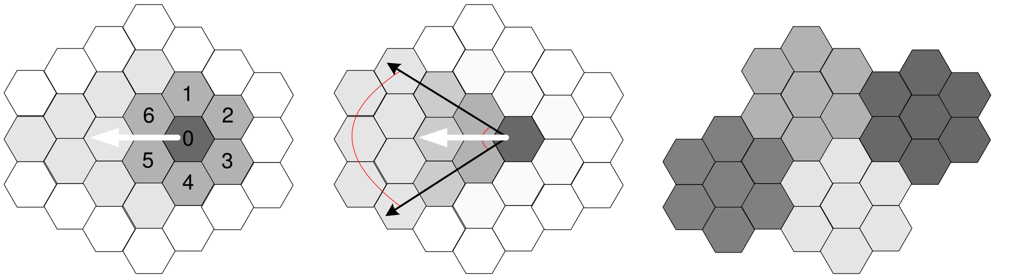
There are two approaches in dynamic reservation schemes, namely, local and distributed (collaborative), depending on whether they use local information or gather information from neighbors to adjust the reservation threshold. In local schemes, each cell estimates the state of the network using local information only, while in distributed schemes each cell gathers network state information in collaboration with its neighboring cells.

Local Schemes

We categorize local admission control schemes into *reactive* and *predictive* schemes. By reactive approaches we refer to those admission policies that adjust their decision parameters, *i.e.*, threshold and reservation level, as a result of an event such as call arrival, completion or rejection. The well-known guard channel (cell threshold, cut-off priority or trunk reservation) scheme (GC) scheme and all its extensions (*e.g.*,two threshold scheme [35]) fall in this category. Predictive approaches refer to those policies that predict future events and adjust their parameters in advance to prevent undesirable QoS degradations. Reactive schemes can be further categorized into *parameter-based* schemes [10, 16] and *measurement-based* schemes [36–38]. Measurement-based schemes are particularly promising for packet-based traffic where parameter-based schemes are known to be inefficient [39].

Distributed Schemes

The fundamental idea behind all distributed schemes [11–15, 17, 40] is that every mobile terminal with an active connection exerts an influence upon the cells in the vicinity of its current location and along its direction of travel [12]. A group of cells which are geographically or logically close together form a *cluster*, as shown in Figure 2.3. Either each mobile terminal has its own cluster independent of other terminals or all the terminals in a cell share the same cluster. Typically, the



(a) Shadow cluster [12]. (b) Most likely cluster [15]. (c) Virtual connection tree [41].

Figure 2.3: Three examples of cluster definition.

admission decision for a connection request is made in cooperation with other cells of the cluster associated to the mobile terminal asking for admission. In Figure 2.3(a) a cluster is defined assuming that a terminal affects all the cells in the vicinity of its current location and along its trajectory, while in Figure 2.3(b) it is assumed that those cells that form a sector in the direction of mobile terminal's trajectory are most likely to be affected (visited) by the terminal. And, Figure 2.3(c) shows a static cluster which is fixed regardless of the terminal mobility.

In general, distributed CACs can be categorized into *partially distributed* or *completely distributed* based on the involvement of cells in the decision making process. In partially distributed schemes, all the necessary information is gathered from the neighboring cells, but the processing is local. The virtual connection tree concept introduced in [41] is an example of an implicitly distributed scheme. Despite the fact that information is gathered from a set of neighboring cells, the final decision is made locally in the network controller. In completely distributed schemes, not only information is gathered from the neighboring cells, but also the neighboring cells are involved in the decision making process. The shadow cluster concept introduced in [12] is an example of an explicitly distributed scheme.

2.3.5 Performance Comparison

Looking at existing CAC schemes, there are many assumptions and parameters involved in each scheme. Therefore, it is extremely difficult to develop a unified

CAC scheme		Efficiency	Overhead	Complexity	Adaptivity
Local	Reactive	Low	Low	Low	Moderate
	Predictive	Moderate	Low	Moderate	Moderate
Distributed	Implicit	High	Very High	High	High
	Explicit	High	High	Very High	High

Table 2.1: Comparison of dynamic CAC schemes.

framework for evaluating and comparing the performance of CAC schemes using analytical or simulation techniques. For the comparison purposes, we do not use quantitative values for these criteria instead we use qualitative values. These qualitative values, *e.g.*,"Very High", "High", "Moderate" and "Low", are sufficient for a relative comparison of the CAC schemes discussed in this section.

We use the following criteria in our comparison:

1. *Efficiency:* Efficiency refers to the achieved utilization level of network capacity given a specific set of QoS requirements.

2. *Complexity:* Shows the computational complexity of a CAC scheme for a given network configuration, mobility patterns, and traffic parameters.

3. *Overhead:* Refers to the signalling overhead induced by a CAC scheme on the fixed interconnection network among base stations.

4. *Adaptivity:* Defined as the ability of a CAC scheme to react to changing network conditions, *i.e.*,traffic load changes.

5. *Stability:* Stability is the CAC insensitivity to short term traffic fluctuations. If an adaptive CAC reacts too fast to any load change then it may lead to unstable control. For example during a period of time all connection requests are accepted until a congestion occurs and then all requests are rejected.

Table 2.1 shows a comparison of different dynamic CAC schemes. In general, there is a tradeoff between the efficiency and the complexity of local and distributed schemes. Table 2.2 compares three major distributed CAC schemes. In this table, *basic distributed* was proposed by Naghshineh and Schwartz [11], *shadow cluster* refers to the work of Levin *et al.* [12] and *stable dynamic* is due to Wu *et al.* [17].

CAC scheme	Efficiency	Complexity	Stability
Basic distributed	Moderate	Moderate	Moderate
Shadow cluster	High	High	Moderate
Stable dynamic	Very High	High	High

Table 2.2: Comparison of distributed CAC schemes.

2.4 System Model

As shown in Figure 2.4, we consider a cellular system which carries both voice and data traffic. We assume that wireless bandwidth is channelized where a channel can be a frequency, a time slot or a code sequence. We define the basic bandwidth unit (BU) as the smallest amount of bandwidth that can be allocated to a call, *e.g.*, a channel. In this chapter we focus on call-level QoS parameters, therefore only call-level traffic dynamics are required for resource allocation and admission control. More specifically, we assume that the *effective bandwidth* [42–44] concept is applied to each call. When employing this concept, an appropriate effective bandwidth is assigned to each call and each call is treated as if it required this effective bandwidth throughout the active period of the call. The feasibility of admitting a given set of connections may then be determined by ensuring that the sum of the effective bandwidths is less than or equal to the total available bandwidth, *i.e.*, the cell capacity.

We assume that each voice call requires b_v BUs and each data call requires b_d BUs for the whole duration of the call. In the system under consideration, voice handoff calls have the highest priority, then come new voice calls, and lastly the new and handoff data calls are considered. As mentioned earlier, there is no prioritization of handoff data calls, and hence handoff data calls are treated the same as new data calls.

The considered system is not required to be uniform. Each cell can experience a different load, *e.g.*, some cells can be over-utilized while others are under-utilized. Let $k = \{v, d\}$ denote the type of traffic, *i.e.*, $k = v$ for voice and $k = d$ for data traffic. Below is the notation which will be used throughout this chapter.

- M: number of cells in the network

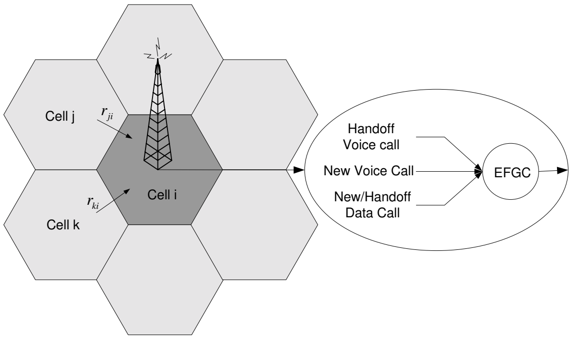


Figure 2.4: Integration of voice and data at the base station of a cellular network.

- $\mathcal{A}_i$: the set of the adjacent cells of cell i
- c_i: the capacity of cell i in terms of BUs
- $R_i(t)$: bandwidth requirements (used capacity) in cell i at time t in terms of BUs
- p_{f_i}: voice handoff failure probability in cell i
- p_{QoS}: target voice handoff failure probability to be guaranteed
- k: the service index for voice and data with $k = v$ for voice and $k = d$ for data
- λ_i^k: type-k new call arrival rate into cell i
- $1/\mu_k$: type-k mean call duration
- $1/h_k$: type-k mean cell residency time
- T: length of the control period
- b_k: bandwidth requirement of type-k calls in terms of BUs
- $N_i^k(t)$: number of active type-k calls in cell i at time t
- r_{ji}: routing probability from cell $j \in \mathcal{A}_i$ to cell i
- b_i^k: type-k call blocking probability in cell i

- a_i^k: type-k call acceptance ratio in cell i
- α_i: relative priority of voice traffic over data traffic in cell i defined as $\alpha_i = a_i^v/a_i^d$
- α_{QoS}: target relative priority of voice traffic over data traffic to be guaranteed
- p_b^k: network-wide type-k call blocking probability
- p_d: network-wide voice call dropping probability
- $E[z]$: the mean of random variable z
- $V[z]$: the variance of random variable z
- $\tilde{z}$: time-averaged value of random variable z
- $\hat{z}$: measured (observed) value of random variable z

Let random variables t_{d_k} and t_{r_k} denote the call duration (call holding time) and cell residency time of a typical type-k call, respectively. Note that t_{d_k} and t_{r_k} are independent random variables. Similar to [10, 16, 17, 19–29], we assume that t_{d_k} and t_{r_k} are exponentially distributed. In the real world, the cell residence time distribution may not be exponential but exponential distributions provide the mean value analysis, which indicates the performance trend of the system. Furthermore, our proposed admission control algorithm involves a periodic control where the length of the control period is set to much less than the average cell residency time of a call to make the algorithm insensitive to this assumption.

2.4.1 Multiple Handoffs Probability

As mentioned earlier, in order to make the optimal admission decision, distributed schemes regularly exchange some information with other cells in the network. Those cells involved in the information exchange form a *cluster*. Due to the intercell information exchange, base station interconnection network incurs a high signalling overhead. Moreover, as the cluster size increases the operational complexity of the control algorithm increases too. In particular, two major factors affect the overhead and complexity of distributed CAC schemes; (1) frequency of information exchange, and, (2) depth of information exchange, *i.e.*, how many cells away information is exchanged.

To reduce the overhead, distributed CAC schemes typically have a periodic structure in which only at the beginning of control periods information exchange is triggered. Moreover, information exchange is typically restricted to a cluster of neighboring cells. Note that, if the control interval is too small then frequent communications increases the signalling overhead. On the other hand, if the control period is too long then the state information stored locally may become stale. Similarly, if the cluster is too small then the exchanged information will poorly reflects the state of the network. On the other hand, a big cluster will lead to higher overhead. An efficient CAC scheme must compromise between the frequency and depth of information exchange.

In this work, we set the control interval in such a way that the probability of having multiple handoffs in one control period becomes negligible. Therefore, we can effectively assume that only those cells directly connected to a cell can influence the number of calls in that cell during a control period. In a sense, we reduce the control interval in favor of a smaller cluster size. We claim that using this technique, the signalling overhead will not increase, while the collected information on the network status will be sufficiently accurate for the purpose of a stochastic admission control. The reason is that: first, by decreasing the control interval, the probability of multiple handoffs decays to zero exponentially (see section 2.6.3); second, a cluster shrinks quadratically with decreasing the depth of information exchange (see below).

Without loss of generality, consider a symmetric network where each cell has exactly $\mathcal{A}$ neighbors. Consider cell i and all the cells around it forming circular layers as shown in Figure 2.5. From cell i, all the cells up to layer n are accessible with n handoffs assuming that cell i forms layer 0. The number of cells reachable by n handoffs from cell i denoted by $M(n)$ is given by

$$\begin{aligned} M(n) &= 1 + \mathcal{A} + \cdots + n\mathcal{A} \\ &= 1 + \frac{1}{2}n(n+1)\mathcal{A}\,. \end{aligned} \tag{2.1}$$

Therefore, by slightly reducing the control interval, we essentially achieve the

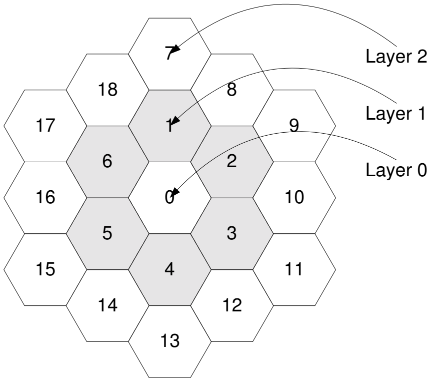


Figure 2.5: A cellular system with 3 layers.

same control accuracy but with reduced signalling overhead. The problem of choosing the proper control interval will be further addressed in section 2.6.3.

2.4.2 Handoff Failure and Call Dropping Probabilities

Although call dropping probability is more meaningful for mobile users and service providers, calculating the handoff failure probability is more convenient. Therefore, our calculations are based on the handoff failure probability, p_f, which can be related to the call dropping probability, p_d, as follows (refer to [10] for more details):

$$p_d = \sum_{H=1}^{\infty} (P_h^v)^H (1-p_f)^{H-1} p_f = \frac{P_h^v p_f}{1 - P_h^v (1-p_f)}, \tag{2.2}$$

where H is the number of possible handoffs during the life of a call, and P_h^v is the handoff probability of a voice call before the call completes which can be computed by the following equation:

$$\begin{aligned} P_h^v &= \Pr(t_{d_v} > t_{r_v}) \\ &= \int_{t=0}^{\infty} \Pr(t_{d_v} > t_{r_v} | t_{r_v}) \Pr(t_{r_v} = t)\, dt \\ &= \int_{t=0}^{\infty} h_v \exp(-\mu_v t) \exp(-h_v t)\, dt = \frac{h_v}{\mu_v + h_v} \end{aligned} \tag{2.3}$$

therefore,

$$p_f = \frac{p_d}{1-p_d} \left(\frac{\mu_v}{h_v} \right) . \tag{2.4}$$

It means that for a given p_d, the equivalent p_f can be easily computed based on (2.4). Therefore, in this work it is assumed that a target handoff failure probability p_{QoS} must be guaranteed for voice calls. Notice that, exponential assumption is a necessary condition in deriving (2.3). Interested readers are referred to [45, 46] for the handoff probability under general call duration and cell residency distributions.

2.4.3 Time-Dependent Handoff and Stay Probabilities

We compute here some useful probabilities required for the rest of our discussion. Let $P_h^k(t)$ denote the probability that a type-k call hands off by time t and remains active until t, given that it has been active at time 0. Also, let $P_s^k(t)$ denote the probability that a type-k call remains active in its home cell until time t, given that it has been active at time 0. Then,

$$\begin{aligned} P_h^k(t) &= \Pr(t_{r_k} \leq t)\ \Pr(t_{d_k} > t) \\ &= (1 - \exp(-h_k t))\ \exp(-\mu_k t), \end{aligned} \tag{2.5}$$

and,

$$\begin{aligned} P_s^k(t) &= \Pr(t_{r_k} > t)\ \Pr(t_{d_k} > t) \\ &= \exp(-(\mu_k + h_k)t)\,. \end{aligned} \tag{2.6}$$

These equations are valid as far as the memoryless property of call duration and cell residency is satisfied. On average, for any call which arrives at time $t' \in (0, t]$, the average handoff and stay probabilities $\tilde{P}_h^k$ and $\tilde{P}_s^k$ are expressed as

$$\tilde{P}_h^k(t) = \frac{1}{t}\int_0^t P_h^k(t - t')\, dt', \tag{2.7}$$

$$\tilde{P}_s^k(t) = \frac{1}{t}\int_0^t P_s^k(t - t')\, dt'\,. \tag{2.8}$$

These integrals can be easily computed with respect to (2.5) and (2.6). Finally, let $P_{ji}^k(t)$ denote the time-dependent handoff probability and $\tilde{P}_{ji}^k(t)$ denote the average time-dependent handoff probability from cell j to cell i where $j \in \mathcal{A}_i$. It is obtained that

$$P_{ji}^v(t) = P_h^v(t)\, r_{ji}, \tag{2.9}$$

$$\tilde{P}_{ji}^v(t) = \tilde{P}_h^v(t)\, r_{ji}, \tag{2.10}$$

because voice handoff calls are always accepted if there is enough free bandwidth. Similarly,

$$P^d_{ji}(t) = a^d_i \big[P^d_h(t)\, r_{ji}\big], \tag{2.11}$$

$$\tilde{P}^d_{ji}(t) = a^d_i \big[\tilde{P}^d_h(t)\, r_{ji}\big], \tag{2.12}$$

because data calls are always subject to an acceptance ratio a^d_i in cell i.

In next section, we will use the computed probabilities to find the maximum acceptance ratios for voice and data calls with respect to the prespecified call dropping probability (p_{QoS}) and relative voice/data acceptance probability (α_{QoS}).

2.5 Admission Control Algorithm

The proposed distributed algorithm, EFGC, consists of two components. The first component is responsible for retrieving the required information from the neighboring cells and computing the control parameters. Using the computed control parameters, the second component enforces the admission control locally in each cell. The following sections describe these two components in detail.

2.5.1 Distributed Control Algorithm

As mentioned earlier, to reduce the signalling overhead EFGC has a periodic structure. All the information exchange and control parameter computations happen only once at the beginning of each control period of length T. Several steps involved in EFGC distributed control are described below:

1. At the beginning of a control period, each cell i sends the following information to its adjacent cells:

    (a) the number of active voice and data calls present in the cell denoted by $N^v_i(0)$ and $N^d_i(0)$, respectively.

    (b) the number of new voice calls, N^v_i, and new/handoff data calls, N^d_i, which were admitted in the last control period.

2. Each cell i receives $N^k_j(0)$ and N^k_j from every adjacent cell $j \in \mathcal{A}_i$.

3. Now, cell i uses the received information and those available locally to compute the acceptance ratios a_i^v and a_i^d using the technique described in section 2.5.3.

4. Finally, the computed acceptance ratios a_i^v and a_i^d are used to admit call requests into cell i using the algorithm presented in section 2.5.2.

Assume that all the cells have the same number of adjacent cells. Let $\mathcal{A}$ denote the number of adjacent cells. Also, assume that all the required information can be sent from one cell to another cell in one message. Then, the signalling overhead in terms of the number of exchanged messages in one control period is $\mathcal{A}$ messages per cell.

2.5.2 Local Admission Control Algorithm

Let (m, n) denote the state of cell i, where there are m voice calls and n data calls active in the cell. Define $\mathcal{S}_i$ as the state space of cell i governed by EFGC scheme. Then $\mathcal{S}_i$ can be expressed as

$$\mathcal{S}_i = \{(m,n) | mb_v + nb_d \le c_i\}\,. \tag{2.13}$$

Let $a_i^k(m,n)$ denote the acceptance ratio for type-k calls where the cell state is (m,n). Figure 2.6 shows the state transition diagram of the EFGC scheme in cell i for a typical state $(m,n) \in \mathcal{S}_i$. In this figure, ν_i^k is the type-k handoff arrival rate into cell i. At each state there are two acceptance ratios for voice and data calls in such a way that

$$\begin{cases} a_i^v(m,n) = 0, & \text{if } (m+1,n) \notin \mathcal{S}_i \\ a_i^d(m,n) = \frac{1}{\alpha_i}\, a_i^v(m,n), & \text{if } (m,n) \in \mathcal{S}_i \end{cases} \tag{2.14}$$

There is a service differentiation (α_i) between voice and data calls that governs the relation between these two acceptance ratios. We assume that this service differentiation is specified apriori (α_{QoS}) and EFGC should maintain it regardless of traffic conditions.

For an accurate control, the call blocking probability in each period is given by complementing the acceptance ratio. Therefore, by averaging acceptance ratios

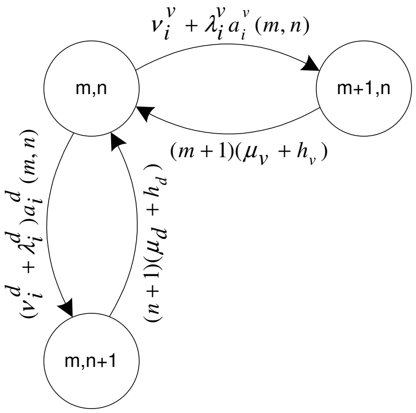

$v_i^v + \lambda_i^v a_i^v (m,n)$
m,n
m+1,n
$(m+1)(\mu_v + h_v)$
$(v_i^d + \lambda_i^d) a_i^d (m,n)$
$(n+1)(\mu_d + h_d)$
m,n+1

Figure 2.6: Extended fractional guard channel transition diagram.

over a number of control periods, the call blocking probability is expressed as $b_i^k = 1 - E[a_i^k]$. Consequently, the average network-wide call blocking probability for the considered network is given by

$$p_b^k = \frac{\sum_{j=i}^{M} \lambda_i^k b_i^k}{\sum_{j=i}^{M} \lambda_i^k} \,. \tag{2.15}$$

The pseudo-code for the local admission control in cell i is given by the algorithm of Figure 2.7. In this algorithm, x_k is a type-k call requesting b_k BUs. The corresponding type-k acceptance ratio is a_i^k. Also, $\texttt{rand}(0,1)$ is the standard random generator function. In the next section, we will present a technique to compute the acceptance ratio vector $a_i = (a_i^v, a_i^d)$ in order to complete this algorithm.

2.5.3 Computing Acceptance Ratios

It is assumed that by setting the control interval T to an appropriate value, each call experiences at most one handoff during a control period (see section 2.6.3 for more detail). Therefore, immediate neighbors of cell i, *i.e.*, $\mathcal{A}_i$, are those which will affect the number of calls and consequently the bandwidth usage in cell i during a control period.

The proposed approach for computing the acceptance ratios includes the following steps:

```
if ($x_k$ is a *voice* handoff call) then
  if ($R_i(t) + b_v \le c_i$) then
    accept call;
  else
    reject call;
  end if
else /* new voice or new/handoff data call */
  if ($(R_i(t) + b_k \le c_i)\&(\texttt{rand}(0,1) < a_i^k)$) then
    accept call;
  else
    reject call;
  end if
end if
```

Figure 2.7: Local call admission control algorithm in cell i.

1. Each cell i uses the information received from its adjacents and the information available locally to find the time-dependent mean and variance of the number of calls in the cell.

2. The computed mean and variance of the number of calls is used to find the mean and variance of the bandwidth requirement process in the cell.

3. Having the mean and variance of the bandwidth requirement process, the actual time-dependent bandwidth requirement process is approximated by a Gaussian distribution.

4. The tail of this Gaussian distribution is used to find the time-dependent handoff failure in each cell i.

5. Time-dependent handoff failure is averaged over control interval of length T to find an average handoff failure probability for the whole period.

6. Using the computed handoff failure probability and the prespecified QoS constraints, *i.e.*, p_{QoS} and α_{QoS}, acceptance ratios a_i^v and a_i^d are computed.

The number of calls in cell i at time t is affected by two factors: (1) the number of background (existing) calls which are already in cell i or its adjacent cells, and, (2) the number of new calls which will arrive in cell i and its adjacent cells during the period $(0, t]$ $(0 < t \le T)$. Let $g_i^k(t)$ and $n_i^k(t)$ denote the number of background and new type-k calls in cell i at time t, respectively. A background type-k call in cell i will remain in cell i with probability $P_s^k(t)$ or will handoff to an adjacent cell j with probability $P_{ij}^k(t)$. A new type-k call which is admitted in cell i at time $t' \in (0, t]$ will stay in cell i with probability $\tilde{P}_s^k(t)$ or will handoff to an adjacent cell j with probability $\tilde{P}_{ij}^k(t)$. Therefore, the number of background calls which remain in cell i and the number of handoff calls which come into cell i during the interval $(0, t]$ are binomially distributed. For a binomial distribution with parameter q, the variance is given by $q(1-q)$. Using this property it is obtained that

$$V_s^k(t) = P_s^k(t)\,(1 - P_s^k(t)), \tag{2.16}$$

$$V_{ji}^k(t) = P_{ji}^k(t)\,(1 - P_{ji}^k(t)), \tag{2.17}$$

$$\tilde{V}_s^k(t) = \tilde{P}_s^k(t)\,(1 - \tilde{P}_s^k(t)), \tag{2.18}$$

$$\tilde{V}_{ji}^k(t) = \tilde{P}_{ji}^k(t)\,(1 - \tilde{P}_{ji}^k(t))\,. \tag{2.19}$$

where, $V_s^k(t)$ and $V_{ji}^k(t)$ denote the time-dependent variance of stay and handoff processes, and, $\tilde{V}_s^k(t)$ and $\tilde{V}_{ji}^k(t)$ are their average counterparts, respectively.

The number of type-k calls in cell i is the summation of the number of background calls, $g_i^k(t)$, and new calls, $n_i^k(t)$, of type k. Therefore, the mean number of type-k active calls in cell i at time t is given by

$$E[N_i^k(t)] = E[g_i^k(t)] + E[n_i^k(t)], \tag{2.20}$$

where,

$$E[g_i^k(t)] = N_i^k(0)P_s^k(t) + \sum_{j\in\mathcal{A}_i} N_j^k(0)P_{ji}^k(t), \tag{2.21}$$

$$E[n_i^k(t)] = (a_i^k\lambda_i^k t)\tilde{P}_s^k(t) + \sum_{j\in\mathcal{A}_i} (a_j^k\lambda_j^k t)\tilde{P}_{ji}^k(t)\,. \tag{2.22}$$

Similarly the variance is given by

$$V[N_i^k(t)] = V[g_i^k(t)] + V[n_i^k(t)], \tag{2.23}$$

where,

$$V[g_i^k(t)] = N_i^k(0)V_s^k(t) + \sum_{j\in\mathcal{A}_i} N_j^k(0)V_{ji}^k(t), \tag{2.24}$$

$$V[n_i^k(t)] = (a_i^k\lambda_i^k t)\tilde{V}_s^k(t) + \sum_{j\in\mathcal{A}_i} (a_j^k\lambda_j^k t)\tilde{V}_{ji}^k(t)\,. \tag{2.25}$$

Note that given the arrival rate λ_i^k and the acceptance ratio a_i^k, the actual new call arrival rate into cell i is given by $\lambda_i^k a_i^k$ (see section 2.6.2). Therefore, the expected number of call arrivals during the interval $(0, t]$ is given by $a_i^k \lambda_i^k t$.

Knowing the bandwidth requirement of each type of calls, the mean and variance of bandwidth usage in cell i at time t are given by

$$E[R_i(t)] = b_v E[N_i^v(t)] + b_d E[N_i^d(t)], \tag{2.26}$$

$$V[R_i(t)] = b_v^2 V[N_i^v(t)] + b_d^2 V[N_i^d(t)]\,. \tag{2.27}$$

As we mentioned in section 2.2, the cellular system considered in this chapter is a broadband wireless system with a capacity of several Mbps. In practice, 3G systems and beyond can be considered as broadband wireless systems (for example a UMTS system can support up to 2 Mbps) [2,3]. With this range of cell capacity it is reasonable to apply the central limit theorem (this will be further discussed in section 2.7.3). Thus, the bandwidth usage in each cell can be approximated by a Gaussian distribution with mean $E[R_i(t)]$ and variance $V[R_i(t)]$. That is

$$R_i(t) \sim \mathbf{G}\big(E[R_i(t)],\ V[R_i(t)]\big)\,. \tag{2.28}$$

Therefore, the original problem of maintaining a target handoff failure probability p_{QoS} is reduced to maintaining the bandwidth usage below the available capacity c_i at any point in time $t \in (0, T]$. Approximating the handoff failure probability by the overload probability, the time-dependent handoff failure probability $P_{f_i}(t)$ can be computed as follows:

$$P_{f_i}(t) = \Pr\big(R_i(t) > c_i\big), \tag{2.29}$$

therefore,

$$P_{f_i}(t) = \frac{1}{2}\,\mathrm{erfc}\left(\frac{c_i - E[R_i(t)]}{\sqrt{2\,V[R_i(t)]}}\right), \tag{2.30}$$

where erfc(c) is the complementary error function defined as

$$\mathrm{erfc}(c) = \frac{2}{\sqrt{\pi}} \int_c^\infty e^{-t^2}\,dt\,. \tag{2.31}$$

Then the average handoff failure probability over a control period is given by

$$\tilde{P}_{f_i} = \frac{1}{T}\int_0^T P_{f_i}(t)\,dt\,. \tag{2.32}$$

Finally, to guarantee the target handoff failure p_{QoS}, we should have

$$\tilde{P}_{f_i} = p_{\mathrm{QoS}} \,. \tag{2.33}$$

To solve (2.33) for $a_i = (a_i^v, a_i^d)$ we need one more equation. This equation can be derived with respect to the required service differentiation. Given the service condition $a_d = f(a_v)$, the acceptance ratio vector $a_i = (a_i^v, a_i^d)$ can be found by numerically solving (2.33). Function f is such that $0 \leq f(a_i^v) \leq 1$ and $f(0) = 0$. In addition, f is uniformly increasing over $[0, 1]$. The boundary condition is that $a_i \in [0, 1] \times [0, 1]$, hence if $\tilde{P}_{f_i}$ is less than p_{QoS} even for $a_i^v = 1$ then a_i is set to $(1, f(1))$. Similarly, if $\tilde{P}_{f_i}$ is greater than p_{QoS} even for $a_i^v = 0$, then a_i is set to $(0, 0)$. We only consider a constant service differentiation function denoted by α_i, where $a_i^d = a_i^v/\alpha_i$.

Finally, (2.33) can be solved using the bisection method [47]. Let ξ denote the required numerical precision. Then, the computational complexity of this technique is $O(\log 1/\xi)$, given that all mathematical operations (including exponentiation and integration) can be performed in $O(1)$.

2.6 Control Parameters

In previous sections, we assumed that several parameters are known to the admission control algorithm apriori. Among these parameters are the call arrival rates, mean call durations, mean cell residency times and routing probabilities. In practice, all these parameters can be extracted from measured field data using an estimation technique. Measurement and estimation units are used for providing the required parameters to the admission control unit as shown in Figure 2.8. One useful estimation technique is presented in the following subsection.

2.6.1 Parameter Estimation

A common technique for estimating the mean values from measurement data is the *exponentially weighted moving average* (EWMA) technique. Let z denote a control parameter to be estimated, *e.g.*, arrival rate, and $\hat{z}$ its measured (observed) value.

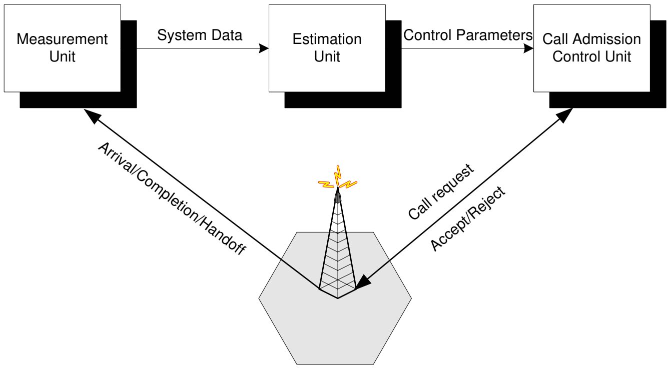


Figure 2.8: Control unit diagram.

A moving average estimator for z at nth step is given by

$$z(n) = (1 - \epsilon_1)\,\hat{z}(n-1) + \epsilon_1 z(n-1),$$

where ϵ_1 is a weighting factor that should be specified with respect to the sampled observations of z. In general, a small value of ϵ_1 can keep track of the changes more accurately, but is too sensitive to temporary fluctuations. On the other hand, a large value of ϵ_1 is more stable but could be too slow in adapting to real traffic changes. By using this estimator, it can be verified that $E[z] = E[\hat{z}]$. However, EFGC is independent of the estimation technique, and hence, it is possible to use more sophisticated estimation techniques to achieve more accurate estimations (refer to [48, 49]).

We use the EWMA technique to compute the new call arrival rate λ into a cell of the network. The only unknown parameter is the estimation coefficient ϵ_1. As mentioned before, the accuracy of the EWMA estimation depends on ϵ_1. The goal is to choose ϵ_1 in such a way to minimize the estimation error. To measure the estimation error, we use the mean squared error (MSE) of the estimations as expressed by

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}\left(\lambda(i) - \hat{\lambda}(i)\right)^2 \tag{2.34}$$

where N is the number of measurements. In our experiments we found that $\epsilon_1 = 0.8$ minimizes the prediction error and achieves sufficiently accurate predictions.

2.6.2 Actual New Call Arrival Rate

In section 2.5.3, we used products $a_j^k \lambda_j^k$ to compute the mean and variance of the number of calls in cell i ($j \in \mathcal{A}_i$). Let us define the *actual new call arrival rate* into cell j, denoted by $\bar{\lambda}_j^k$, as follows

$$\bar{\lambda}_j^k = a_j^k \lambda_j^k \,. \tag{2.35}$$

In order to compute a_i^k for the new control period we need to know $\bar{\lambda}_j^k$ for every adjacent cell j ($j \in \mathcal{A}_i$). Similarly, cell j needs to know $\bar{\lambda}_i^k$ in order to be able to compute a_j^k. Therefore, every cell depends on its adjacents and vice versa. To break this dependency, instead of using the actual value of $\bar{\lambda}_j^k$, each cell i estimates the actual new call arrival rates of its adjacents for the new control period.

Let $\bar{\lambda}_j^k(n)$ denote the actual new call arrival rate into cell j during the nth control period. Also, let $N_j^k(n)$ denote the number of new calls that were accepted in cell j during the nth control period. Similar to [11, 17], an estimator for $\bar{\lambda}_j^k$ is expressed as

$$\bar{\lambda}_j^k(n+1) = (1-\epsilon_2)\frac{N_j^k(n)}{T} + \epsilon_2 \bar{\lambda}_j^k(n), \tag{2.36}$$

where, $\bar{\lambda}_j^k(n+1)$ is the actual new call arrival rate into cell j at the beginning of the $(n+1)$th control period. Note that $\bar{\lambda}_j^k(n)$ is known at the beginning of the $(n+1)$th control period. In our simulations we found that $\epsilon_2 = 0.3$ leads to a good estimation of the actual new call arrival rate.

2.6.3 Control Interval

The idea behind at-most-one handoff assumption is that by setting control interval appropriately, the undesired multiple handoffs during a control period can be avoided. As discussed in section 2.5, this minimizes the signalling overhead and operational complexity of EFGC. In this section, we address the control interval selection problem.

Consider a symmetric network where each cell has exactly $\mathcal{A}$ neighbors, and the probability of handoff to every neighbor is the same. Then, the routing probability

r_{ij} from cell i to cell j is given by

$$r_{ij} = \begin{cases} 1/\mathcal{A}, & j \in \mathcal{A}_i, \\ 0, & j \notin \mathcal{A}_i\,. \end{cases} \tag{2.37}$$

Let $q(n)$ denote the probability that an active call experiences n handoffs during time interval T. Also, let $q_{ij}(n)$ denote the probability that a call originally in cell i moves to cell j over a path consisting of n handoffs during time interval T. Define δ as the multiple handoffs probability from cell i to cell j. We then can write

$$\delta = \sum_{n=2}^{\infty} q_{ij}(n)\,. \tag{2.38}$$

Our goal is to find a relation between T and δ in order to be able to control δ by controlling T.

For an effective control (p_f in the range of 10^{-4} to 10^{-2}) we can assume that p_f is effectively zero. Similarly, if $\delta \approx p_f$ for a given T, we can assume that the multiple handoffs probability is zero. Since cell residency is exponential, the number of handoffs a call experiences during an interval is Poisson distributed with mean hT, given that the call is active during the whole interval. Therefore, it is obtained that

$$q(n) = \frac{(hT)^n}{n!} e^{-(h+\mu)T}\,. \tag{2.39}$$

In order to compute $q_{ij}(n)$ based on (2.39), we need to find the probability of moving from cell i to cell j by n handoffs. Let $L_{ij}(n)$ denote the number of paths consisting of n handoffs from i to j, then

$$q_{ij}(n) = \frac{L_{ij}(n)}{\mathcal{A}^n}\, q(n)\,. \tag{2.40}$$

Consider the network depicted in Figure 2.5. Let $T = 20\,s$, $1/\mu = 180\,s$, $1/h = 100\,s$ and $\mathcal{A} = 6$. Table (2.6.3) shows the maximum probability of multiple handoffs from any cell j to cell 0, $P_{j0}(n)$, based on the number of handoffs, n. For each n, we have also determined which layer has the maximum paths to cell 0. Interestingly, cell 0 has the most paths to itself through other cells. We have also illustrated in Figure 2.9 the impact of the control interval T on the multiple handoffs probability δ for the same set of parameters.

n	Layer	$\max\{L_{j0}(n)\}$	$\max\{P_{j0}(n)\}$
0	0	1	0.73263
1	0	1	0.02442
2	0	6	0.00244
3	1	15	0.00007
4	0	90	0.00000
5	0	360	0.00000

Table 2.3: Multiple handoffs probability for $T = 20\,s$.

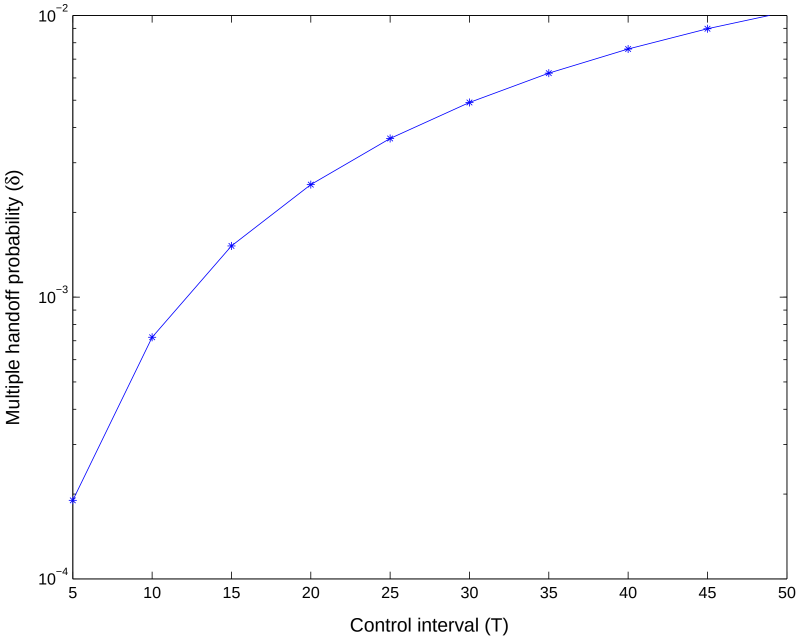


Figure 2.9: Effect of T on multiple handoffs probability.

Consider cell i and all the cells around it forming circular layers. From cell i, all the cells up to layer n are accessible with n handoffs assuming that cell i forms layer 0. It can be shown that

$$L_{ij}(n) \leq \mathcal{A}^{n-1}, \quad n \geq 1 \tag{2.41}$$

because for $n \geq 1$, at each level there are at least $\mathcal{A}$ cells which have the same number of paths to the destination cell i. Therefore

$$q_{ij}(n) \leq \frac{1}{\mathcal{A}} \frac{(hT)^n}{n!} e^{-(h+\mu)T}, \quad n \geq 1\,. \tag{2.42}$$

Using (2.38) and (2.42), it is obtained that

$$\begin{aligned} \delta &\leq \sum_{n=2}^{\infty} \frac{1}{\mathcal{A}} \frac{(hT)^n}{n!} e^{-(h+\mu)T} \\ &= \frac{e^{hT} - hT - 1}{\mathcal{A} e^{(h+\mu)T}}\,. \end{aligned} \tag{2.43}$$

Using the Taylor expansion of exponential terms for $\delta \ll \frac{1}{\mathcal{A}}(\frac{h}{\mu+h})$, it is obtained that

$$T \leq \frac{\mathcal{A}\delta(\mu+h) + h\sqrt{2\mathcal{A}\delta}}{\mathcal{A}\delta(\mu+h)^2 - h^2}, \tag{2.44}$$

which finally leads to the following simple relation

$$T \approx \frac{\sqrt{2\mathcal{A}\delta}}{h}\,. \tag{2.45}$$

2.7 Simulation Results

2.7.1 Greedy EFGC

The basic EFGC introduced in section 2.5 may seem to be too conservative about accepting data calls. We refer to this restrictive version of EFGC by EFGC-REST (or simply REST). REST is a conservative approach which aims at satisfying the specified priority function f over time. In other words, REST always uses the acceptance ratio $a_i = (a_i^v, f(a_i^v))$ regardless of the congestion situation to impose an exact priority function.

It is observed that in some states of the system it is possible to increase the acceptance ratio of data calls beyond the limit returned by the service differentiation function. For example when the network is not congested (at light traffic loads), we found that by increasing the priority of data traffic the overall utilization of the wireless bandwidth is increased while the handoff failure remains almost untouched. This relaxed version is called EFGC-UTIL (or simply UTIL) due to its greedy behavior in maximizing the bandwidth utilization. To find the data acceptance ratio in cell i, UTIL follows the following steps:

1. Find a_i^v using (2.33),
2. If $(a_i^v == 1)$ then find the maximum value of $a_i^d \in [f(1), 1]$ which satisfies (2.33),

It is worth noting that the computational complexity of EFGC-UTIL is the same as EFGC-REST, *i.e.*, $O(\log 1/\xi)$.

2.7.2 Simulation Parameters

Simulations were performed on a two-dimensional cellular system consisting of 19 hexagonal cells (see Figure 2.5). Opposite sides wrap-around to eliminate the finite size effect. It is assumed that mobile users move along the cell areas according to a uniform routing pattern. In other words, all neighboring cells have the same chance to be chosen by a call for handoff, *i.e.*, $r_{ji} = 1/6$. For ease of illustrating the results, the simulated system is uniform, *i.e.*, input load is the same for every cell, although EFGC as well as the simulation program are designed to handle the nonuniform case as well. Therefore, unless explicitly specified, the subscript i is omitted hereafter.

The common parameters used in the simulation are as follows. All the cells have the same capacity $c = 5$ Mbps, which is equal to 160 BU assuming each BU is equal to 32 Kbps (encoded voice using ADPCM requires 32 Kbps). Target handoff failure probability for voice calls is $p_{\text{QoS}} = 0.01$[1] and $T = 20\,s$. We use normalized load in simulations which is simply the total arrival load per BU. Let ρ denote the

[1] It has been show that 1% call dropping is acceptable in practical systems.

Type	Priority	$1/\mu$ (s)	$1/h$ (s)	BU	Load
voice	1	180	100	1	60%
data	0.5	1000	800	2	40%

Table 2.4: Voice/Data service parameters.

total normalized arrival load into a cell, then

$$\rho = \frac{1}{c}\big(\rho_v + \rho_d\big), \tag{2.46}$$

where, ρ_v and ρ_d are, respectively, voice and data load given by

$$\rho_v = b_v\lambda_v/\mu_v, \tag{2.47}$$

$$\rho_d = b_d\lambda_d/\mu_d\,. \tag{2.48}$$

For each load, simulations were done by averaging over 8 samples, each for 10 hours of simulation time. Load distribution between voice and data traffic is fixed over time. At any load, 60% of the load is due to voice calls and the remaining 40% is composed of data calls. Table 2.4 summarizes service and traffic parameters for both traffic types. In this table, *priority* refers to the relative priority (service differentiation) of voice and data calls. It means that new voice calls have higher priority than data calls for the admission control algorithm. In particular, the probability of accepting a new voice call is at least twice the probability of accepting a data call (new/handoff) at any time and any load. Equivalently, this is achieved by setting $\alpha_{\text{QoS}} = 2$. As mentioned earlier, this relative priority can be any service differentiation function. In our simulations, for the sake of simplicity we have chosen a constant service differentiation function.

We have also implemented the DTR scheme introduced in section 2.2 for comparison purposes. Since DTR is designed for a static traffic pattern, the handoff failure probability increases rapidly with the network load when the guard channels for handoff are few, but remains too low when the guard channels are many. Here, we choose the two thresholds in such a way that DTR achieves its objectives when the network starts to get overloaded. Hence, the voice threshold is set to 155 BUs and the data threshold is set to 151 BUs. Using these thresholds at load 2, p_f and $\alpha = a_v/a_d$ were found to be 0.01 and 2, respectively.

2.7.3 Gaussian Approximation

When the network is not congested and each cell has only a few active calls, it is clear that Gaussian approximation is not good. However, at light loads the admission algorithm does not require a high precision estimation of the load since there is no congestion in the network. As the load increases the number of active calls in each cell increases rapidly until no more calls can be accepted. Due to the high capacity of broadband systems, it is expected to have enough active calls in each cell so that central limit theorem can be applied.

Other researchers have also successfully applied Gaussian approximation for similar purposes. Naghshineh and Schwartz [11] and Epstein and Schwartz [14] used the same kind of approximation. The main difference is that we extend their single point approximation at the end of the control period to a time dependent approximation over the whole control period. The authors of [17] also realized that for large system sizes, as is the case in this chapter, the cell occupancy distribution evolves into a Gaussian distribution. Investigating the tail behavior of the bandwidth usage distribution is beyond the scope of this study, instead we rely on the results from other researchers [11, 14, 17, 42].

2.7.4 Results and Analysis

Effect of arrival load

The first set of simulation results show the main performance parameters of EFGC. Figure 2.10 shows the handoff failure probability for the three schemes for a wide range of loads. Both UTIL and REST maintain a constant failure probability independent of the load. For DTR, it grows very rapidly with the load (which was expected). With light loads (load < 2), DTR and REST have almost the same handoff failure probability while UTIL has slightly higher handoff failure probability. But with high loads (load > 2), UTIL and REST converge to exactly the same handoff failure probability while DTR has much higher handoff failure probability. Figure 2.12(b) shows that, although REST has better failure probability in light loads, this is accomplished at the expense of the data call blocking probability. However, even in this region (load < 2), UTIL satisfies the target handoff failure

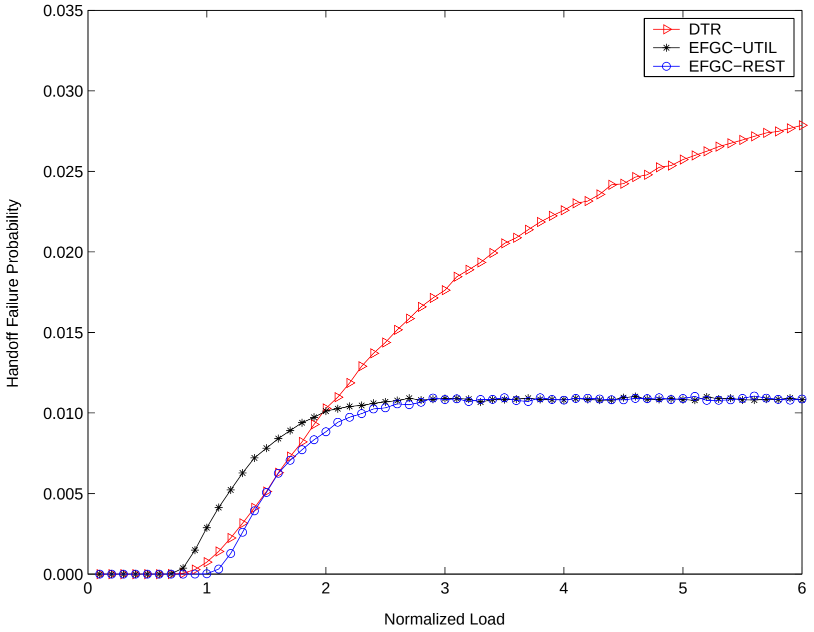


Figure 2.10: Voice handoff failure probability.

probability p_{QoS}.

One of the objectives of EFGC is to maintain the relative service priority between voice and data calls. In our simulations, this relative priority is fixed and indicates that the acceptance probability of new voice calls should be twice the acceptance probability of new data calls. Figure 2.11 gives the service differentiation $\alpha = a_v/a_d$ for different loads. It shows that EFGC maintains an almost constant service priority between the two types of traffic. More precisely, REST preserves $\alpha = 2$ for the whole range of loads while UTIL has $\alpha = 1$ in light loads and $\alpha = 2$ in high loads as expected. This can be explained by the fact that in light loads UTIL accepts data calls as long as there is free bandwidth (without violating the target voice handoff failure probability). As the load increases, service priority of DTR increases rapidly. Figure 2.12(b) shows that at high loads almost no data calls are accepted. In other words, DTR is not fair and leads to starvation of data traffic. It is worth mentioning that, although in this simulation the service differentiation is fixed, the EFGC can satisfy more complex priority disciplines such as state dependent priorities.

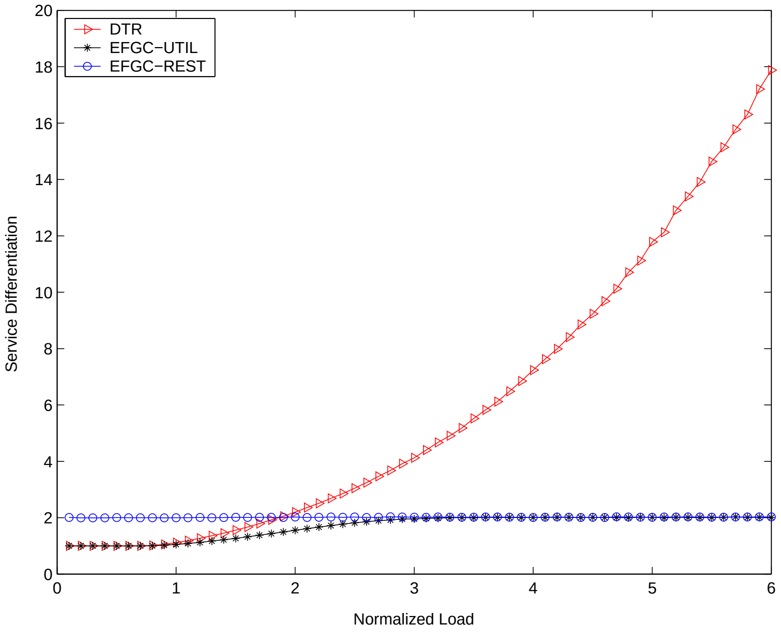
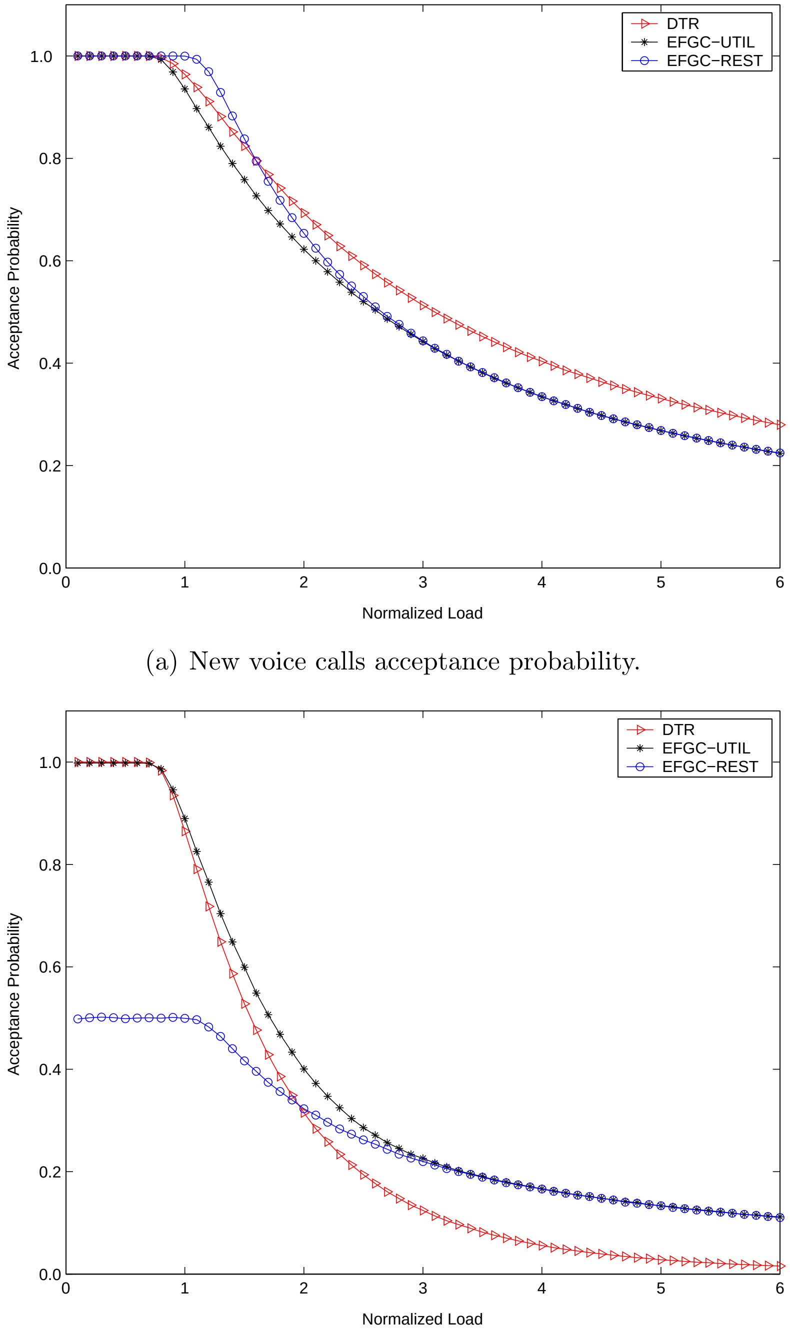


Figure 2.11: Voice/Data relative acceptance probability (α).

Figure 2.12 shows the new voice and new/handoff data call acceptance probabilities respectively. Again for high loads, UTIL and REST converge to the same result but the difference in their performance at light loads is significant. For data traffic at light loads the acceptance probability of UTIL is almost twice that of REST. This explains why the utilization of UTIL is superior to REST. It can be seen that DTR has slightly higher acceptance probability for voice but much lower acceptance probability for data in comparison to UTIL and REST.

Finally, Figure 2.13 shows the wireless bandwidth utilization under the three bandwidth allocation mechanisms. Although DTR performs poorly in terms of handoff failure probability and service priority, its utilization is slightly better than EFGC. Interestingly, UTIL has exactly the same utilization level as DTR at light loads but higher than that of REST. In this simulation, voice traffic constitutes the larger portion of the total load. As the percentage of data traffic increases, the utilization of DTR is expected to drop. This will be investigated next.



(a) New voice calls acceptance probability.

(b) New/handoff data calls acceptance probability.

Figure 2.12: Acceptance probability of voice and data.

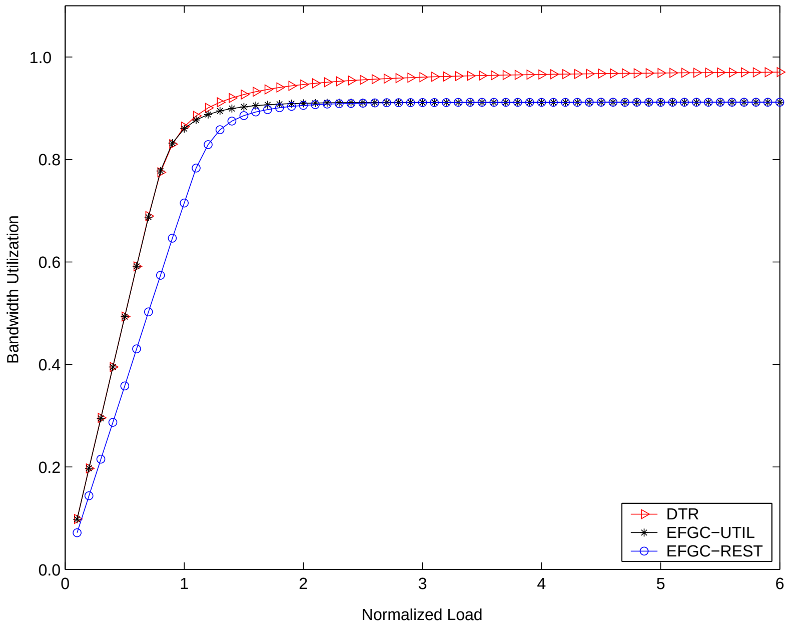


Figure 2.13: Wireless bandwidth utilization.

Effect of load sharing

In previous simulations, the load sharing factor $\beta (\beta > 0)$ is set to 1.5, where

$$\beta = \frac{\text{arriving data traffic load } (\rho_v)}{\text{arriving voice traffic load } (\rho_d)}. \tag{2.49}$$

Due to the priority of voice calls over data calls, varying β will affect the behavior of EFGC. As shown in Figures 2.14-2.16, EFGC is insensitive to the load sharing factor. In these plots, the X axis indicates the load sharing factor β. It is assumed that most of the traffic is composed of voice calls, hence β varies between 0.5 and 5.

For this set of simulations, normalized arrival load is set to 1.5 Erlang and voice priority is set to 2 ($\alpha = 2$). As expected, DTR is not able to adjust to changes in load shares although the total load is fixed. Interestingly as β increases, EFGC-UTIL and EFGC-REST converge to the same value for handoff failure probability. The reason is that by increasing β, voice traffic will dominate data traffic. Therefore, a larger portion of the available bandwidth is allocated to voice traffic in such a

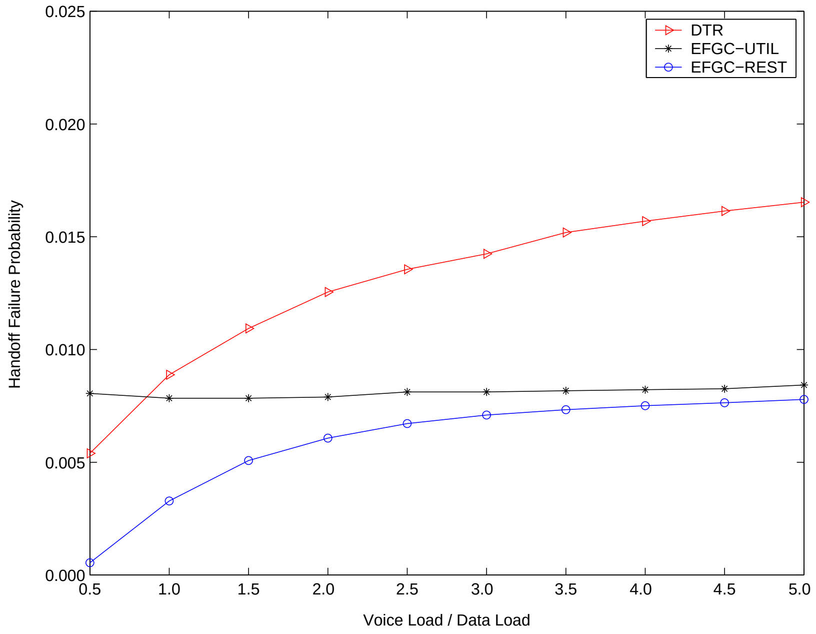


Figure 2.14: Effect of load sharing (β) on handoff failure probability.

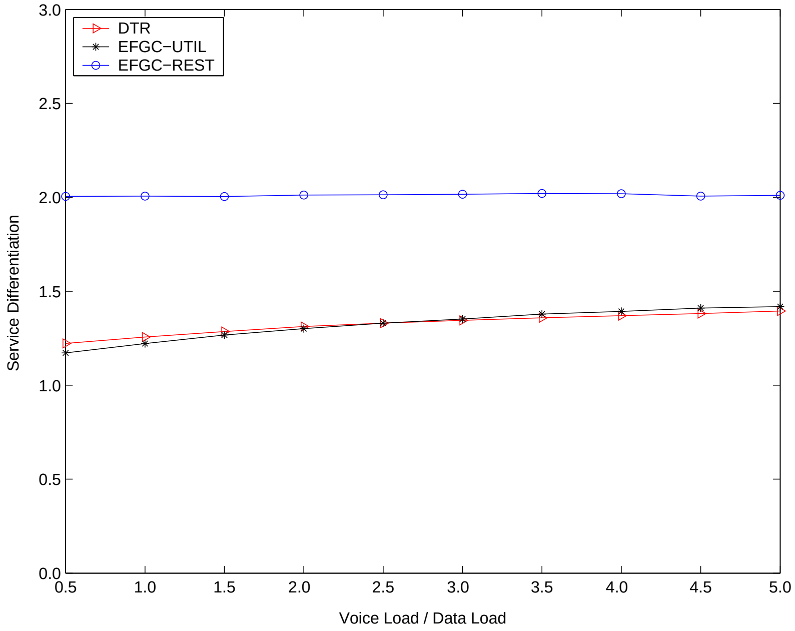


Figure 2.15: Effect of load sharing (β) on relative acceptance probability (α).

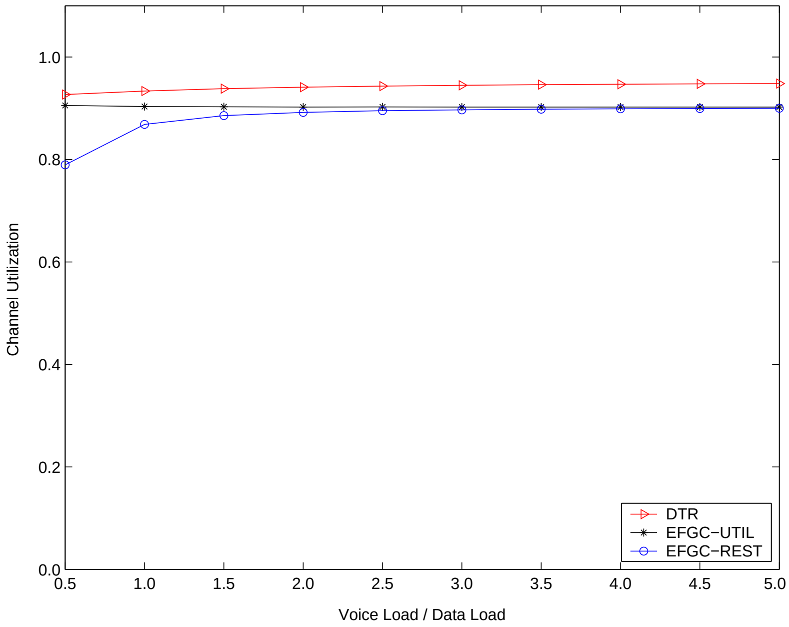


Figure 2.16: Effect of load sharing (β) on bandwidth utilization.

way that there is no extra free bandwidth to be assigned to data traffic (more than their guaranteed share).

The primary goal of the following set of simulations is to show the stability of EFGC under various QoS requirements (p_{QoS} and α_{QoS}) and the insensitivity of EFGC to the exponential assumption we made about the cell residency time.

Effect of voice priority

Figures 2.17-2.19 show the effect of changing the relative priority of data calls and voice calls. In this set of plots, the X axis indicates the quantity $1/\alpha$, where

$$1/\alpha = \frac{\text{data calls acceptance probability } (a_d)}{\text{voice calls acceptance probability } (a_v)}. \tag{2.50}$$

In the simulations, the total arrival load is set to 1.5 Erlang which consists of 60% voice traffic and 40% data traffic (*i.e.*, a load sharing factor of 1.5). It is found that regardless of α, EFGC is able to satisfy the target α_{QoS} while providing the desired service differentiation. The straight lines in Figure 2.18 indicate that any value of service differentiation can be strictly guaranteed with EFGC.

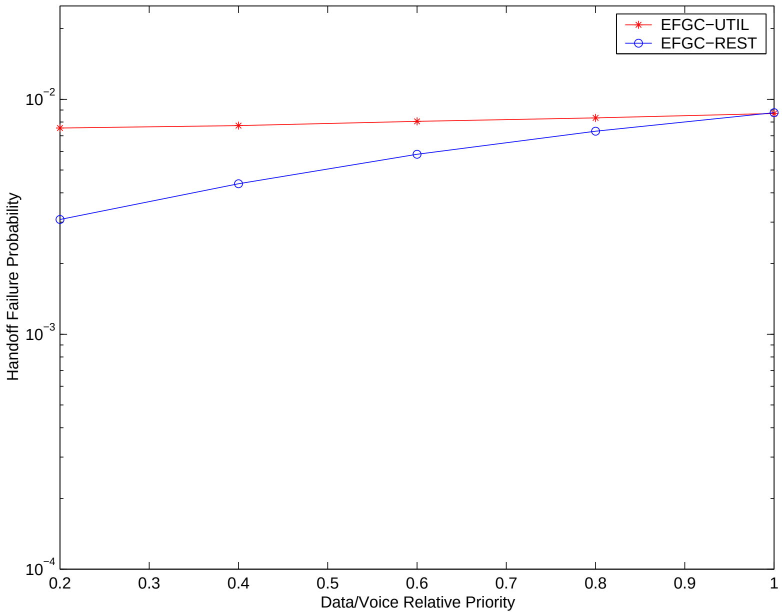


Figure 2.17: Effect of voice priority on handoff failure probability.

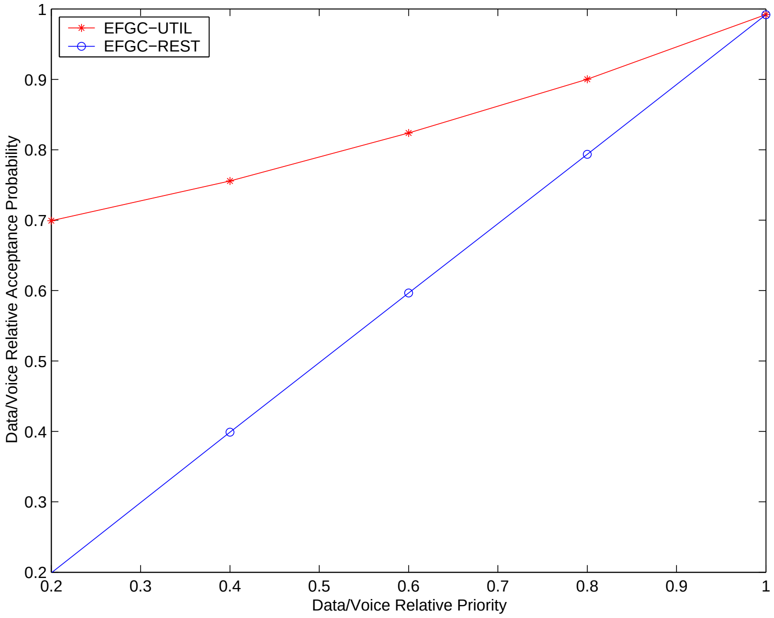


Figure 2.18: Effect of voice priority on relative acceptance probability ($1/\alpha$).

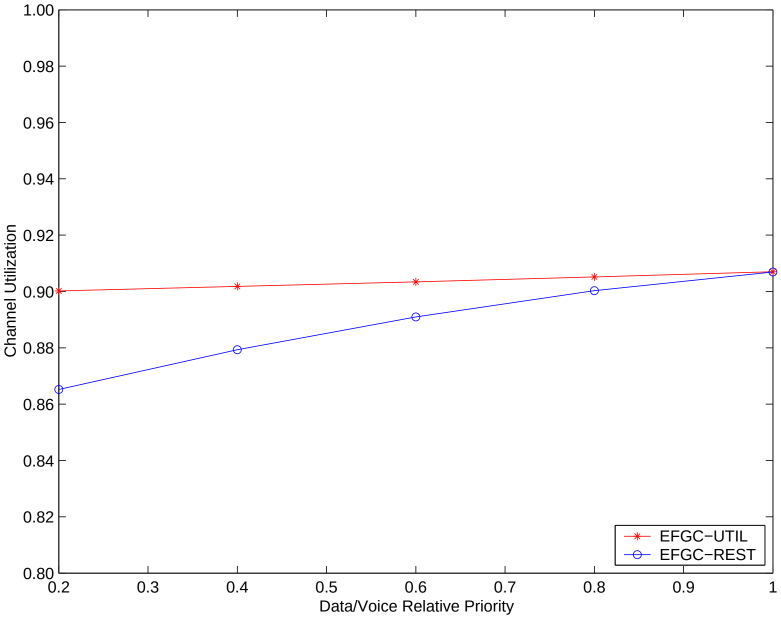


Figure 2.19: Effect of voice priority on bandwidth utilization.

As indicated in these figures, UTIL and REST converge to the same control policy as α tends towards 1. This was expected because the two schemes differ from each other with respect to α. In this case, available resources are completely shared among voice and data traffic and channel utilization is maximized. However, for large values of α (small values of $1/\alpha$), UTIL has a superior performance over REST. For example, at $\alpha = 1/0.2$, UTIL has 4% better utilization.

Effect of handoff failure probability (QoS)

In cellular systems, the target p_{QoS} is typically set to 1%. To show the adaptiveness of EFGC, simulations were performed for $p_{\text{QoS}} = [0.2\%, 1\%, 5\%]$. Notice that $p_{\text{QoS}} = 0.2\%$ is an extremely low handoff failure probability. As shown in Figures 2.20-2.22, handoff failure and service differentiation are fully satisfied regardless of the target QoS requirements. In particular, Figure 2.20 shows the stability of EFGC under different target dropping requirements.

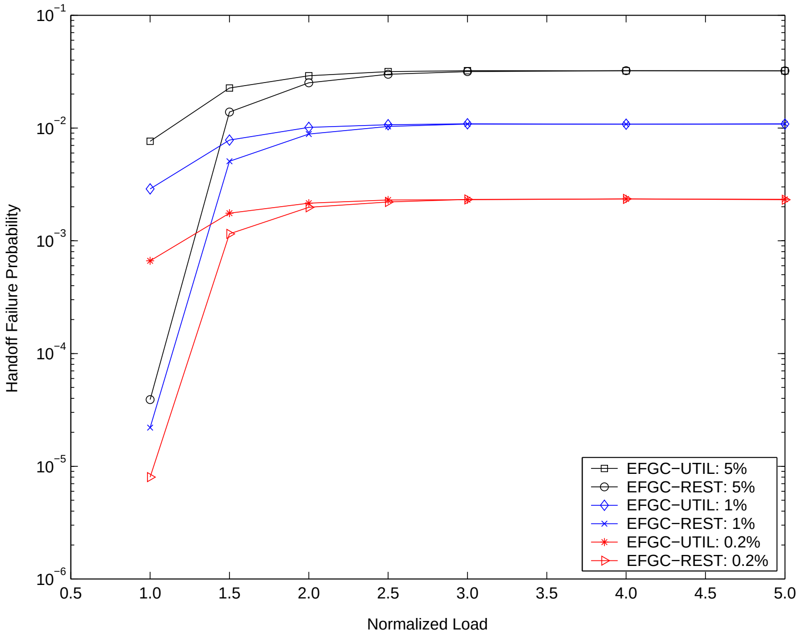


Figure 2.20: Effect of QoS on handoff failure probability.

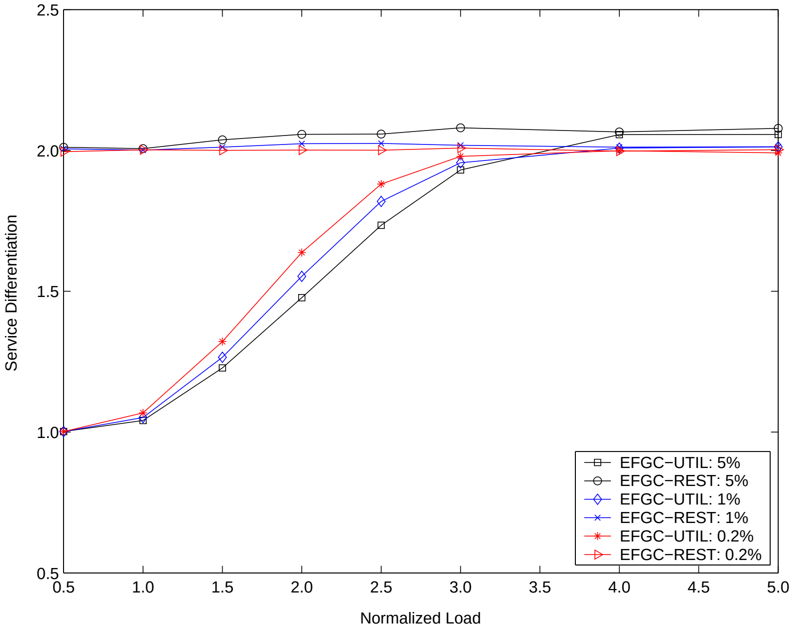


Figure 2.21: Effect of QoS on relative acceptance probability (α).

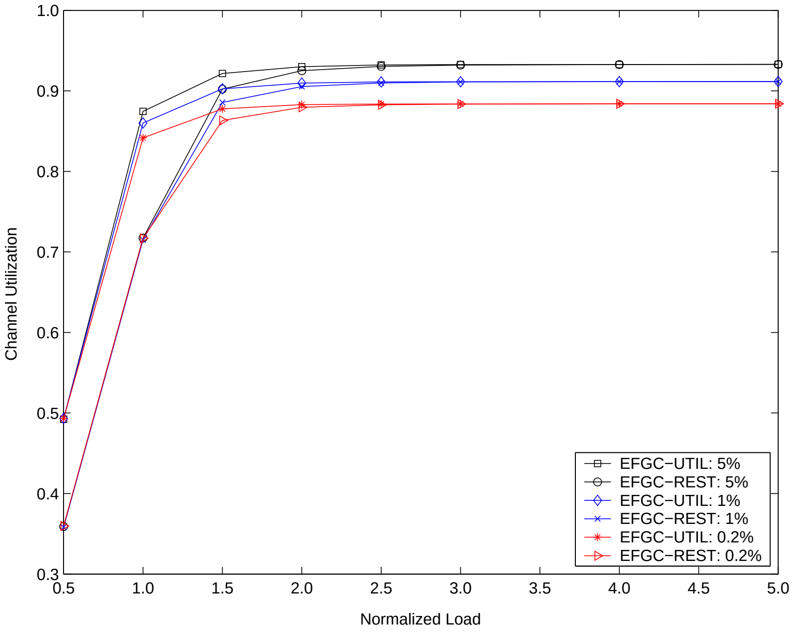


Figure 2.22: Effect of QoS on bandwidth utilization.

Effect of non-exponential cell residency

The first part of our analysis, which gives the equations describing the mean and variance of channel occupancy (*i.e.*, number of busy channels in a cell), is based on the exponential cell residency time assumption. This assumption may not be correct in practice and needs more careful investigation as pointed out in [50–52] and references there in. Although exponential distributions are not accurate in practice but the models based on the exponential assumption are tractable and do provide mean value analysis which indicates the system performance trend.

Using real measurements, Jedrzycki and Leung [50] showed that a lognormal distribution is a more accurate model for cell residency time. We now compare the results obtained under exponential distribution with those obtained under more realistic lognormal distribution. The mean and variance of both distributions are the same (refer to Table 2.4). Figures 2.23-2.25 shows that the exponential cell residency achieves sufficiently accurate control. In other words, the control algorithm is rather insensitive to this assumption due to its periodic control in which the length of the control interval is much less than the average cell residency time.

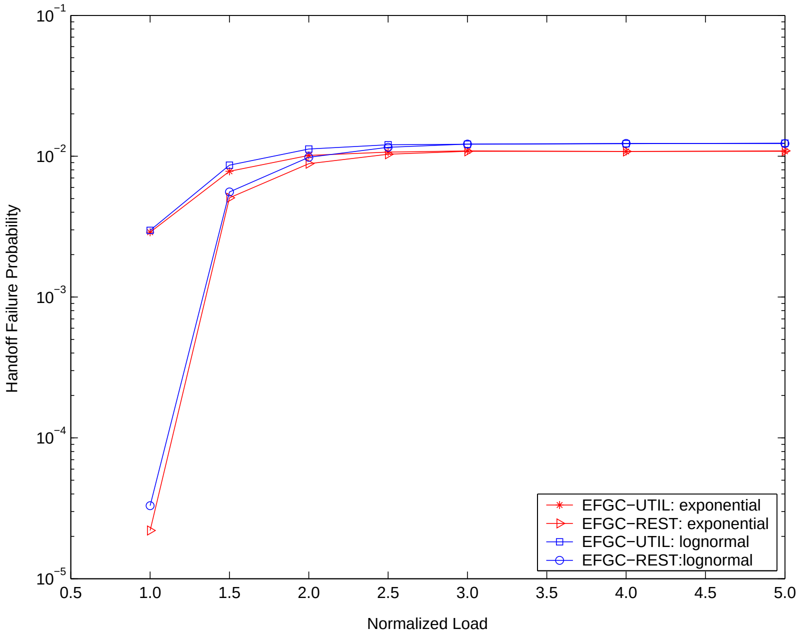


Figure 2.23: Effect of non-exponential cell residency on handoff failure probability.

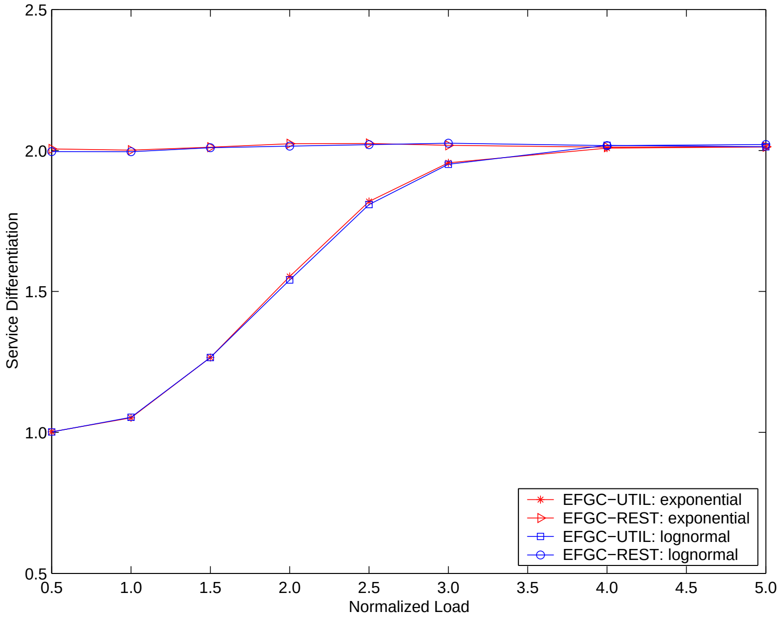


Figure 2.24: Effect of non-exponential cell residency on relative acceptance probability (α).

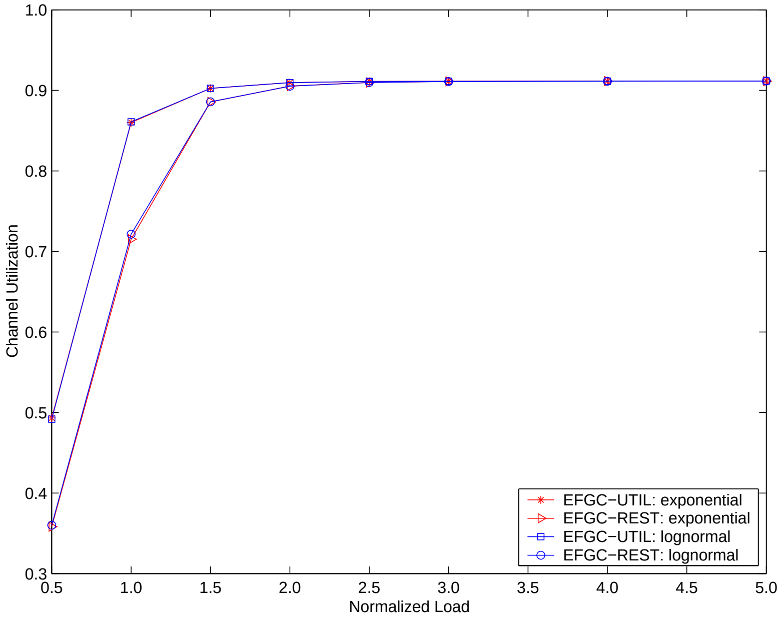


Figure 2.25: Effect of non-exponential cell residency on bandwidth utilization.

2.8 Conclusion

In this chapter, we proposed a new admission control algorithm for voice/data integration in broadband wireless networks. Our algorithm is a natural extension of the well-known fractional guard channel proposed for voice cellular systems. EFGC always achieves the predetermined call dropping probability for voice calls while keeping the relative blocking probability of voice and data calls within a target threshold. We then described two versions of the EFGC, namely EFGC-UTIL and EFGC-REST. EFGC-UTIL follows a greedy approach to maximize the bandwidth utilization while EFGC-REST maintains the relative service priority. Both versions converged to the same result for high traffic loads. The major advantage of EFGC is its insensitivity to network traffic load. The dropping probability of voice calls and relative blocking probability of voice and data calls is maintained at a stable level over a wide range of traffic loads. From the simulation results, we conclude that EFGC-UTIL is a better candidate for integrated voice/data cellular networks.

In this chapter, we focused on the impact of mobility on network performance. In next chapter, we study the impact of wireless channel on TCP throughput. The results can be used to further extend the admission control algorithms proposed in this chapter by providing more accurate estimation for bandwidth requirements of data calls.

Chapter 3

Transport Perspective: TCP Optimization

TCP is the dominant transport protocol over both wired and wireless links. It is however, well known that TCP is not suitable for wireless networks and several solutions have been proposed to rectify this shortcoming. In this chapter we explore cross-layer optimization of the *rate adaptation* feature of cellular networks to optimize TCP throughput. Modern cellular networks incorporate RF technology that allows them to dynamically vary the wireless channel rate in response to user demand and channel conditions. However, the set of data rates as well as the scheduler policy are typically chosen to optimize throughput for *inelastic* applications.

In order to optimize such a system for TCP, we propose a *two state* TCP-aware scheduler that switches between two rates as a function of the TCP sending rate. We develop a fluid model of the steady-state TCP behavior for such a system and derive analytical expressions for TCP throughput that explicitly account for rate variability as well as the dependency between the scheduler and TCP. Using the model we choose RF layer parameters that, in conjunction with the TCP-aware scheduler, improve long-term TCP throughput by the order of $15\% - 25\%$. We also compare our analytical results against those obtained from *ns-2* simulations and confirm that our model indeed closely approximates TCP behavior in such an environment.

3.1 Chapter Organization

The remainder of the chapter is structured as follows. Section 3.3 discusses the related work in more depth and contrasts it with our current work. Section 3.4 presents our system model of a TCP-aware RF channel scheduler. Section 3.5 presents our TCP model that captures the correlation between the TCP session and the scheduler. We derive analytical expressions for TCP throughput that explicitly accounts for the presence of two channel states. Section 3.7 validates the accuracy of the model against *ns-2* simulations. We also demonstrate the utility of this model through numerical optimization of the channel rates to maximize TCP throughput. Our conclusions and future work is presented in Section 3.8.

3.2 Introduction

Modern digital communication technologies combined with powerful mobile processors now allow wireless channel schedulers in cellular networks to rapidly change the allocated channel resources in response to channel conditions as well as user demands. This is achieved by controlling various parameters (and combinations thereof) such as the coding rate, spreading factor, modulation scheme[1] and link layer re-transmission rate. For ease of exposition, we shall refer to these parameters as *RF control variables.* These variables essentially trade-off data rates for improved *frame error rates* (FER) and vice versa.

Cellular networks typically specify various combinations of the RF control variables that result in a set of *allowed* data rates and corresponding FER. The RF scheduler dynamically assigns rates from this allowed set based on its rate adaptation policy. For example, in the CDMA2000 1xRTT network [53], the scheduler can dynamically transit between four different data rates during a mobile's session in response to buffer content and channel conditions by varying the spreading factor through the *Walsh code*[2] length. A shorter *Walsh* code lowers the spreading factor,

[1] Signal power is typically used to offset noise and interference rather that directly increase data rate.

[2] In CDMA systems, the Walsh code is an orthogonal code used to identify each user and mitigate other-user interference.

which results in higher data rates but at the cost of lower SINR (Signal to Interference and Noise Ratio) and hence higher frame error rates. Channel schedulers in modern third generation cellular networks such as W-CDMA [54] and 1xEV-DO [55], can allocate from ten different transmission rates in each time slot to each user. For such networks, all three control knobs, coding rate, spreading rate and modulation are used to achieve higher data rates, either at the cost of higher frame error rates, or when channel conditions permit.

In practice, the above mentioned factors that decide the set of allowed data rates as well as the scheduler's rate adaptation policy are chosen to optimize the *raw* physical layer *goodput* of a user. The set of data rates is obtained by choosing a combination of the RF control variables for each channel condition[3] that produces the highest channel data rate (under that particular channel condition) for a target Frame Error Rate (FER). Similarly, scheduling policies typically involve the assignment of the highest possible data rate allowed (for a given channel condition) from the set of given data rates that can clear the buffer backlog[4]. While ideal for *inelastic* constant rate applications, this methodology for resource allocation can produce sub-optimal performance of *elastic* applications and protocols, in particular TCP, that adapt their rate in response to feedback from the receiver.

As is well known, TCP, by far the most dominant transport protocol, uses the additive increase multiplicative decrease (AIMD) algorithm that gradually increases its transmission rate based on receiver feedback and rapidly throttles back when it perceives losses (either due to congestion or channel errors). Given this complex relation between TCP throughput and the channel transmission and loss rate, the same trade-off in channel capacity and frame errors that works for inelastic traffic may yield data rates and FERs that degrade TCP throughput. Similarly, a scheduler's rate adaptation policy that always aims at clearing buffer backlog can be sub-optimal for TCP. For example, when the TCP source has a small window and is ramping up its rate, it is very sensitive to losses but not to the assigned channel rate. In such a state if the RF scheduler allocates a high channel rate at the expense of a higher FER (perhaps due to a sudden accumulation of buffer

[3] A channel condition is defined by a particular range of SINR values and is a function of fading, interference *etc.*

[4] The CDMA2000 1xRTT is an example of a commercial system with such features.

backlog as a result of jitter in the network), the TCP source cannot fully utilize the high rate and in fact may drop its window or time-out due to the channel errors. Conversely, for larger windows (higher TCP sending rates), it may be advisable to allocate higher channel rates even at the expense of higher bit error rates, since low channel rates will inevitably result in packet losses due to congestion. A detailed study on the interaction of the CDMA2000 1xRTT scheduler and TCP based on extensive measurements was conducted in [56]. One of the main conclusions of the work was that variable rate of the wireless channel is the dominant factor in determining TCP behavior. On a related note, a previous study found that sharp bandwidth oscillations induced by rate adaptation of the RF scheduler in CDMA networks that are agnostic to TCP result in throughput degradation [57].

The above discussion clearly motivates the case for a “TCP-aware” scheduler as a means to improve TCP throughput on a wireless channel[5]. In order to achieve this objective, such a scheduler should be able to: (a) choose control variables such as coding rate that yield the optimal set of data rates (and corresponding frame error rates) from the perspective of TCP throughput, and (b) adapt its rate in a manner that is cognizant of TCP dynamics. The proposal of a simple scheduler that captures the above mentioned properties and analysis of its performance forms the main objective of this chapter. Specifically, we propose a two-rate wireless channel scheduler that changes its rate in response to the TCP sending rate and build an analytical model to compute the bulk TCP throughput of a session in such an environment. We then show how the model can be used to optimize control variables like coding rate to obtain the set of data rates that maximizes TCP throughput.

Several previous studies have explored the subject of cross-layer optimization of error coding rate, signal power, *etc.* to maximize TCP throughput. However, none of them have considered TCP-aware rate adaptation. References [58–61] have investigated the impact of RF control variables on TCP throughput, but only in *static* scenarios that involve choosing a *single* instance of the RF control variables. The authors of [59,62] consider the impact of dynamic rates on TCP via simulations. However, in all these works, the codes are assumed to change *only* in response to channel conditions and not user demand. The work closest in nature to ours is [63]

[5] Such an approach falls in the realm of RF cross-layer optimization.

that explicitly accounts for TCP state in the cross-layer optimization. Their focus, however, is on *signal power* adaptation that does not incorporate rate changes and requires significant channel information and computational complexity. We dwell in more detail on related work in Section 3.3.

Our contribution can be summarized as follows:

1. We propose a simple two-state wireless channel scheduler that changes its state in response to TCP sending rate. Each scheduler state results in a different transmission rate, round trip time and FER. This system is used to study the benefits of cross-layer optimization of the dynamic rate adaptation feature of modern cellular networks with respect to TCP throughput.

2. We develop analytical expressions for the steady-state throughput of a long-lived TCP session in such an environment. Our model explicitly captures the dependency of the scheduler on TCP sending rate as well the impact of the presence of two distinct rates and frame error rates on TCP.

3. Our analytical model includes both types of TCP configurations. One, when the TCP window size is allowed to grow large enough so that the session experiences both channel and congestion losses, and two, when the maximum window size is constrained by the receiver advertised window size and hence TCP experiences only channel related losses.

4. We demonstrate how these analytical expressions can be utilized for the selection of RF control variables that determine the channel rates and corresponding FER to maximize TCP throughput. For example, we identify the optimal coding rates to be used in each of the two states when coding rate is used to control data rate and FER. The model is also applied to determine the optimal spreading factors, which is representative of rate control in current CDMA networks. Our studies show that throughput improvements of the order of $15\% - 25\%$ can be obtained for a single TCP session by optimization of the rate adaptation feature.

3.3 Related Work

Numerous approaches have been proposed in the literature to optimize TCP performance in wireless networks. These approaches can be broadly categorized as either *TCP enhancement* approaches or *link-layer* optimization approaches.

TCP enhancement consists of approaches that either introduce end-to-end TCP modifications or *split* the TCP connection with the help of an intelligent agent. Some examples of the former are TCP Westwood [64] that adapt the sending rate to estimated queue sizes, TCP-Freeze [65] which freezes the TCP state and drops sending rate to zero if it detects signal degradation and the Eifel timer [66] to combat spurious time-outs. The latter category comprise intelligent agents placed in the network which are aware of RF conditions and control TCP behavior accordingly. We list below a brief non-exhaustive list of such techniques.

M-TCP [67] proposes a specialized transport between the split point and the mobile while Snoop [68] caches and retransmits lost packets to prevent the TCP source from throttling back due to channel errors. W-TCP [69] is similar to Snoop except that it timestamps packets for accurate estimation of RTT in the presence of retransmissions. Explicit Bad State Notification (EBSN) [70], and Explicit Loss Notification (ELN) [71] send explicit feedback regarding channel state and packet losses to the TCP sender to prevent spurious time-outs and window drops. The ACK-regulator [72] and window-regulator [73] are two techniques to prevent buffer over- and under-flows. They involve agents residing on base stations that monitor the buffer and based on this information control the TCP sending rate either through ACKs or by adapting the maximum receiver window size. We refer the reader to [74, 75] for a more detailed survey.

The framework presented in this chapter is a link-layer optimization approach that, rather than modify TCP to adapt to RF dynamics, adapts the RF layer to TCP dynamics. In this view, our work is closer in philosophy to previous literature that optimizes link layer parameters like Forward Error Correction (FEC) (or coding rate), Automatic Repeat reQuest (ARQ) and RF scheduling to improve TCP throughput.

References [58, 59] analyzed the trade-off between TCP throughput and the amount of FEC added by the link layer. They showed that there exists a coding rate that maximizes TCP throughput though they only consider channel error losses. Reference [60] also conducted a similar study but included the impact of signal power and ARQ as well. Baccelli *et al.* [61] developed an analytical model of TCP that includes the impact of congestion losses due to a finite capacity channel. They used this model to study the impact of both coding rate and processing gain on TCP throughput. See also [76–78] for other cross-layer optimization techniques proposed to improve TCP throughput. All these previous studies however consider a static scenario with only a *single* coding rate and cannot be used to analyze dynamic rate variations.

Adaptive coding on the fly has been studied by the authors of [59, 62] via simulations. In both cases, however, the scheduler behavior is agnostic to TCP *state* and each coding rate is chosen based on expressions for the *long term* throughput of TCP in wireline networks. This ignores the correlation over short time scales between TCP and the scheduler. Kandukuri *et al.* [57] examined the impact of bandwidth oscillations introduced by the CDMA2000 1xRTT RF scheduler on TCP throughput through simulations and laboratory experiments. They observed that the large bandwidth oscillations cause time-outs which can be reduced by using large window sizes. Chan *et al.* [73] proposed a flow-level scheduler, the Short Flow Priority (SFP) scheduler that assigns higher priority to short-lived TCP flows in order to improve TCP performance. They showed through simulations that it performs better that the combination of Proportional Fair (PF) for user-level scheduling and FIFO for flow-level scheduling. Reference [79] proposed a TCP rate aware multi-user EV-DO scheduler to minimize time-outs that is shown to perform well through simulations. Neither, however considers the impact of optimization of RF level parameters on TCP throughput or the correlation between the TCP rate and the scheduler's state. Mattar *et al.* [56] performed a detailed characterization of TCP behavior on CDMA2000 1xRTT channels and the role of key factors like the RLP (Radio Link Protocol) layer and channel conditions. They identified the strong correlation between the TCP sending rate and the channel rates of the CDMA2000 scheduler, as a key characteristic of such networks.

References [80] and [72] have modeled TCP in the presence of variable round trip times and packet losses on wireless links. However, they assume that the variability is independent of TCP dynamics which is at odds with the environment considered here. Finally, the authors of [63] studied optimization of transmission power to maximize TCP throughput. They explicitly consider TCP dynamics in the selection of the transmission power level. However the resulting solutions are quite complex requiring detailed TCP state knowledge. Furthermore, our focus is different from this work since we study the impact of rate adaptation which is not considered by the authors.

3.4 System Model

The next two sections are devoted to presentation of the system model. In this section we propose a TCP-aware RF scheduler that allocates resources based on the TCP sending rate. The motivation for such a scheduler is two-fold. First, it takes into consideration our observations that for optimal performance a scheduler must be aware of TCP dynamics and presents simple guidelines for such a scheduler. Secondly, the same model also allows us to study practical schedulers over long time scales that vary their resource allocation in response to the TCP sending rate. For ease of exposition, we use the CDMA2000 1xRTT system as an example of a practical system to motivate our proposed RF scheduler, although our explanations also apply equally to other wireless systems that dynamically adapt wireless channel rate in response to user sending rate.

3.4.1 The CDMA2000 1xRTT System

Figure 3.1 depicts the wireless hop in a typical cellular network. It comprises of a base station, mobile devices, and a buffer at the base station for each user. The RF scheduling algorithm resides at the base station (or in the case of CDMA2000 1xRTT, the Base Station Controller) and determines the rate allocated to each mobile session. For the purposes of this work, we focus on one such mobile session that involves a TCP bulk transfer on the downlink.

According to CDMA2000 standards [53], whenever, the buffer level for a particular user at the base station exceeds a threshold the scheduler dynamically increases its rate to a pre-specified higher rate by assigning a *supplemental channel*[6]. The

[6] This is also accompanied by a slight increase in signal power to combat fading etc.

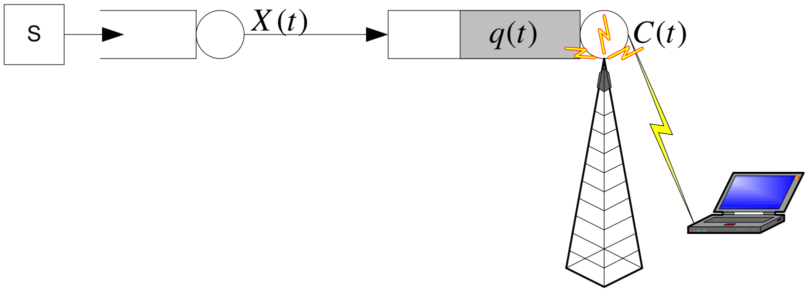


Figure 3.1: Illustration of a cellular hop.

higher rate supplemental channel is achieved by reducing the Walsh code length, which reduces the spreading factor thus increasing data rate. However, the high rate comes at the expense of increased interference (due to smaller code length) which results in higher frame error rates. For example, in CDMA2000 1xRTT, the *fundamental channel* has a data rate of 9.6 Kbps and a target FER of $1\% - 2\%$. In comparison the supplemental channel offers rates ranging from 19.2 Kbps to 153.6 Kbps but with a higher target FER of $5\% - 10\%$. On a different note, assignment of a higher rate to a mobile session prevents other users in the same cell from transmitting at higher rates since the smaller Walsh code is no longer orthogonal to all longer codes that contain it as a prefix as well as resulting in higher inter-cell interference due to increased power. Hence the allocation of a higher rate channel must be done judiciously.

3.4.2 TCP-Aware RF Scheduler: Proposed System Model

The TCP-aware RF scheduler we consider resides at base station. It is quite similar in operation to the 1xRTT scheduler described above, with the exception that it assigns a channel rate based on the *user's TCP sending rate*[7] rather than buffer content[8]. Specifically, whenever the TCP sending rate exceeds (or drops below) the current channel rate, the base station increases (decreases) the channel rate

[7] This of course implies the assumption that the scheduler can measure TCP sending rate, which is defined in the next section.

[8] Alternatively, one could view the scheduler as inspecting buffer content over time scales of a single RTT.

(accompanied by the corresponding trade-off with FER, signal power, *etc.*). If the user exceeds the maximum possible channel rate, the session experiences packet loss due to dropped packets. For analytical tractability, we assume that the scheduler can switch between at most *two* rates. Extending the model to three or more rates is ongoing work.

We present the operations of the scheduler as well as the assumptions regarding the system in more detail below:

1. The RF scheduler decides which of two channel rates C_0 and C_1 are to be assigned based upon the user's TCP sending rate. We assume $C_0 \leq C_1$. If the TCP sending rate is below C_0, the scheduler assigns a channel rate of C_0, otherwise it assigns C_1[9]. In other words:

$$C(t) = \begin{cases} C_0 & \text{if } X(t) \leq C_0 \\ C_1 & \text{otherwise} \end{cases}$$

   where $X(t)$ is the TCP sending rate at time t and $C(t)$ is the assigned channel capacity. The motivation for such a scheduler was presented in Section 3.2. Intuitively, when the TCP sending rate is small, a lower frame error rate is essential (to avoid time-outs, *etc.*). The scheduler can exploit the knowledge of the low TCP sending rate to trade-off channel capacity for channel integrity. At higher TCP sending rates, it is more appropriate to assign a larger channel capacity at the expense of a higher FER. This is because, even though packet loss probability due to channel errors increases, a larger channel capacity prevents packet loss due to congestion, which would have happened with probability *one* were capacity not increased, allowing TCP to transmit at high rates for a longer time

2. The packet error probability is implicitly assumed to be a function of the assigned rate and denoted by $p_0(p_1)$ when the assigned rate is $C_0(C_1)$. This is an important feature representative of current wireless systems where an

[9] In practice a higher rate allocation may not always be possible due to heavy congestion scenarios or deep fading but is typically feasible in low load scenarios. This phenomenon can be easily and naturally incorporated in our model by modeling the *denial* of a higher channel as a random variable.

increase in channel rate typically comes at the cost of increased packet error probability[10]. Hence we assume $p_0 \leq p_1$ since $C_0 \leq C_1$. For simplicity, we refer to (p_i, C_i) together as a *state* or *mode*. We will dwell in more detail on the relation between p and C in Section 3.7.

3. We assume the presence of power control to primarily combat fast fading and interference effects. This is true in current systems where fast closed loop power control tracks a specified target SINR (or equivalently target FER).

4. We assume no (or a very small) buffer at the base station. Hence, TCP experiences congestion if its sending rate exceeds the maximum channel rate (C_1). Although per-user buffers exist at the base stations, they are quite small ($\approx$ 25 Kbytes) in order to accommodate hand-offs. Hence, the assumption of a zero buffer is not unreasonable.

We note that there are several other features of TCP, for example, timeout values, window size, sending state, *etc.* which could potentially be utilized to improve upon our proposed scheduler. Incorporation of such features however, would make the scheduler complex to implement as well as to study. Our aim here is to propose a system that involves minimal modifications to schedulers used with current technologies like CDMA2000 1xRTT, EV-DO, *etc.* and can be studied analytically. From this perspective, we believe that our proposal to incorporate knowledge of only TCP sending rate satisfies both goals.

Another issue worth mentioning is that modern cellular systems have four or more *modes*, *i.e.*, they can support up to four or more different channel rates. However, obtaining succinct analytical expressions even for three is quite difficult. Hence, as a starting point we study two modes to demonstrate the impact of selecting $(C0, C1)$ on TCP performance.

Finally, we emphasize that at this stage, no specific assumptions have been made regarding how the two channel rates are achieved nor how they result in the specific channel error probabilities. Indeed the specific relation is not required in the TCP model and only the actual variables (p_i, C_i) are required. The channel

[10] Of course, our model also covers the simpler scenario where the packet error probability does not change, say due to significant increase in signal power.

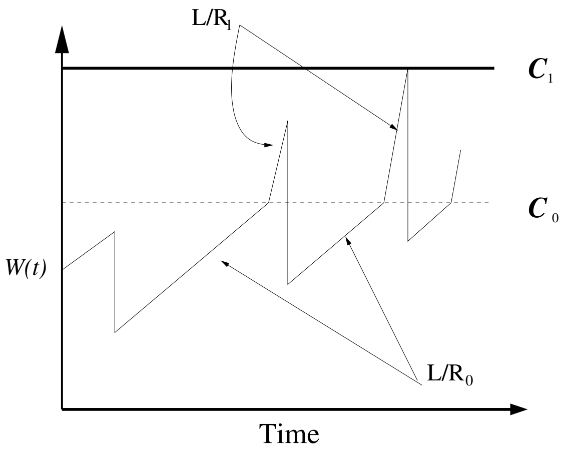


Figure 3.2: TCP window evolution over a variable rate channel.

rates and packet error probabilities are a function of the underlying technology that is used, *e.g.*, adaptive modulation, spreading, *etc.*. This issue is addressed in detail in Section 3.7.

3.5 TCP Model for Variable Channel

In the previous section, we presented the system model for a simple TCP-aware RF scheduler. Such a system results in two distinct operating regimes with different channel capacities, round trip times and packet error probabilities. In this section, we present our TCP model for such a system.

Existing TCP models typically assume that the Round Trip Time (RTT) and packet loss statistics are independent of TCP dynamics in throughput calculations. However, in the wireless environment considered in this work, there exists a strong correlation between the scheduler and TCP sending rate. In particular, the channel capacity C, which affects RTT, as well as packet loss probability p are *functions of the TCP sending rate.* As an illustration, Figure 3.2 depicts the evolution of TCP window size in steady-state when serviced by the proposed TCP-aware RF scheduler. The scheduler assigns rates based on the TCP sending rate and as can be seen, this in turn affects the window *growth* rates. In Section 3.7 we show that ignoring this dependency can result in large errors in throughput prediction.

In order to tackle the impact of the proposed RF scheduler on TCP throughput, we use the model developed by Baccelli *et al.* [61] for a *single fixed rate* as a starting point and develop a model that accounts for the two rate regime. We present a more detailed explanation of our model below:

1. We assume that the TCP version is TCP Reno and model the TCP window growth in steady-state as a fluid process where the window size grows linearly in the absence of loss.

2. The sender is assumed to always have data to send and, for analytical tractability, we ignore time-outs and slow start.

3. Let $W(t)$ denote the window size of TCP at time t and $R(t)$ the round trip time at time t. In the absence of a buffer, if the scheduler is in mode $i = 0, 1$ at time t, we approximate the round trip time $R(t)$ as:

   $$R(t) = R_i = a + L/C_i,$$

   where a is the propagation delay, L is the packet length and C_i is the channel capacity in mode i.

4. Let $X(t)$ denote the instantaneous TCP sending rate at time t in bits/sec, then

   $$X(t) = \frac{W(t)}{R(t)}\,.$$

5. In congestion avoidance mode, the TCP window size increases by roughly one packet (L bits) every R_i seconds in mode i when there is no packet loss. We approximate this in our fluid model with a linear growth rate of L/R_i. Consequently, the sending rate grows at a linear rate of L/R_i^2 bits/sec^2 in the absence of loss. To see this, note that the rate of increase of $X(t)$ is given by:

   $$\frac{X(t+R_i) - X(t)}{R_i} = \frac{W(t+R_i) - W(t)}{R_i^2} = \frac{L}{R_i^2}\,.$$

6. As in [61, 81] we assume that the channel losses can be modeled by an inhomogeneous Poisson process with rate $p_i X(t)$, $i = 0, 1$ at time t.

7. If the TCP sender is not constrained by the receiver window, then TCP experiences congestion with probability one when its sending rate, $X(t)$, exceeds C_1. However, if TCP is constrained by the receiver window size, the sending rate stops increasing once it reaches the maximum advertised window and it experiences only channel related losses.

We need to approach the problem differently depending on whether the two modes (channel rates) satisfy either $C_1 \leq 2C_0$ or $C_1 > 2C_0$. For the important case[11] of $C_1 \leq 2C_0$ we derive expressions for mean TCP throughput for both, the unconstrained rate case as well as the constrained rate case in Sections 3.6.2 and 3.6.3 respectively. Expressions for the TCP throughput for the case when $C_1 > 2C_0$ can be obtained in a similar way but are omitted here due to the fact that most performance improvements are obtained in the region $C_1 \leq 2C_0$.

For ease of exposition, we define some notation.

1. Let $f_i(x,t)$ denote the probability density function of rate $X(t)$ at time t in mode i. Define $P_i(x,t)$ as follows

   $$P_i(x,t) = \mathbb{P}\left\{x \leq X(t) \leq x + dx, C(t) = C_i\right\},$$

   where $C(t)$ is the instantaneous channel rate at time t.Then,

   $$P_i(x,t) = f_i(x,t)dx\,.$$

   It is clear from the above definition that

   $$f(x,t) = f_0(x,t) + f_1(x,t)\,.$$

   where, $f(x,t)$ is the probability density function of $X(t)$.

2. We also define the terms:

   $$\begin{aligned} \delta_i &= \frac{L}{R_i^2}, \quad i = 0, 1, \text{ and,} \\ \gamma_i &= \frac{p_i}{L}, \quad i = 0, 1\,. \end{aligned}$$

[11] In practice, most wireless rates obey the relationship $C_1 = 2C_0$.

3.6 TCP Throughput

3.6.1 Impact of Variable Channel

Before proceeding with the analysis, it is worthwhile discussing an important aspect of the variable rate environment that directly affects the behavior of the TCP sending rate at the *channel transition points*. Let at some time t^-, the TCP sending rate increases to $X(t^-) = C_0$. The corresponding window size is given by $W(t^-) = C_0 \cdot R_0$. As per the proposed policy, the scheduler would then assign a rate of C_1 to the TCP session at time t^+ resulting in a new RTT of R_1. Since TCP is a window-based protocol, the *window size* will be continuous at the rate transition point. Specifically, we have $W(t^+) = W(t^-) = C_0 R_0$. Consequently, the *new* sending rate would be given by

$$\begin{aligned} X(t^+) &= \frac{W(t^+)}{R_1} = \frac{R_0 \cdot C_0}{R_1} \\ &= \frac{R_0}{R_1} C_0 = \frac{R_0}{R_1} X(t^-) \,. \end{aligned}$$

In other words, the TCP sending rate experiences a discontinuous jump by a factor of $g = R_0/R_1$ when the channel capacity transitions from C_0 to C_1. Similar arguments can be used to show that if $X(t) \geq C_0$ and TCP experiences a loss, the sending rate drops by a factor $1/2g$. This aspect of the TCP sending rate must be accounted for in the analysis. In the next two sub-sections, we demonstrate how this feature is incorporated in our model.

3.6.2 Case I: The Rate Unconstrained Case

Recall that in the *rate unconstrained* case, the sender side TCP window size is not limited by the receiver. Hence, TCP experiences congestion loss, with probability one, whenever its sending rate $X(t)$ exceeds C_1. For purposes of analysis we partition the range of the sending rate $X(t)$ into four different regions as shown in Figure 3.3. The discontinuity in $X(t)$ when the channel rate transitions from C_0 to C_1 is clearly visible in the figure at C_0. An interesting observation worth mentioning is that because of the discontinuity, the sample path of $X(t)$ never resides

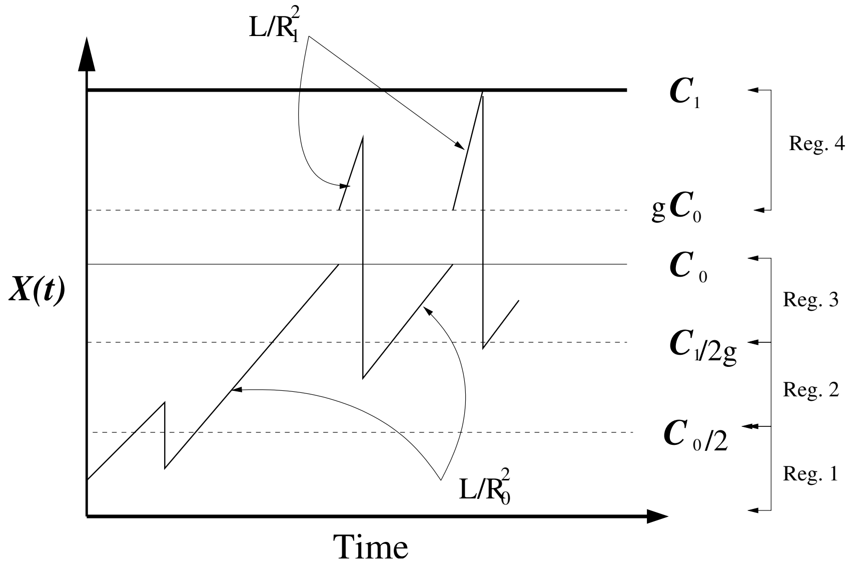


Figure 3.3: Rate evolution and regimes for $C_1 \leq 2C_0$.

in the region (C_0, gC_0). This can be contrasted with the *window size* evolution in Figure 3.2 which has no discontinuities.

In each of these four regions, we use techniques from fluid modeling and develop the forward equations for the probability density function $f_i(x,t)$ of the TCP sending rate by evaluating the probability of a certain event at time $t+dt$ as a function of the events at time t.

We briefly outline, as an example, the derivation of the forward difference equation for the first region $0 \leq X(t) \leq C_0/2$. Let at some time $t+dt$, the sending rate $X(t+dt)$ attains a value, say x. Then only one of *two* possible events could have occurred at time t. One, at time t the TCP sending rate was at value $x - \frac{L}{R_0^2}dt$ and it experiences no loss during the interval $(t, t+dt]$, consequently growing by an amount $(\frac{L}{R_0^2})dt = \delta_0 dt$. Two, at time t, $X(t) = 2x$ and the sender experiences a loss event, resulting in the rate dropping to $2x/2 = x$. Combining these two disjoint events yields the following forward difference equation:

$$P_0(x, t+dt) = (1 - \gamma_0 x dt) P_0(x - dx, t) + (2\gamma_0 x dt) P_0(2x, t)\,.$$

Therefore,

$$f_0(x, t+dt)(\delta_0 dt) = (1 - \gamma_0 x dt) f_0(x - \delta_0 dt, t)(\delta_0 dt) + (2\gamma_0 x) f_0(2x, t)(2\delta_0 dt)\,.$$

Utilizing similar observations and also incorporating regime transitions, one can write the forward difference equations for each of the regimes as shown below:

1. $0 < x < C_0/2$

$$\begin{aligned} f_0(x, t+dt) &= f_0(x - \delta_0 dt, t)(1 - \gamma_0 x dt) + f_0(2x, t) 4\gamma_0 x dt, \\ f_1(x, t+dt) &= 0\,. \end{aligned}$$

2. $C_0/2 < x < C_1/2g$

$$\begin{aligned} f_0(x, t+dt) &= f_0(x - \delta_0 dt, t)(1 - \gamma_0 x dt) + f_1(2x, t) 4\gamma_1 x dt, \\ f_1(x, t+dt) &= 0\,. \end{aligned}$$

3. $C_1/2g < x < C_0$

$$\begin{aligned} f_0(x, t+dt) &= f_0(x - \delta_0 dt, t)(1 - \gamma_0 x dt), \\ f_1(x, t+dt) &= 0\,. \end{aligned}$$

4. $gC_0 < x < C_1$

$$\begin{aligned} f_0(x, t+dt) &= 0, \\ f_1(x, t+dt) &= f_1(x - \delta_1 dt, t)(1 - \gamma_1 x dt)\,. \end{aligned}$$

After rearrangement we have,

1. $0 < x < C_0/2$

$$\begin{aligned} f_0(x, t+dt) - f_0(x - \delta_0 dt, t) &= -\gamma_0 x dt f_0(x - \delta_0 dt, t) + f_0(2x, t) 2\gamma_0 x dt, \\ f_1(x, t+dt) &= 0\,. \end{aligned}$$

2. $C_0/2 < x < C_1/2g$

$$\begin{aligned} f_0(x, t+dt) - f_0(x - \delta_0 dt, t) &= -\gamma_0 x dt f_0(x - \delta_0 dt, t) + f_1(2x, t) 2\gamma_1 x dt, \\ f_1(x, t+dt) &= 0\,. \end{aligned}$$

3. $C_1/2g < x < C_0$

$$\begin{aligned} f_0(x,t+dt) - f_0(x-\delta_0 dt,t) &= -\gamma_0 x dt f_0(x-\delta_0 dt,t), \\ f_1(x,t+dt) &= 0\,. \end{aligned}$$

4. $gC_0 < x < C_1$

$$\begin{aligned} f_0(x,t+dt) &= 0, \\ f_1(x,t+dt) - f_1(x-\delta_1 dt,t) &= -\gamma_1 x dt f_1(x-\delta_1 dt,t)\,. \end{aligned}$$

Letting $dt \to 0$, we can obtain a set of partial differential equations for $f_i(x,t)$. Notice that

$$\begin{aligned} \lim_{dt\to 0} \frac{f(x,t+dt) - f(x-Adt,t)}{dt} &= \lim_{dt\to 0} \frac{f(x,t+dt) - f(x,t) + f(x,t) - f(x-Adt,t)}{dt} \\ &= \frac{\partial}{\partial t} f(x,t) + A\frac{\partial}{\partial x} f(x,t) \end{aligned} \tag{3.1}$$

If $f(x,t)$ has a steady-state distribution $f(x)$ then the right hand side of (3.1) is simply equal to $A\frac{d}{dx}f(x)$ as $t \to \infty$. Under the assumption that $X(t)$ has a steady-state distribution, we obtain the following system of differential equations for the distribution $f_i(x)$, where we denote $\frac{d}{dx}f_i(x)$ by $\dot{f}_i(x)$:

1. $0 < x < C_0/2$

$$\begin{aligned} \delta_0 \dot{f}_0(x) &= -\gamma_0 x f_0(x) + 4\gamma_0 x f_0(2x), \qquad (3.2) \\ f_1(x) &= 0\,. \end{aligned}$$

2. $C_0/2 < x < C_1/2g$

$$\begin{aligned} \delta_0 \dot{f}_0(x) &= -\gamma_0 x f_0(x) + 4g\gamma_1 x f_1(2gx), \qquad (3.3) \\ f_1(x) &= 0\,. \end{aligned}$$

3. $C_1/2g < x < C_0$

$$\begin{aligned} \delta_0 \dot{f}_0(x) &= -\gamma_0 x f_0(x), \qquad (3.4) \\ f_1(x) &= 0\,. \end{aligned}$$

4. $gC_0 < x < C_1$

$$\begin{aligned} f_0(x) &= 0, \\ \delta_1 \dot{f}_1(x) &= -\gamma_1 x f_1(x)\,. \end{aligned} \tag{3.5}$$

Equally important, we also obtain the following boundary conditions:

1. At $x = \frac{C_0}{2}$,

$$f_0((C_0/2)^+) = f_0((C_0/2)^-)\,. \tag{3.6}$$

To see how this is obtained, writing out the forward difference equation for $x = C_0/2$, we have

$$\begin{aligned} f_0(C_0/2, t+dt)(\delta_0 dt) &= f_0(C_0/2 - \delta_0 dt, t)(\delta_0 dt)(1 - \gamma_0 C_1/2dt) \\ &\quad + f_0(2C_0, t)(2\delta_0 dt)(\gamma_0 C_0 dt)\,. \end{aligned}$$

In the limit as $t \to \infty$ and $dt \to 0$, we obtain

$$f_0((C_0/2)^+) = f_0((C_0/2)^-)\,. \tag{3.7}$$

2. At $x = \frac{C_1}{2g}$,

$$\frac{1}{R_0^2} \cdot f_0((C_1/2g)^+) = \frac{1}{R_0^2} \cdot f_0((C_1/2g)^-) + \frac{1}{R_1^2} f_1(C_1^-)\,. \tag{3.8}$$

This is a critical boundary condition that highlights the impact of different channel rates. The forward difference equation at $x = C_1/2g$ can be written as:

$$\begin{aligned} f_0(C_1/2g, t+dt)(\delta_0 dt) &= f_0(C_1/2g - \delta_0 dt, t)(\delta_0 dt)(1 - \gamma_0 C_1/2g \cdot dt) \\ &\quad + f_1(C_1 - \delta_1 dt, t)(\delta_1 dt)(1 - \gamma_1 C_1 dt) \end{aligned}$$

This expression comes about because the probability that at time $t + dt$, the rate is $C_1/2g$, is equal to the probability that at time t, the rate was in the neighborhood of $C_1/2g$ *and* there was no loss *or* the rate was in the neighborhood of C_1 and again there was no loss (in which case the rate would hit C_1 and instantaneously drop down to $C_1/2g$). An important point worth noting in the above relation is the impact of different growth rates $1/R_0^2$ and $1/R_1^2$. In the limit, as $t \to \infty$ and $dt \to 0$, we obtain the above boundary condition.

3. A similar application of the forward difference equation yields the boundary condition at $x = C_0$ to be ,

$$\frac{1}{R_1^2} \cdot f_1((gC_0)^+) = \frac{1}{R_0^2} \cdot f_0(C_0^-)\,. \tag{3.9}$$

The above differential equations could be solved numerically in order to obtain the rate distribution $\{f_i(t)\}$. However, the actual quantity of interest for optimization is usually the mean TCP throughput, and fortunately analytical expressions for it can be obtained, far more easily by the use of Mellin transforms. They were previously used in similar settings by the authors of [61, 82, 83] and we will follow their general approach.

Let $\widehat{f}(u)$, $u \geq 1$ be the Mellin transform of some probability distribution function $f(x)$ defined by

$$\widehat{f}(u) = \int_0^\infty f(x) x^{u-1} dx\,. \tag{3.10}$$

Then the mean of X denoted by $\overline{X}$ is simply given by $\overline{X} = \widehat{f}(2)$.

Define ϕ_{ij} as follows

$$\phi_{ij} = \frac{\gamma_i}{\delta_j} = \frac{p_i R_j^2}{L^2}\,. \tag{3.11}$$

Multiplying (3.2)–(3.5) by x^u, integrating them over the respective limits and summing them yields, after some algebraic manipulations:

1. $0 < x < C_0/2$

$$\begin{aligned} f_0((C_0/2)^-)(C_0/2)^u - u \int_0^{C_0/2} f_0(x) x^{u-1} dx \\ = -\phi_{00} \int_0^{C_0/2} f_0(x) x^{u+1} dx + \frac{\phi_{00}}{2^u} \int_0^{C_0} f_0(y) y^{u+1} dy \end{aligned}$$

2. $C_0/2 < x < C_1/2g$

$$f_0((C_1/2g)^-)(C_1/2g)^u - f_0((C_0/2)^+)(C_0/2)^u - u\int_{C_0/2}^{C_1/2g} f_0(x)x^{u-1}dx$$
$$= -\phi_{00}\int_{C_0/2}^{C_1/2g} f_0(x)x^{u+1}dx + \frac{\phi_{10}}{2^u \cdot g^{u+1}}\int_{gC_0}^{C_1} f_1(y)y^{u+1}dy$$

3. $C_1/2g < x < C_0$

$$f_0(C_0^-)(C_0)^u - f_0((C_1/2g)^+)(C_1/2g)^u - u\int_{C_1/2g}^{C_0} f_0(x)x^{u-1}dx$$
$$= -\phi_{00}\int_{C_1/2g}^{C_0} f_0(x)x^{u+1}dx$$

4. $gC_0 < x < C_1$

$$f_1(C_1^-)(C_1)^u - f_1(gC_0^+)(gC_0)^u - u\int_{gC_0}^{C_1} f_1(x)x^{u-1}dx$$
$$= -\phi_{11}\int_{gC_0}^{C_1} f_1(x)x^{u+1}dx$$

Summing all the above equations together and substituting the boundary conditions yields

$$(\frac{R_0}{R_1})^2(C_0)^u f_1(gC_0^+) - (gC_0)^u f_1(gC_0^+) - (\frac{R_0}{R_1})^2(C_1/2g)^u f_1(C_1^-) + (C_1)^u f_1(C_1^-)$$
$$- u\widehat{f_0}(u) - u\widehat{f_1}(u) = -\phi_{00}(1-\frac{1}{2^u})\widehat{f_0}(u+2) + (\frac{\phi_{10}}{2^u g^{u+1}} - \phi_{11})\widehat{f_1}(u+2), \quad (3.12)$$

therefore,

$$\Big((\frac{R_0}{R_1})^2 - g^u\Big)(C_0)^u f_1(gC_0^+) - \Big((\frac{R_0}{R_1})^2\frac{1}{2^u g^u} - 1\Big)(C_1)^u f_1(C_1^-)$$
$$- u\widehat{f_0}(u) - u\widehat{f_1}(u) = -\phi_{00}(1-\frac{1}{2^u})\widehat{f_0}(u+2) + (\frac{\phi_{10}}{2^u g^{u+1}} - \phi_{11})\widehat{f_1}(u+2)\,. \quad (3.13)$$

Fortunately, the differential equation for region $x \in (gC_0, C_1)$ can be directly solved to obtain a very simple expression. We have,

$$f_1(x) = V_0 e^{-\phi_{11}x^2/2} \quad C_0 < x < C_1$$

Due to the simple form of $f_1(x)$, $\widehat{f_1}(u)$ can be directly computed from $f_1(x)$. Let us define $\Delta(u)$ as follows

$$\begin{aligned}\Delta(u) &= \int_{C_0}^{C_1} e^{-\frac{\phi_{11}}{2}x^2} x^{u-1} dx \\ &= \frac{1}{2}(\frac{\phi_{11}}{2})^{-\frac{u}{2}} \int_{\frac{\phi_{11}}{2}(C_0)^2}^{\frac{\phi_{11}}{2}(C_1)^2} e^{-y} y^{\frac{u}{2}-1} dy \\ &= \frac{1}{2}(\frac{\phi_{11}}{2})^{-\frac{u}{2}} \Big[\gamma\big(u/2, \frac{\phi_{11}}{2}(C_1)^2\big) - \gamma\big(u/2, \frac{\phi_{11}}{2}(C_0)^2\big)\Big]\end{aligned} \tag{3.14}$$

where γ is the lower incomplete gamma function defined as

$$\gamma(a,x) = \int_0^x e^{-t} t^{a-1} dt\,.$$

Then,

$$\widehat{f_1}(u) = V_0 \Delta(u)\,. \tag{3.15}$$

Furthermore,

$$\begin{aligned}f_1(C_1^-) &= V_0 e^{-\phi_{11}C_1^2/2} \\ f_1(gC_0^+) &= V_0 e^{-\phi_{11}(gC_0)^2/2}\,.\end{aligned}$$

By substituting (3.15) in (3.13) and rearranging the result, it is obtained that

$$\begin{aligned}\widehat{f_0}(u) &= \frac{\phi_{00}}{u}(1 - \frac{1}{2^u})\widehat{f_0}(u+2) + \frac{V_0}{u}\psi(u) \\ \widehat{f_1}(u) &= V_0\Delta(u)\end{aligned} \tag{3.16}$$

where

$$\begin{aligned}\psi(u) = &(g^2 - g^u)(C_0)^u e^{-\frac{\phi_{11}}{2}g^2C_0^2} \\ &- \Big((\frac{1}{2g})^u g^2 - 1\Big)(C_1)^u e^{-\frac{\phi_{11}}{2}C_1^2} \\ &- u\Delta(u) - (\frac{\phi_{10}}{2^u g^{u+1}} - \phi_{11})\Delta(u+2)\,.\end{aligned}$$

By expanding the recursive relation in (3.16), we obtain the following expression for $\widehat{f_0}(u)$:

$$\widehat{f_0}(u) = V_0 \sum_{k\geq 0} (\phi_{00})^k \Pi_k(u) \psi(u+2k), \tag{3.17}$$

where,

$$\Pi_k(u) = \frac{1}{u+2k}\prod_{i=0}^{k-1}\Big(\frac{1-2^{-u-2i}}{u+2i}\Big).$$

It is left to find the unknown constant V_0. Substituting $u = 1$ in (3.10), one easily obtains the normalization condition: $\widehat{f}(1) = \widehat{f_0}(1) + \widehat{f_1}(1) = 1$. Utilizing (3.16) we have,

$$V_0 = \frac{1}{\Delta(1) + \sum_{k\geq 0}(\phi_{00})^k\Pi_k(1)\psi(1+2k)}. \tag{3.18}$$

Finally, the mean TCP throughput is given by $\widehat{f}(2) = \widehat{f_0}(2) + \widehat{f_1}(2)$ which can be expressed as

$$\overline{X} = \frac{\Delta(2) + \sum_{k\geq 0}(\phi_{00})^k\Pi_k(2)\psi(2+2k)}{\Delta(1) + \sum_{k\geq 0}(\phi_{00})^k\Pi_k(1)\psi(1+2k)}. \tag{3.19}$$

3.6.3 Case II: The Rate Constrained Case

In the previous sub-section we assumed that sending rate of the source was not constrained by the receiver. Hence the sender side window can grow till it exceeds the maximum bandwidth-delay product of the channel in which case it experiences a congestion loss. However, often the receiver advertises a window size W_{max} that is smaller than the peak bandwidth-delay product. Consequently, the sender size TCP window stops growing once it reaches this advertised window. Equivalently, the sending rate stops growing once it hits C_1[12]. In such a state, if we assume per-user isolation (which is common in modern cellular systems), the source does not experience congestion and the TCP window will drop only due to channel errors.

This behavior is shown in Figure 3.4. From the figure it is clear that a density mass exists at C_1 and the stationary rate density distribution function $f(x)$ has a discontinuity at C_1. We let

$$f_1(C_1) = f_1(x)\mid_{x=C_1} = A\delta(x - C_1) \tag{3.20}$$

where A is some proportionality constant and $\delta(x)$ is the Dirac Delta function.

[12] For simplicity, we assume that the receiver advertised window is perfectly sized, that is $W_{max} = C_1 \cdot R_1$. Of course, our model also works when $W_{max} \leq C_1 \cdot R_1$, in which case we simply choose a new peak capacity $C_1' = \frac{W_{max}}{R_1}$.

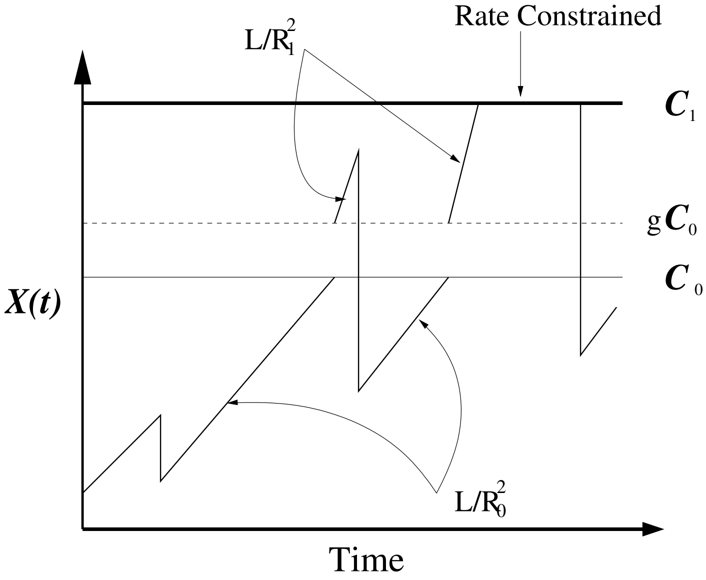


Figure 3.4: TCP evolution when $W_{max} = C_1 \cdot R_1$.

The discontinuity does not affect the differential equations. It does however affect the boundary conditions at $C_1/2g$ and C_1^- (the boundary condition at C_0^- remains unchanged). Solving for the first discontinuity by conditioning on $X(t + dt) = C_1/2g$ yields,

$$\begin{aligned} f_0(C_1/2, t+dt)(\delta_0 dt) &= f_0(C_1/2 - \delta_1 dt, t)(\delta_0 dt)(1 - \gamma_1(C_1/2)dt) \qquad (3.21) \\ &\quad + \Pr\{X(t) = C_1\}(\gamma_1 C_1 dt) \\ &= f_0(C_1/2 - \delta_1 dt, t)(1 - \gamma_1 C_1 dt)(\delta_0 dt) + A(\gamma_1 C_1 dt) \end{aligned}$$

Letting $t \to \infty$ and $dt \to 0$, we have

$$\frac{1}{R_0^2} f_0((C_1/2g)^+) = \frac{1}{R_0^2} f_0((C_1/2g)^-) + \gamma_1 C_1 A \qquad (3.22)$$

where, recall that $\gamma_1 = p_1/L$. Similarly, conditioning on $X(t + dt) = C_1$, we have

$$\begin{aligned} \Pr\{X(t+dt) = C_1\} &= f_1(C_1 - \delta_1 dt, t)(\delta_1 dt)(1 - \gamma_1(C_1)dt) \qquad (3.23) \\ &\quad + \Pr\{X(t) = C_1\}(1 - \gamma_1 C_1 dt) \end{aligned}$$

Letting $t \to \infty$ and $dt \to 0$ we have,

$$\frac{L}{R_1^2} f(C_1^-) = \gamma_1 C_1 A,$$

or,

$$f_1(C_1^-) = \frac{R_1^2}{L} \gamma_1 C_1 A \,. \tag{3.24}$$

Using (3.24) we can re-write the boundary condition at $x = C_1/2g$ ((3.22)) as

$$\frac{1}{R_0^2} f_0((C_1/2g)^+) = \frac{1}{R_0^2} f_0((C_1/2g)^-) + \frac{L}{R_1^2} f_1(C_1^-) \,.$$

This is essentially the same boundary condition as the one for the unconstrained rate case in Section 3.6.2 (the change is hidden in $f(C_1^-)$). Consequently, we can proceed in exactly the same manner as in Section 3.6.2 over the region $[0, C_1)$ to derive the mean throughput.

This yields, as before,

$$\widehat{f_0}(u) = V_0 \sum_{k \geq 0} (\phi_{00})^k \Pi_k(u) \psi(u + 2k),$$
$$\widehat{f_1}(u) = V_0 \Delta(u)$$

where all the variables retain their original meanings defined in the previous subsection.

The difference from the unconstrained rate case is that the normalization and throughput relations are now given by

$$\widehat{f_0}(1) + \widehat{f_1}(1) + \int_{C_1^-}^{C_1^+} A\delta(x - C_1)dx = 1$$

therefore,

$$\widehat{f_0}(1) + \widehat{f_1}(1) + A = 1 \tag{3.25}$$

and,

$$\overline{X} = \widehat{f_0}(2) + \widehat{f_1}(2) + \int_{C_1^-}^{C_1^+} xA\delta(x - C_1)dx$$

therefore,

$$\overline{X} = \widehat{f_0}(2) + \widehat{f_1}(2) + C_1 A\,. \tag{3.26}$$

Through the direct solution of the differential equation in the region $gC_0 < x < C_1$ and the boundary condition at $x = C^-$, we can relate A and V_0 as follows.

$$f_1(C_1^-) = V_0 e^{-\phi_{11} C_1^2/2} = \frac{R_1^2}{L}\gamma_1 C_1 A,$$

therefore,

$$V_0 = \frac{R_1^2}{L}\gamma_1 C_1 e^{\phi_{11} C_1^2/2} \cdot A\,. \tag{3.27}$$

For simplicity, denote $Z = \frac{R_1^2}{L}\gamma_1 C_1 e^{\phi_{11} C_1^2/2}$. By plugging this relation and the expressions for $\widehat{f_0}(u)$ and $\widehat{f_1}(u)$ in the normalization relation, we derive an expression for A;

$$A = \frac{1}{1 + Z\Big(\Delta(1) + \sum_{k\geq 0}(\phi_{00})^k \Pi_k(1)\psi(1+2k)\Big)}\,.$$

The mean throughput can now be computed to be:

$$\overline{X} = \frac{C_1 + Z\Big(\Delta(2) + \sum_{k\geq 0}(\phi_{00})^k \Pi_k(2)\psi(2+2k)\Big)}{1 + Z\Big(\Delta(1) + \sum_{k\geq 0}(\phi_{00})^k \Pi_k(1)\psi(1+2k)\Big)}\,. \tag{3.28}$$

where $\psi(u)$ retains the definition from the previous sub-section.

3.7 Numerical Results

In this section, we verify the accuracy of our model by comparison against *ns-2* simulations and then demonstrate how the model can be used to optimize resources on a wireless channel. Before evaluating the performance of the model, we must address the issue of how rate adaptation by the RF scheduler affects both channel capacity and packet error probability. Note that our model does not assume any specific dependencies between the various channel rates and packet error probabilities, rather it assumes these are pre-determined input variables.

As mentioned in Section 3.2, in current CDMA systems rate adaptation is achieved by changing any of three variables: error coding rate, spreading factor or modulation scheme. In this work we study the impact of the first two variables, though our model is equally applicable to any other means of controlling the data rate. We initially assume that the channel rate and packet error are controlled solely by the coding rate. This is primarily to provide a simple basis to evaluate the accuracy of the model as well as showcase its utility. Section 3.7.1 presents the relationship we use to quantify the impact of coding rate on channel capacity and packet error probability. We use this relationship to: (i) evaluate the accuracy of our model in Sections 3.7.2 and 3.7.4, and (ii) choose coding rates to maximize TCP throughput in Section 3.7.3.

Finally, in Section 3.7.5, we apply the model to scenarios where resource allocation is controlled by changing the *spreading factor*. This is representative of current CDMA networks where spreading factor is the dominant control knob to adapt data rates. We evaluate the gain in TCP throughput and resultant trade-off with energy consumption as a function of the spreading factors.

Another potential control factor in wireless networks is the retransmission mechanism typically deployed at radio link layer to mitigate high frame error rates by retransmitting erroneous frames. Our model can easily incorporate the impact of link-layer retransmissions as a trade-off between channel rate and packet error probability similar to the coding rate studied here, though, it cannot account for the latency introduced by link layer retransmissions. However, extensive measurements conducted over a commercial 1xRTT network [56] found that the impact of rate changes of the scheduler on TCP sending rate is far more dominant, while there is little or no correlation between TCP round-trip times and the link-layer retransmissions primarily due to large RTTs and very fast re-transmissions. Hence we do not study the impact of link layer re-transmissions in this work.

3.7.1 Packet Error Probability: Variable Coding

As mentioned previously, we assume that the capacity C_i and packet error probability p_i in mode i are functions of the coding rate ρ_i. Hence, for a given bit error probability, the TCP throughput is a function of the two coding rates, *i.e.*,

$\overline{X}(\rho_0, \rho_1)$. The relation between the capacity and coding rate is straightforward and is given by

$$C_i = \rho_i \cdot C^*, \tag{3.29}$$

where C^* is the uncoded channel capacity.

The packet error probability however is strongly dependent on not just the coding rate, but also the coding scheme used. Consequently, one must either choose a specific coding scheme, or resort to bounds on the achievable packet error probability for a given coding rate.

One such bound is the Gilbert-Varshamov [84] bound. This was used in [61] and we will also use it as an approximation. The Gilbert-Varshamov bound is a bound on the parameters of a code of length B and information bit length $K \leq B$. It specifies that there exists a minimum Hamming distance d between any two codewords that must satisfy

$$2^B \leq 2^K \sum_{j=0}^{d-1} \binom{B}{j} \tag{3.30}$$

where $d - 1 \leq B/2$. Such a code can correct at most $t = \lfloor (d-1)/2 \rfloor$ errors. Hence, the above relation bounds the maximum number of correctable errors for any coding rate.

This relationship is used to determine the packet error probability for a wireless system as follows. Let p_e denote the bit error probability for the wireless channel in consideration. Suppose that TCP packets have fixed size of L bits. The TCP packets are broken up into RLP (Radio Link Protocol)[13] frames of size B bits for transmission over the wireless channel. Each radio block is assumed to have K bits of information and $(B - K)$ bits for coding. Hence the coding rate is $\rho = K/B$. The packet error probability is

$$p = 1 - (1 - p_b)^{\lceil \frac{L}{\rho B} \rceil}, \tag{3.31}$$

where p_b is the radio block error probability. Using the Gilbert-Varshamov bound, for a given coding rate, we can determine the maximum number of error bits, t, that are correctable using the coding scheme with rate ρ. Then,

[13] The RLP layer is described in the CDMA2000 standards.

$$p_b = \sum_{j=t+1}^{B} \binom{B}{j} p_e^j (1-p_e)^{B-j} \,. \tag{3.32}$$

Obviously, packet error probability increases as the coding rate increases. We note that in the uncoded case $\rho = 1$ and $t = 0$ in which case, $p_b = 1 - (1 - p_e)^B$. As an example, in Figures 3.5(a), we have plotted the packet error probability as a function of coding rate ρ.

3.7.2 Adaptive Coding Evaluation: Rate Unconstrained

We begin our evaluation ofthe TCP model by comparing its accuracy against simulations of the TCP-aware RF scheduler implemented in *ns-2*. In this sub-section we study the case when the TCP sending rate is not limited by the receiver advertised window. To be concrete in this section, we set the raw channel rate C^* to 128 Kbps and the two-way propagation delay a to 200 ms. TCP packet size is set to 1024 bits. A TCP packet is divided into radio blocks of size 256 bits for transmission over the wireless channel.

The same parameters were also used for the *ns-2* simulation. TCP Reno was chosen (since the model fits that version best) and packet errors were assumed to be independent and identically distributed[14]. Since we do not account for time-outs in our model, we only simulated scenarios with large window sizes that result in few timeouts. Indeed, a previous study has recommended the use of large window sizes in cellular networks to precisely avoid such time-outs due to bandwidth oscillations [57]. The duration of each simulation run was 1000 seconds. The bit error probability was held constant for the duration of the simulation, which is true under perfect power control. Hence the packet error probability is solely a function of the coding rate. We ran 20 simulations with different random seeds for each data point and the results reported are in the 95% confidence interval.

It is worth mentioning that we limited ourselves to low packet error probabilities in all the scenarios. The reasons for this are two-fold. First, TCP is known to perform well only for low packet errors (less than 5%) and hence it represents

[14] Typically in wireless channels, errors are bursty affecting multiple packets. From the perspective of our model, this simply shows up as a single congestion event.

the region of interest. The second reason has to do with modeling the packet loss process as an inhomogeneous Poisson process which is reasonable only for low packet error probabilities. To see this, let us assume that the TCP window size is at its maximum (say CR) and compute the probability that at least one of the packets sent in the time window of one round trip, R, is lost. This can be written as

$$\mathbb{P}\left\{\text{at least one packet loss}\right\} = 1 - (1-p)^{CR}\,.$$

If the product pRC is sufficiently small, the above expression can be approximated as

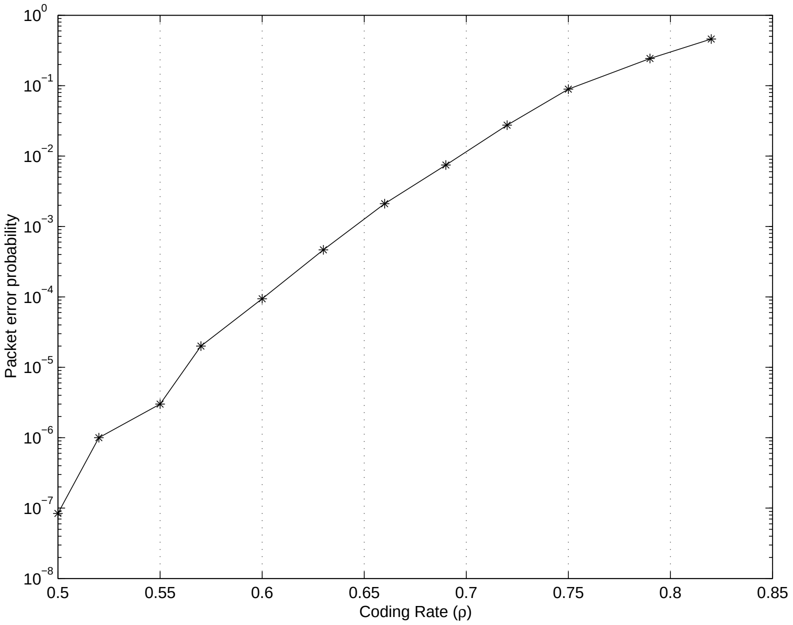
$$1 - (1-p)^{CR} \approx pCR\,.$$

Observe that this is also an approximation for a single congestion event in a time bin R for the inhomogeneous Poisson loss process for small values of p. Consequently, one can expect a reasonable match between the model and simulated loss process. In our numerical experiments, we found that $pRC < 0.1$ leads to accurate prediction of TCP throughput. We now present our comparison results.

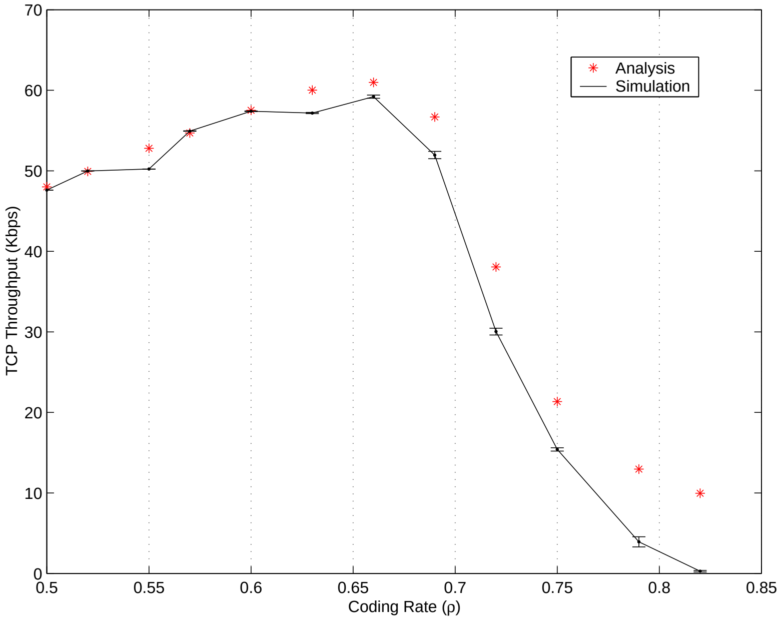
To begin, we show the results of comparison of the model and simulations for the special case of $\rho_0 = \rho_1$, *i.e.*, there is only a *single rate*. In this case, our model simplifies to the scenario considered in [61] where a *single* static coding rate is utilized. We plot the packet error probability as a function of coding rate in Figure 3.5(a) for a target bit error probability of 10^{-2} and the corresponding throughputs of both the model and simulation as a function of coding rate in Figure 3.5(b). One observes the close agreement between the model and simulation. The exact nature of the curves are discussed in the next section. We also studied the performance for target bit-error probabilities of 10^{-3} and 10^{-4} which showed similar results. The results are depicted in Figures 3.6 and 3.7.

Figures 3.8–3.10 present results for the more general case when ρ_0 and ρ_1 are different. To show the accuracy of the model, we have depicted the percentage of error between the model and simulation across all feasible[15] coding rates (ρ_0, ρ_1) on a three dimensional grid for different bit error probabilities $(10^{-2}, 10^{-3}, 10^{-4})$ in

[15] By “feasible”, we imply $\rho_0 < \rho_1$, $\rho_1 \leq 2\rho_0$ and $p_i \ll 1$.

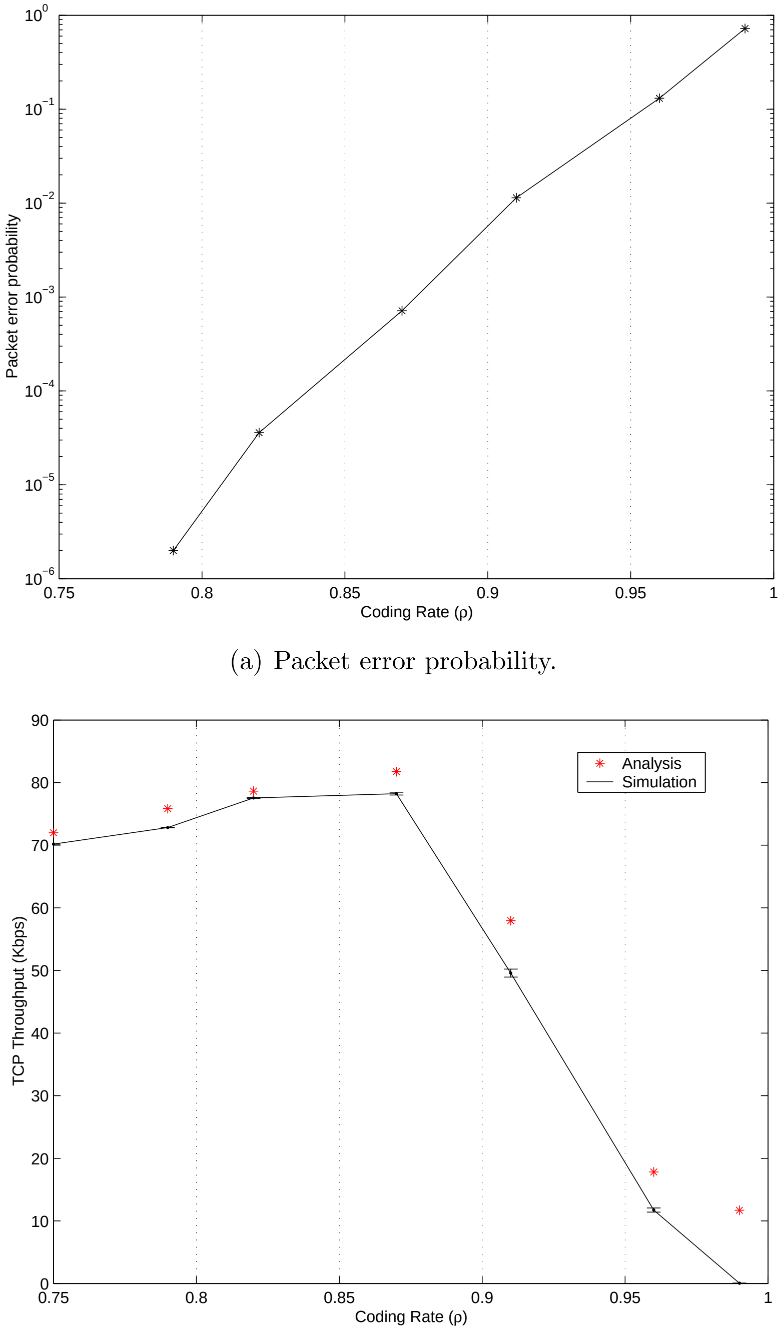


(a) Packet error probability.



(b) TCP throughput.

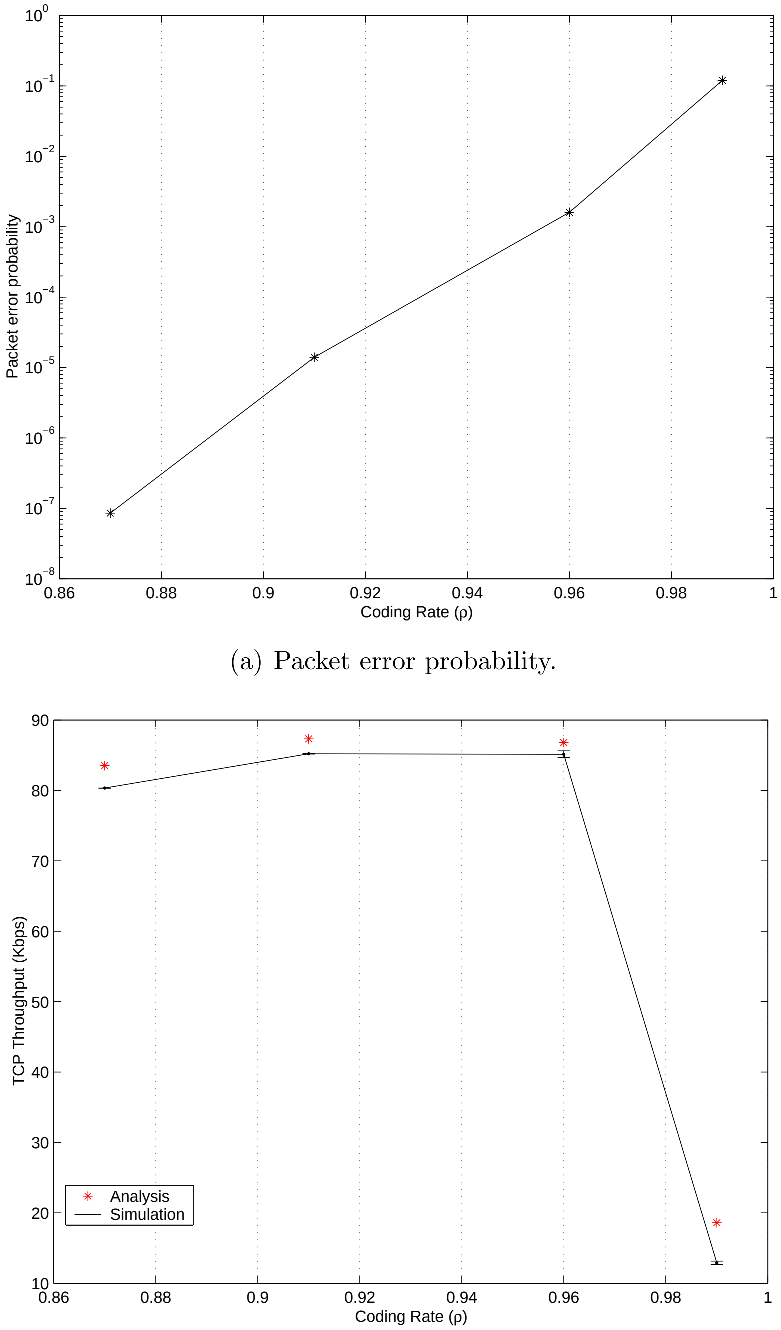
Figure 3.5: *Single-rate case*: $p_e = 10^{-2}$.



(a) Packet error probability.

(b) TCP throughput.

Figure 3.6: *Single-rate case*: $p_e = 10^{-3}$.



(a) Packet error probability.

(b) TCP throughput.

Figure 3.7: *Single-rate case*: $p_e = 10^{-4}$.

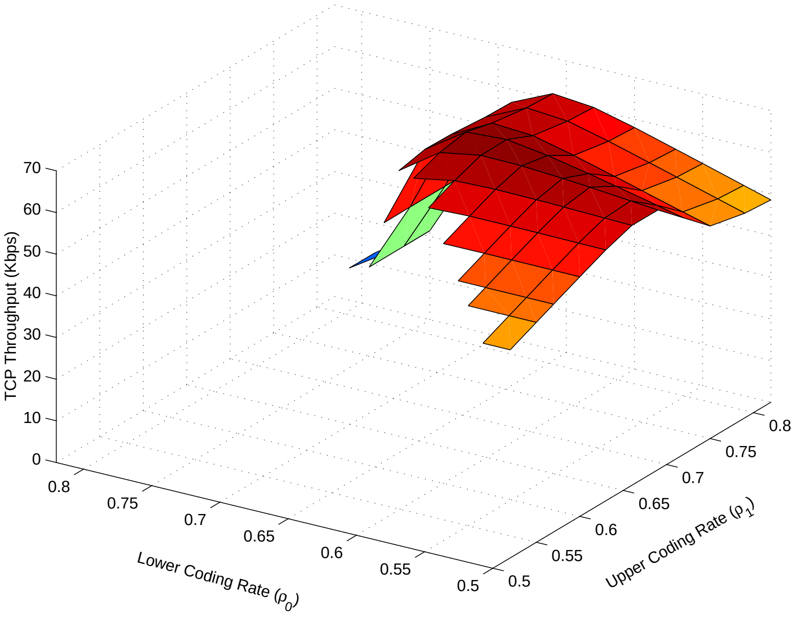


Figure 3.8: *Two-rate case*: TCP throughput ($p_e = 10^{-2}$).

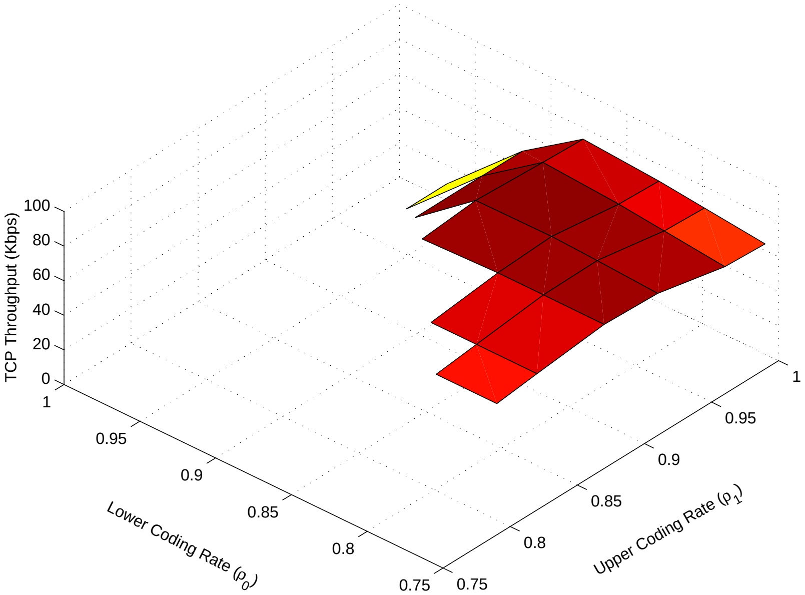


Figure 3.9: *Two-rate case*: TCP throughput ($p_e = 10^{-3}$).

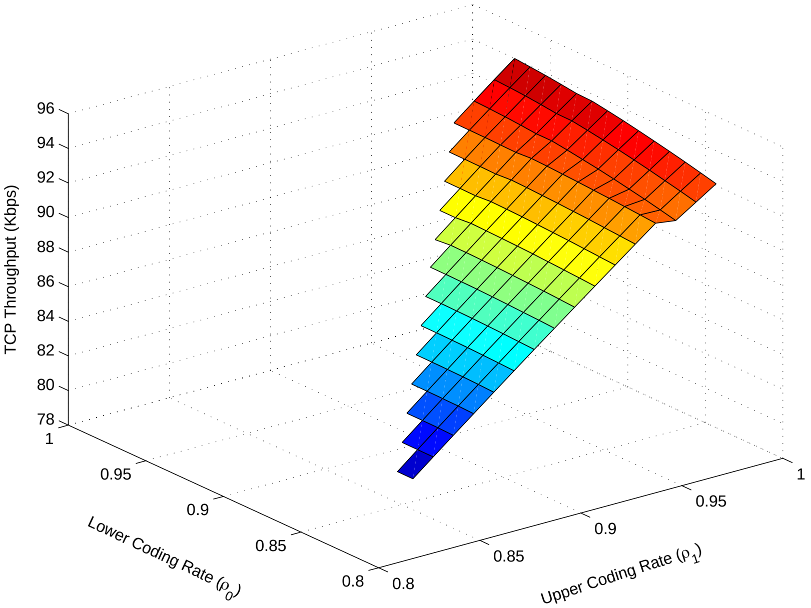


Figure 3.10: *Two-rate case*: TCP throughput ($p_e = 10^{-4}$).

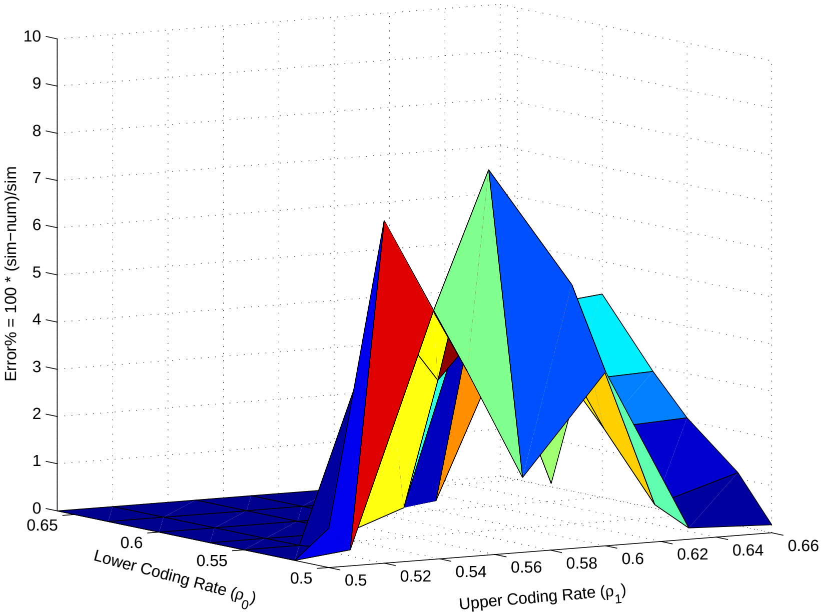


Figure 3.11: Difference between model and simulation ($p_e = 10^{-2}$).

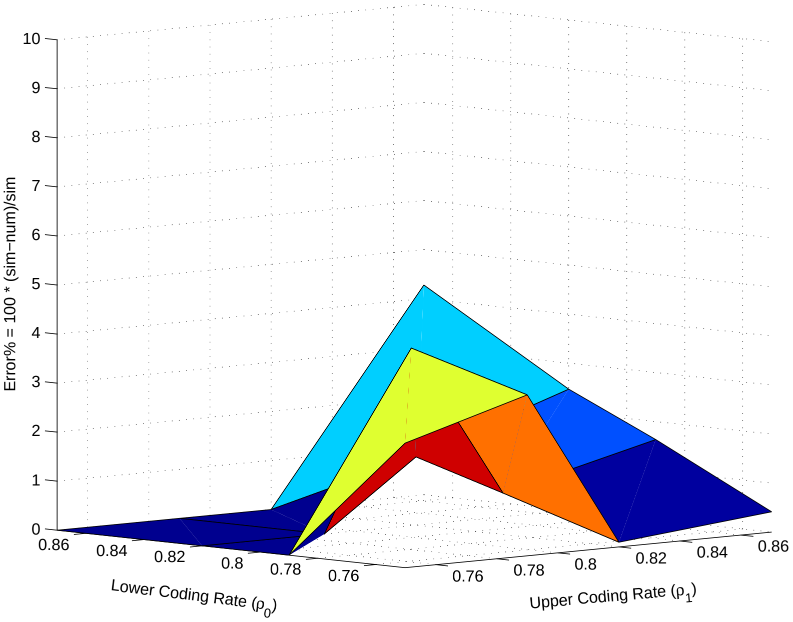


Figure 3.12: Difference between model and simulation ($p_e = 10^{-3}$).

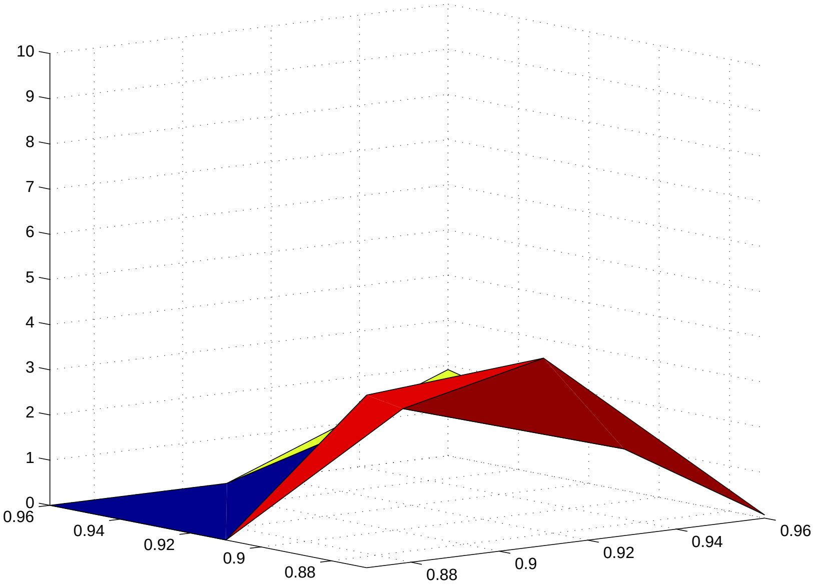


Figure 3.13: Difference between model and simulation ($p_e = 10^{-4}$).

Figures 3.11–3.13. Observe again that the model matches the simulation results closely with errors typically less than 5%.

To demonstrate the importance of capturing the correlation between TCP and the scheduler as well as the presence of two states, we plot in Figure 3.14 the (percentage) difference between the two-rate model (which we have shown above to be accurate) and a single-rate model that takes only one set of parameters, either those due to coding rate ρ_0 or those due to ρ_1 into consideration.

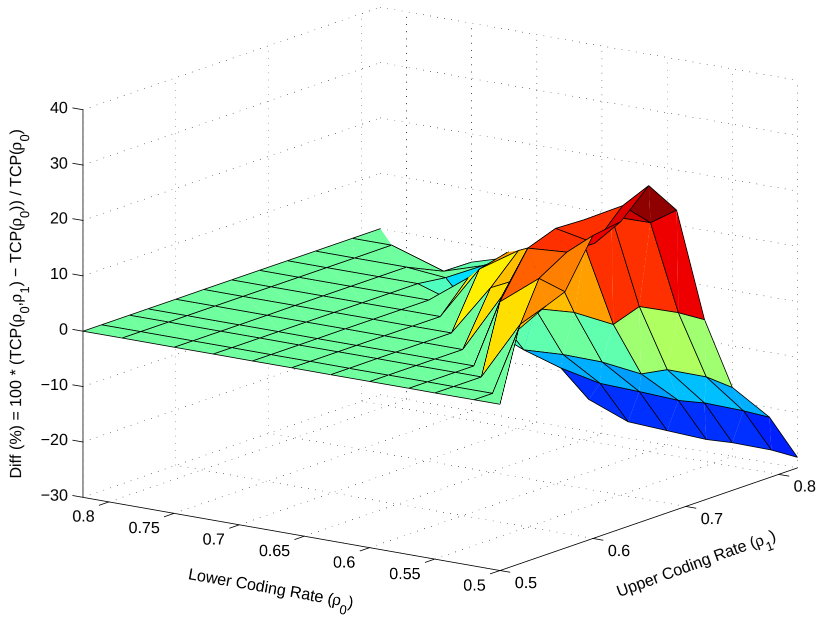
We observe that there exist regions where there is substantial difference in predicted throughput. Hence a single-rate model may not be always feasible to tune performance in such systems.

3.7.3 Optimal Coding Rates

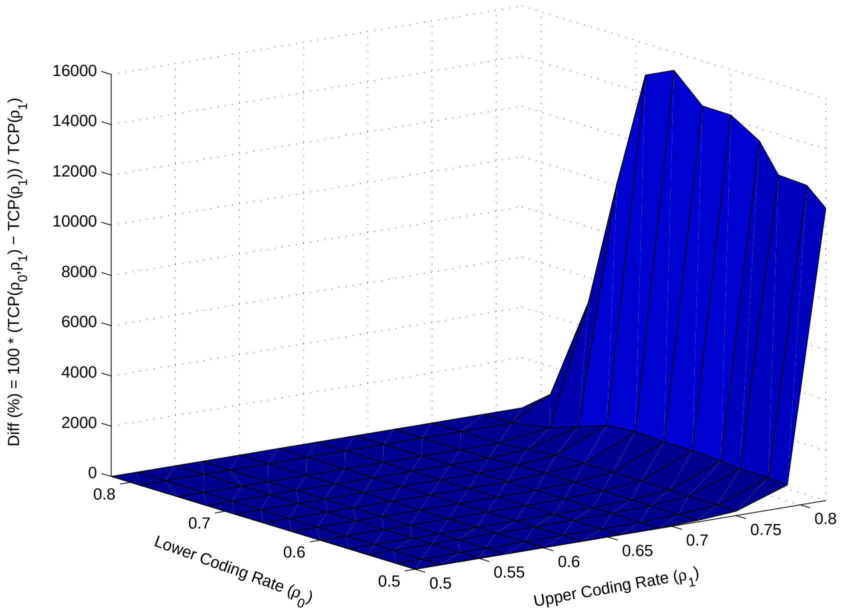
We next turn our attention to the determination of the coding rates that maximize TCP throughput. Recall that in our model for adaptive resource allocation, the scheduler switches between two modes depending on the TCP sending rate. If we assume that the allocation is controlled through the coding rate, then the scheduler is essentially switching between two coding rates ρ_0 and ρ_1.

Clearly, the particular choice of ρ_0 and ρ_1 affects the achieved throughput. Intuitively, if packet error probability is close to 0 then increasing the coding rate increases throughput as well. The reason is that, in this situation, channel rate increases as a linear function of the coding rate whereas the packet error probability remains negligible resulting in increased throughput. However, if the packet error probability is large, then increasing the coding rate decreases throughput and eventually reduces it to 0. Consequently, one expects the existence of a coding rate that maximizes TCP throughput. This behavior is clearly evident in Figure 3.5(b) where the TCP throughput initially increases as the coding rate increases, because the increase in capacity is far more than the increase in packet error probability, and then decreases. Similar curves have also been previously obtained in [61] for finite capacity and [58, 59] for infinite capacity models.

In Table 3.1, we have summarized the results from simulation and analysis for the optimal static coding rates at different bit error rates (BER). Interestingly, the optimal coding rate obtained from the analysis matches that obtained from the



(a) Deviation of single rate model with parameters (C_0, p_0, R_0).



(b) Deviation of single rate model with parameters (C_1, p_1, R_1).

Figure 3.14: Comparison between two-rate and single-rate models ($p_e = 10^{-2}$).

	Analysis		Simulation	
p_e	$\tilde{\rho}$	Throughput (Kbps)	$\tilde{\rho}$	Throughput (Kbps)
10^{-1}	0.125	10.85	0.125	8.89 ± 0.13
10^{-2}	0.660	60.98	0.660	59.20 ± 0.20
10^{-3}	0.870	81.74	0.870	78.24 ± 0.22
10^{-4}	0.910	87.32	0.910	85.21 ± 0.04
10^{-5}	0.960	92.11	0.960	92.68 ± 0.04

Table 3.1: Static coding rates ($C^* = 128$ Kbps).

	Analysis		Simulation	
p_e	$(\tilde{\rho}_0, \tilde{\rho}_1)$	Thr. (Kbps)	$(\tilde{\rho}_0, \tilde{\rho}_1)$	Thr.(Kbps)
10^{-1}	$0.125, 0.125$	10.85	$0.125, 0.125$	8.89 ± 0.11
10^{-2}	$0.630, 0.720$	65.19	$0.600, 0.660$	65.66 ± 0.12
10^{-3}	$0.820, 0.910$	85.21	$0.820, 0.870$	85.35 ± 0.06
10^{-4}	$0.910, 0.960$	91.99	$0.910, 0.960$	92.83 ± 0.15
10^{-5}	$0.960, 0.990$	94.65	$0.960, 0.990$	94.91 ± 0.27

Table 3.2: Optimal adaptive coding rates ($C^* = 128$ Kbps).

simulations. In the table, $\tilde{\rho}$ denotes the optimal coding rate.

We extend these previous results by identifying the optimal *pair of coding rates* within the framework of an adaptive TCP-aware RF scheduler. Table 3.2 presents the pair of optimal coding rates $\tilde{\rho}_0$ and $\tilde{\rho}_1$ obtained from both, the model and simulations, that maximize the TCP throughput $\overline{X}(\rho_o, \rho_1)$ and the corresponding throughput for different target bit error probabilities. Observe that in most cases, there is a close match between the optimal coding rates (and corresponding throughputs) obtained from the model with simulations.

In order to quantify the benefits of using two coding rates, Table 3.3 presents the relative increase in throughput as a consequence of using adaptive coding compared to a single coding rate for each bit error probability. The gain factor is defined as

$$\text{gain} = 100 \times \frac{\overline{X}(\tilde{\rho}_0, \tilde{\rho}_1) - \overline{X}(\tilde{\rho})}{\overline{X}(\tilde{\rho})} \tag{3.33}$$

where $\tilde{\rho}$ represents the optimum coding rate for the single rate case. Hence, for

p_e	Analysis	Simulation
10^{-1}	0%	0%
10^{-2}	6.9%	10.9%
10^{-3}	4.2%	9.1%
10^{-4}	5.3%	9.0%
10^{-5}	2.7%	2.4%

Table 3.3: Gain of adaptive coding over single coding rate.

each bit error probability, the gain factor was computed by comparing throughputs at the *optimal* coding rates for adaptive and static coding respectively. In all the cases, we see that the maximum throughput with adaptive coding is *at least* as much as that of static coding and higher in several situations, with improvements of up to 10%.

Intuitively, the reason adaptive coding yields a gain in throughput over static coding even with the same target bit error probability is that the RF scheduler with adaptive coding exploits knowledge of TCP sending rate. Specifically, when TCP has a small window, it sends at a low rate. Hence, the RF scheduler can offer a smaller rate to the source but with a lower error probability. As the rate increases, the scheduler switches to a higher rate to cope with TCP. This is not possible in the single rate case.

Furthermore, as expected, when p_e is very small or very large, *i.e.*, close to 0 or 1, the achieved throughput gain is negligible. The reason is that if p_e is close to 1 or 0, regardless of the coding rate, the packet error probability will be close to 1 or 0, respectively. It means that there is no room for optimization and, hence, adaptive coding will perform similar to static coding in these situations. However, there exist intermediate scenarios as shown in Table 3.2 where throughput gain is larger, *e.g.*, around 10% at a target bit error probability of $p_e = 10^{-2}$. The numerical results presented in this section confirm that adaptive coding is at least as good as static coding and in fact it can be better depending on the operating regime.

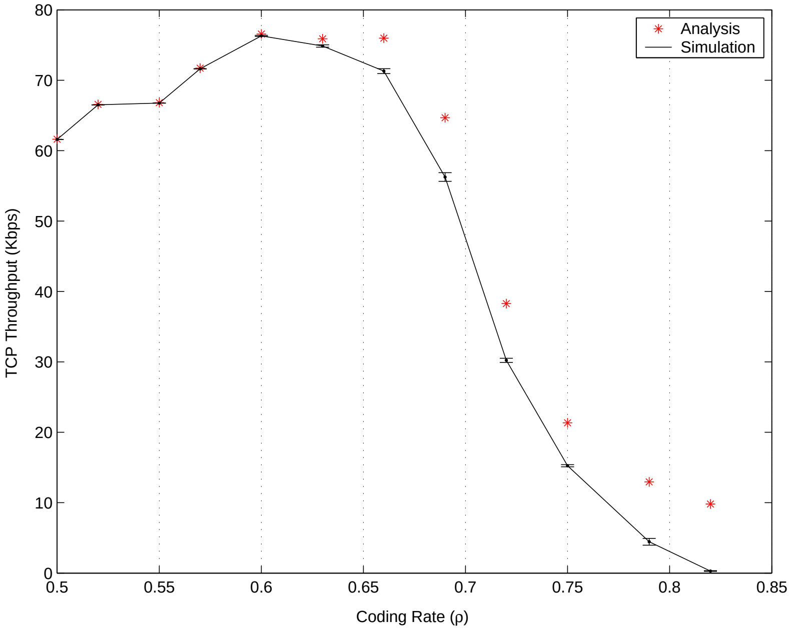
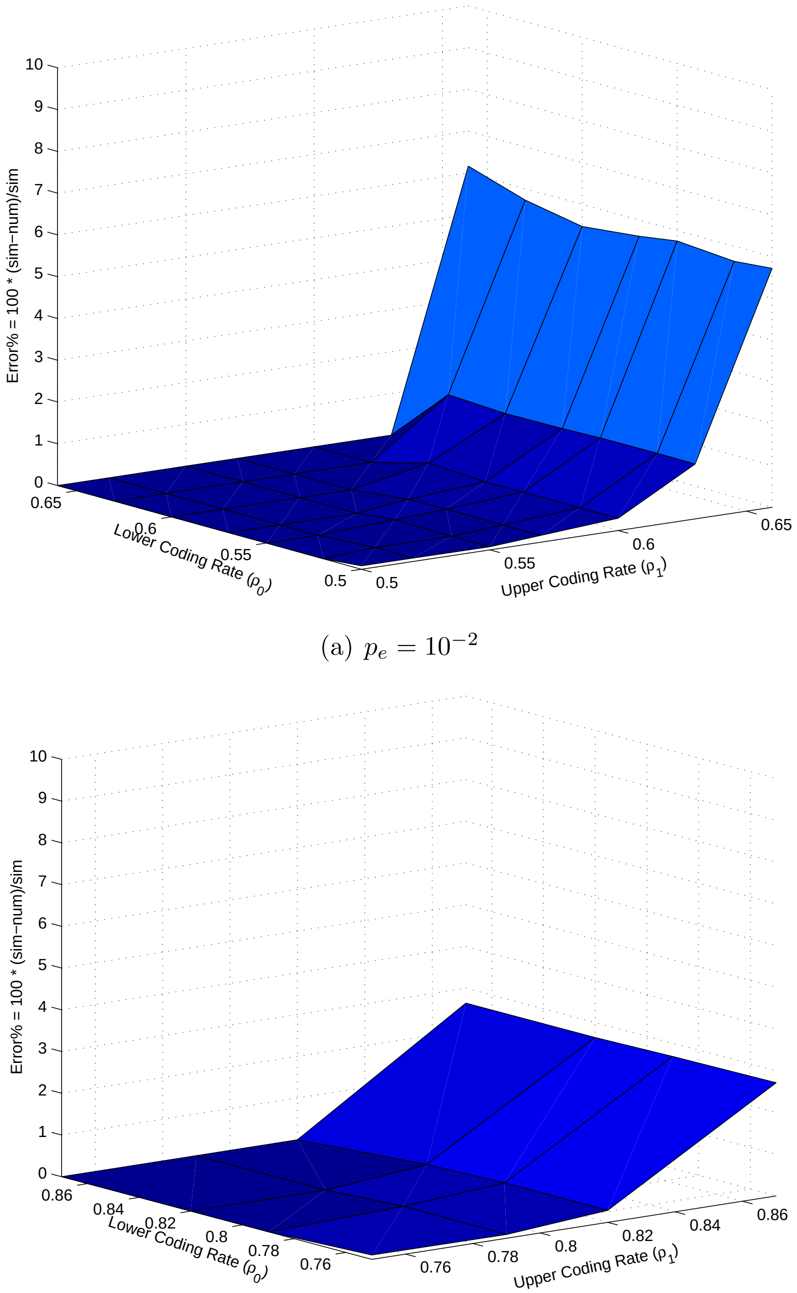


Figure 3.15: *single-rate case*: receiver window limited.

3.7.4 Adaptive Coding Evaluation: Rate Constrained

We have also evaluated the efficacy of our TCP model for the case when the sending rate is constrained by the receiver advertised window size (which translates into a maximum advertised rate). We used the same parameters as in Section 3.7.2 to compare the accuracy of our model against *ns-2* simulations. As before, we first present results for the special case $\rho_0 = \rho_1$ in Figure 3.15. From the figure, it is clear that in the region of low error probabilities, our model closely matches the simulation results.

The percentage differences between the model and simulations for the two-rate case are plotted in Figures 3.16(a) and 3.16(b) for bit error probabilities of 10^{-2} and 10^{-3} respectively. It is clear from the figures that the model is as accurate as in the rate unconstrained case, with the difference from simulation typically less than 7%.



(a) $p_e = 10^{-2}$

(b) $p_e = 10^{-3}$

Figure 3.16: *Two-rate case*: receiver window limited.

3.7.5 Rate Adaptation: Processing Gain Control

In this sub-section, we study scenarios that are more representative of current CDMA cellular networks. In such networks, the dominant RF control variable that governs this trade-off is the *spreading factor*. The spreading factor is defined as the ratio of the CDMA chip rate to the actual *data rate* and relates the channel capacity and the bit error probability in the following manner. Let W denote the *chip rate* of the CDMA system and r_i the spreading factor allocated in mode i. Then the data rate C_i and SINR are given by:

$$C_i = \frac{W}{r_i},$$
$$\frac{E_b^i}{I_0} = \frac{r_i \cdot E_c}{I_0},$$

where E_c is the *chip energy* of the data signal and I_0 the wide-spectrum interference. It should be clear from the above relations that as we decrease r_i, the channel capacity C_i *increases*, but the SINR, which directly affects the bit error probability[16], *decreases*.

Typically, the bit error probability is extremely sensitive to SINR and hence, in practice, some amount of power control is required to prevent a steep rise in the error probability when the spreading factor is decreased. Specifically, the chip energy is "boosted" by a small factor that is a function of the spreading rate to prevent a sharp drop in the quality of service. Consequently, the relationship between the SINR and the spreading factor r_i is modified slightly to:

$$\frac{E_b^i}{N_0} = \frac{E_c \cdot r_i E(r_i)}{I_0}, \tag{3.34}$$

where $E(r_i)$ is a decreasing function of r_i and represents the factor by which the original chip energy E_c is boosted. Hence, in addition to the spreading factor, one can also control the bit error probability by an appropriate choice of the function $E(r_i)$, which we shall denote as the *energy profile*.

We now demonstrate how our model can be used to study the impact of rate adaptation in such an environment. Since signal power is also a resource in the

[16] The exact relation between SINR and bit error is a function of the coding and modulation scheme.

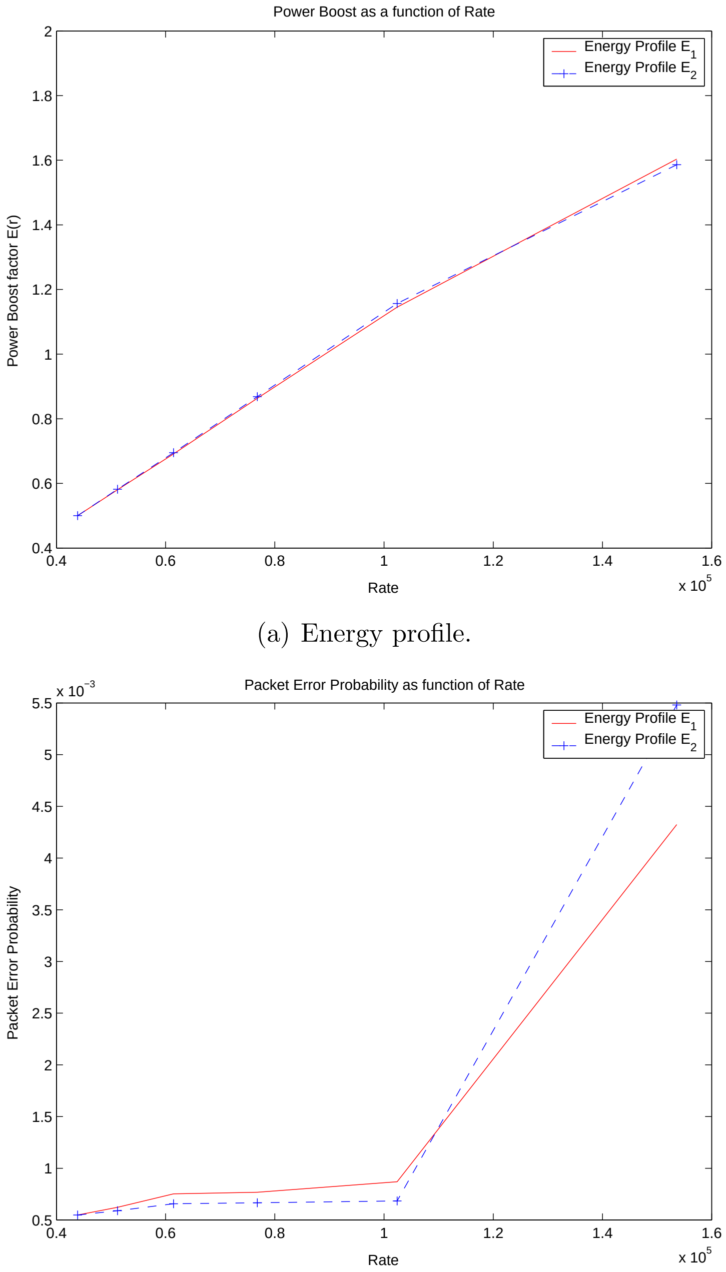
above framework, we must look at both, the mean TCP throughput as well as energy consumption. To quantify the latter, we use the notion of *normalized energy consumption.* It represents the average energy spent per bit as a function of TCP throughput and is computed as follows. Denote the energy spent per bit in mode i by $E_b^i = E_c \cdot (r_i E(r_i))$. Further, assume that the mobile session spends a fraction of time π_i in mode i and achieves an average throughput $\overline{X}_i$ in that mode. Then the normalized energy consumption is given by:

$$\begin{aligned} E_{norm} &= \frac{1}{\overline{\overline{X}}} \sum_i \pi_i \overline{X}_i \cdot E_b^i \\ &= \frac{1}{\overline{\overline{X}}} \sum_i \widehat{f}_i(2) E_b^i \\ &= E_c \frac{1}{\overline{\overline{X}}} \cdot \sum_i \widehat{f}_i(2) r_i E(r_i)\,. \end{aligned} \tag{3.35}$$

In our experiments, we set the chip rate $W = 1.2288$ Mchips/sec and the basic pilot signal quality (E_c/I_0) to (-7) dB based on CDMA2000 standards. We assume QPSK modulation (also a CDMA2000 standard) and a fixed coding rate of 0.6. To limit the complexity of the search, the spreading factor r_i was allowed to take values from the set $\{4, 12, 16\}$, which gives a rate set of $\{76.8, 102.4, 153.6\}$ Kbps. We note that 78.6 Kbps and 153.6 Kbps are the two highest data rates available in CDMA2000 1xRTT with Radio Configuration Type 3 [53].

We tested different *energy profiles*, $E(r_i)$, and present results for two of them, denoted by E_1 and E_2. Figure 3.17(a) shows the two different energy profiles while Figure 3.17(b) shows the resultant packet error probability as a function of the channel rate (in other words, the processing gain). The figures highlight the aforementioned sensitivity of error probability to SINR. Specifically, note that even though the energy profiles E_1 and E_2 are almost similar, the slight difference results in a large change in the packet error probabilities.

To give a sense of the numbers, Table 3.4 and Table 3.5 present the results for the single-rate and two-rate cases, respectively, with the energy profile E_1. The first column in each table represents the channel rates (and hence spreading factors) used. The last column presents the energy consumption for each configuration, measured in multiples of E_c. The *optimal configuration* of spreading factors, which



(a) Energy profile.

(b) Packet error probability.

Figure 3.17: Energy profile and packet error probability as function of processing gain.

	Analysis		Simulation	
C Kbps	Thr. (Kbps)	Energy ($\cdot E_c$)	Thr. (Kbps)	Energy ($\cdot E_c$)
76.8	56.95	13.89	56.78	13.81
102.4	75.10	13.87	75.81	13.75
153.6	**88.59**	**12.83**	**83.20**	**12.83**

Table 3.4: Energy profile E_1: single-rate case ($C^* = 153.6$ Kbps).

	Analysis		Simulation	
(C_0, C_1) Kbps	Thr. (Kbps)	Energy ($\cdot E_c$)	Thr. (Kbps)	Energy ($\cdot E_c$)
$76.8, 102.4$	74.55	13.77	74.91	13.77
$76.8, 153.6$	89.83	13.08	88.19	13.12
$\mathbf{102.4, 153.6}$	**98.18**	**13.26**	**96.94**	**13.28**

Table 3.5: Energy profile E_1: two-rate case ($C^* = 153.6$ Kbps).

was chosen based on the *maximum throughput*, is marked in bold for both the single and the two rate cases. Results for energy profile E_2 are shown in Tables 3.6 and 3.7.

The gain in TCP throughput and difference in energy consumption between the adaptive and static cases, under their respective *optimal configurations*, are shown in Table 3.8 for both energy profiles. The energy savings was computed as:

$$\text{Energy Savings} = 100 \times \frac{E_{norm}^{static} - E_{norm}^{adaptive}}{E_{norm}^{static}}.$$

There are a number of interesting observations regarding the results. From Table 3.8, the analytical model indicates that a small increase in energy consumption, of the order of $3\% - 4\%$, in conjunction with a two rate strategy boosts TCP throughput by $10\% - 15\%$ as compared to the single rate scenario even when the allowed *peak* channel capacities are the same. The rate gain and energy savings predictions of the simulation are comparable to those of the model for E_1 and for rate gain in E_2, but predicts higher benefits than the model for energy savings in the case of E_2. In fact, it predicts a positive energy savings of 4.5%, *i.e.*, less energy is required to achieve a higher throughput when the two-rate scheduler is used. The differences in energy prediction for E_2 are caused due to the model and simulation

	Analysis		Simulation	
C Kbps	Thr. (Kbps)	Energy ($\cdot E_c$)	Thr. (Kbps)	Energy ($\cdot E_c$)
76.8	57.032	13.892	56.867	13.892
102.4	75.472	13.878	**76.171**	**13.878**
153.6	**83.041**	**12.688**	76.021	12.688

Table 3.6: Energy profile E_2: single-rate case ($C^* = 153.6$ Kbps).

	Analysis		Simulation	
(C_0, C_1) Kbps	Thr. (Kbps)	Energy ($\cdot E_c$)	Thr. (Kbps)	Energy ($\cdot E_c$)
$76.8, 102.4$	74.819	13.884	75.234	13.884
$76.8, 153.6$	85.886	13.032	85.176	13.09
$\mathbf{102.4, 153.6}$	**96.253**	**13.29**	**95.113**	**13.32**

Table 3.7: Energy profile E_2: two-rate case ($C^* = 153.6$ Kbps).

picking different *optimal configurations* in the single rate case, while yielding the same configuration in the two rate case. Specifically the model overestimated (by about 9%) the TCP throughput in the single rate case for the highest rate configuration of 153.6 Kbps causing it to pick that as the optimal. In comparison the simulation chose the 102.4 Kbps rate which has higher energy consumption.

Energy	Analysis		Simulation	
Profile	Rate Gain	Energy Savings	Rate Gain	Energy Savings
E_1	10.8%	−3.5%	16.8%	−3.5%
E_2	15.8%	−4.8%	24.8%	4.2%

Table 3.8: Adaptive vs. static spreading factors (max. rate configuration).

3.8 Conclusion

In this chapter, we proposed a two-state TCP-aware scheduler as a means to improve TCP throughput on a wireless channel that can vary its rate dynamically, and analyzed the performance of such a system. Our proposed system adapts its channel rate in response to the TCP sending rate allowing it to trade-off channel rate and FER in a “TCP friendly” way. We developed an analytical model to study TCP throughput in such a system that captures several of its salient features, in particular the interaction between TCP and the scheduler, as well as the presence of two distinct regimes in terms of channel rate, packet error probability and round trip times. We were able to obtain analytical expressions for mean TCP throughput for the case when TCP is not constrained by the receiver window, as well as the case when it is rate limited by the receiver. The accuracy of the analysis was confirmed through comparison with *ns-2* simulations. To demonstrate the utility of our model we applied it to maximize TCP throughput in scenarios where different RF control variables are used. In particular we explored optimization of coding rates as well as the spreading factor. Throughput improvements of the order of $15\% - 25\%$ were observed as compared to previous cases where only a single rate was assumed.

In the previous chapters, we studied the impact of mobility and wireless channel from network and transport protocol perspectives. However, what is visible for users is the application performance. In next chapter, we study the performance of wireless networks from application perspective.

Chapter 4

Application Perspective: Wireless Messaging Systems

In this chapter, we study the performance of wireless networks from application perspective. As two popular wireless applications, we study short messaging service (SMS) and multimedia messaging service (MMS). We develop a mathematical model to evaluate the performance of such systems. Using the model, closed-form expressions for major performance parameters such as message loss and message delay are derived and used for dimensioning temporary storage at message centers.

4.1 Chapter Organization

The rest of this chapter is organized as follows. Section 4.2 is an introduction to the problem considered in this chapter. In Section 4.3, an overview of the multimedia messaging service is presented which covers the network architecture, operations and protocols involved in a multimedia messaging system. Section 4.5 is dedicated to the modeling and analysis of the messaging system. To investigate the accuracy of the presented analysis, simulation and analytical results are presented in Section 4.6. Finally, Section 4.7 reviews some related works and Section 4.8 concludes this chapter.

4.2 Introduction

Short Message Service (SMS) [8] is a globally accepted wireless service that allows mobile subscribers to send and receive alphanumeric messages of up to 140 bytes in length. A distinguishing characteristic of the service is the guaranteed delivery of short messages by the network via a store-and-forward mechanism. Temporary failures are identified, and the short message is stored in the network until the destination becomes available. Despite the enormous popularity of SMS, the content that can be transmitted is limited to short text messages, ring tones, and small graphics.

Due to recent developments in wireless communications, building more flexible and more capable messaging services has become the reality. The Multimedia Messaging Service (MMS), a revolutionary successor to SMS [85], has emerged as the result of research efforts primarily by the Third Generation Partnership Project (3GPP) [9] and Open Mobile Alliance (OMA) [86]. MMS will extend the revenue opportunities for network operators and manufacturers, and lead to lower costs for customers.

To the end user MMS is very similar to SMS as it provides automatic and fast delivery of multimedia messages (MMs) between capable phones and other devices. However, there are important technical differences between SMS and MMS. MMS supports richer content types such as text, graphics, music, video clips and more [87]. The MMS specifications do not mandate any specific content format for MMs. Instead the MMs are encapsulated in a standard way, so that the recipient can identify those content formats it does not support and handle them properly. The standard does not specify a maximum size for an MM either in order to avoid the SMS message size limitation.

With the increasing size and volume of messages being transmitted, the fast and robust delivery of messages becomes a challenging problem. As we will see later, the critical factor affecting the MMS system performance in terms of message delay and loss is the temporary storage of messages at server nodes. Therefore, an important problem in designing an MMS system is the proper sizing of the temporary storage in order to achieve a desirable performance. This requires the modeling of the end-to-end path which will be shown to reduce to modeling the behavior of a single

MMS server. However, a simple application of $M/M/1$ model in this context is not appropriate due to the limited patient time of queued messages.

Unfortunately, the literature on the analysis of messaging systems is quite rare. We are able only to mention the work by Haung [88] on the analysis of optimal buffer size for SMS. To the best of our knowledge, we are the first to address the message delay and loss probability of a multimedia messaging system. Indeed, the analytical method presented in this chapter is applicable to a general class of queueing systems with reneging customers and that includes SMS as well. Using the presented analysis, MMS service providers can choose the right system settings, *e.g.*, storage size, to achieve the best performance, *i.e.*, message loss and delay, for a multimedia messaging system. The main contributions of this chapter are:

1. Description of MMS and comparison with SMTP,
2. A simple approach to compute message delay which enables the derivation of closed-form expressions for both virtual and actual message delay, and,
3. A simulation study of the system sensitivity to different parameters such as load, message expiry and service rate.

4.3 The Multimedia Messaging Service

4.3.1 Network Architecture

Figure 4.1 shows a generalized view of the MMS architecture [89]. The architecture consists of different networks and integrates existing messaging systems within these networks. Mobile stations operate with the Multimedia Messaging Service Environment (MMSE) which provides all the necessary service elements, *e.g.*, delivery, storage and notification functionality under the control of a single administration. Connectivity between these different networks is provided by the Internet Protocol (IP) and its associated set of messaging protocols. This approach enables messaging in 2G and 3G wireless networks to be compatible with messaging systems found on the Internet, *i.e.*, SMTP-based email.

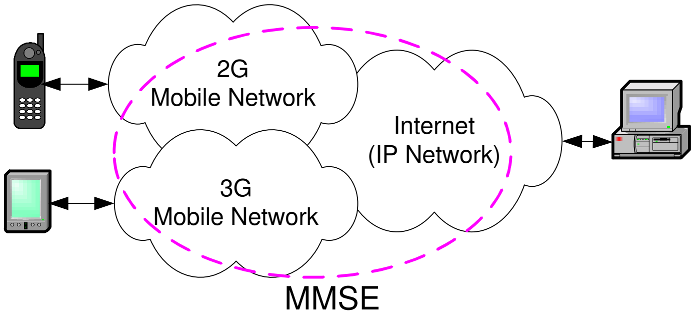


Figure 4.1: General MMS architecture integrating different networks.

Figure 4.2 shows the MMS network architecture consisting of all the elements required for providing a complete MMS to a user [89]. At the heart of this architecture is the MMS Relay/Server (MMS-RS) which is responsible for storage and reliable delivery of messages between possibly different messaging systems, akin to an SMTP mail transfer agent (MTA). The MMS-RS temporarily stores messages until they are successfully delivered. The MMS-RS may be a single logical element or may be separated into MMS Relay and MMS Server elements.

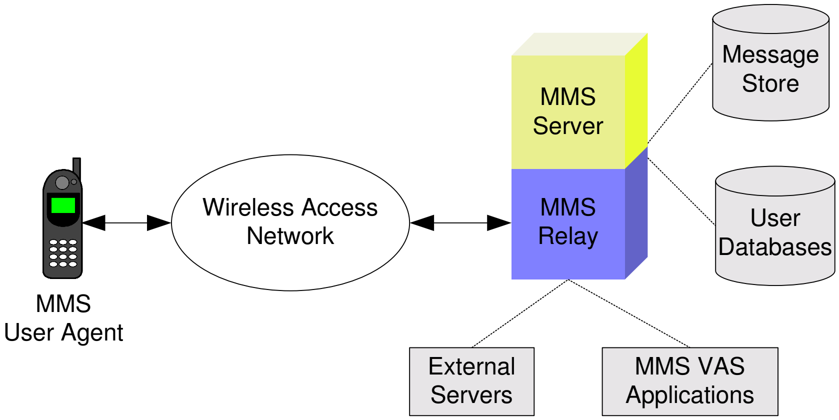


Figure 4.2: MMS network architecture.

The MMS User Agent (MMS-UA) exists within a mobile station. This application akin to an email reader application lets user view, compose and handle (*e.g.*, submit, receive, delete) multimedia messages. The retrieval of MMs from MMS-RS can be either automatic or manual. In automatic mode, an MM is retrieved without user involvement. In manual mode, the user is informed by a notification message and is allowed to make a decision whether to download the MM or not.

The MMS-RS has access to several User Databases, *e.g.*, user profile database, subscription database and home location register (HLR). An optional feature of MMS is the support of persistent network-based storage called an MMBox. The MMS-RS has access to an MMBox in order to store, retrieve or delete messages. Depending on the operator configuration, each subscriber may configure his MMBox to automatically store incoming and submitted messages, or, manually request that specific messages be persistently stored.

The MMS VAS Applications provide value added services to the MMS users. Several External Servers may be included within or connected to an MMSE, *e.g.*, E-Mail server, SMS server and Fax server. The MMS-RS is responsible for providing convergence functionality between External Servers and MMS-UAs. Thus mobile phone users can use an MMS-RS to access email, multimedia attachments, SMS or faxes.

4.4 MMS Operation

4.4.1 Transmission of Multimedia Messages

Sending Messages

A user sends a message by having its MMS-UA submit the message to its home MMS-RS. A message must have the address of the recipient and a MIME content type. Several other parameters may be set for a message including the desired time of expiry for the message and the message priority. Upon reception of a message from an originator MMS-UA, the originator MMS-RS assigns a message identification to the message and sends this message identification to the originator MMS-UA. If an MMBox is supported and enabled for the sender, MMS-RS automatically stores a copy of the message into the sender MMBox, then routes the message towards the recipients.

Receiving Messages

Upon reception of a message, the recipient MMS-RS verifies the recipient profile and generates a notification to the recipient MMS-UA. It also stores the message at least until one of the following events happens:

- the associated time of expiry is reached,
- the message is delivered,
- the recipient MMS-UA requests the message to be forwarded,
- the message is rejected.

If it has been requested, MMS-RS will also store the message in an MMBox, if the MMBox is supported and enabled.

When the recipient MMS-UA receives a notification, it uses the message reference in the notification to reject or retrieve the message, either immediately or at a later time, either manually or automatically, as determined by the operator configuration and user profile. If MMBoxes are supported, the MMS-UA may request retrieval of a message from the user MMBox, based on a message reference received from a previous MMBox operation.

Message Adaptation

Within a request for delivery of a message, the recipient MMS-UA can indicate its capabilities, *e.g.*, a list of supported media types and media formats, for the recipient MMS-RS. On getting a delivery request, the recipient MMS-RS uses the information about the capabilities of the recipient MMS-UA to prepare the message for delivery to the recipient MMS-UA. This preparation may involve the deletion or adaptation of unsupported media types and media formats [90]. Depending on the configuration and the capability of the recipient MMS-UA and the recipient MMS-RS, the MM-UA may use streaming for the retrieval of message contents.

Delivery Reports

Unlike SMPT, if a delivery report has been requested by the originator MMS-UA and if the recipient MMS-UA did not request a delivery report not to be generated, the recipient MMS-RS generates a delivery report and delivers the delivery report to the originator MMS-RS. The recipient MMS-RS stores delivery reports in the network until the originator MMS-RS becomes reachable or until the delivery report expires. A delivery report contains information such as:

- The identification of the original message for which the delivery report has been generated,
- Status information on how the message was handled (*e.g.*, expired, rejected, delivered, forwarded or indeterminate),
- A time stamp showing when the message was handled.

The originator MMS-RS, in turns, stores delivery reports until the originator MMS-UA becomes reachable or until the delivery report expires.

4.4.2 MMS Interworking

Figure 4.3 shows the message routing mechanism in MMS. From an end-to-end perspective, multimedia messages are always routed via both the sender and the recipient's home MMS-RSs. The MMS-RS provides access to the MMSE via the MM1 interface which can be implemented using WAP [91] or applications conforming to MExE [92] (*e.g.*, Java and TCP/IP) as indicated by the 3GPP specification [89]. Whenever an MMS-RS receives a message whose recipient belongs to another MMSE, the originator MMS-RS must forward the message to the recipients MMS-RS. Reference point MM4 between MMS-RSs belonging to different MMSEs is used to transfer messages between them. Interworking between MMS-RSs is based on SMTP. Resolving the destination address to find the recipient MMS-RS IP address is the responsibility of the originator MMS-RS.

The MMS-RS is also connected to External Servers such as email servers via an IP network. This connectivity works in both directions to perform three operations [86]:

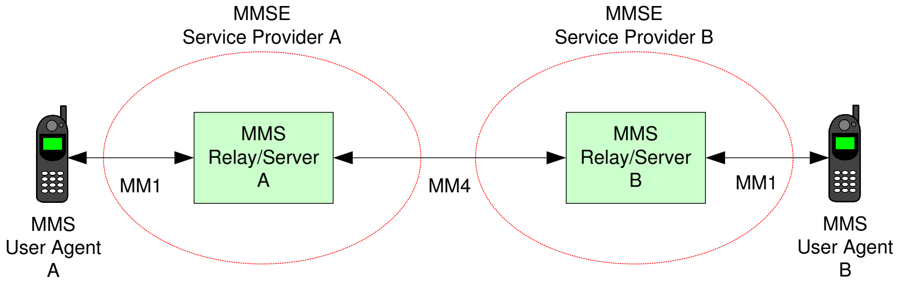


Figure 4.3: Interworking of different MMSEs.

Sending messages to email servers

After converting the message to standard Internet MIME format, the MMS-RS submits the message to the recipient using the SMTP protocol. The MMS specific header fields will be converted into appropriate headers by appending an ‘X-Mms-’ to the header name. This permits MMS-aware systems to understand the fields while not being problematic for non-MMS-aware systems.

Receiving messages from email servers

Received messages will be similarly converted. The MIME part of the message is converted to the MMS format. Similarly, any headers found with a prefix of ‘X-Mms-’ can be converted back to the associated MMS header.

Retrieving messages from email servers

This is normally done through the use of the POP or IMAP protocols. Such retrievals are performed by the MMS-RS, which will then convert the data into an appropriate MMS format.

4.4.3 MMS Addressing

MMS supports the use of email address or MSISDN (E.164) or both to address the recipient of a message. Since MMS interworking across different networks (MMSEs) is provided based on SMTP, each MMSE is assigned a unique DNS domain name.

Addressing at the MM1 Interface

The message addressing on MM1 consists of three addresses: the address of the originator MMS-RS, the address of the recipient and the address of the sender. The address of the originator MMS-RS is the Uniform Resource Identifier (URI) of the MMS-RS given by the service provider. The sender address could be either a user address or a terminal address (*e.g.*, terminal IP addresses). The recipient address can be a user address, a terminal address, or a short code. The user address can be either an MSISDN or email address.

Addressing at the MM4 Interface

For those recipients that appear in a message and belong to an external MMSE, the originator MMS-RS has to send the message to each of the recipients MMS-RS. The MMS-RS has to resolve the recipient MMS-RS domain name to an IP address based on the recipient address. In case of MSISDN addressing, the originator MMS-RS should translate the address to an email address using DNS-ENUM protocol [93].

4.5 Multimedia Messaging Service Analysis

In this section, we analyze the message handling performance of an MMS-RS in isolation as the central element in a multimedia messaging system. We will later discuss end-to-end system performance. The performance parameters of interest are the queueing delay induced by the temporary storage of messages in the MMS-RS, and message loss probability due to storage overflow and message expiration while messages are temporarily waiting in storage. Indeed, the messaging system performance is dramatically affected by the size of the temporary storage. Having a small storage avoids long message delay. However, if the storage is too small then message loss due to storage overflow increases and the MMS-RS utilization deteriorates. Thus, it is crucial to determine an optimal storage size to achieve maximum utilization and minimum message loss probability, and thus prevent messages from being excessively delayed, which is especially needed for multimedia messages.

A conceptual model of the MMS-RS is shown in Figure 4.4, where the MMS-RS is modeled as a queueing system. The assumptions and parameters involved in this

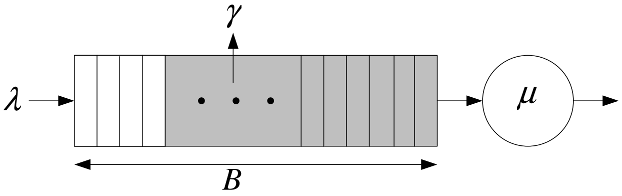


Figure 4.4: A queueing model of the MMS-RS.

model are stated below:

1. The new message arrival into the MMS-RS is Poisson distributed with rate λ.

2. The timeout period of the temporary stored messages in the MMS-RS is assumed to be exponentially distributed with mean $1/\gamma$. A stored message is removed from the temporary storage if it can not be transmitted within its timeout period. A message expiry that happens during the actual transmission of the message is ignored by the MMS-RS.

3. The transmission time of a message is assumed to be exponentially distributed with mean $1/\mu$. Stored messages are served according to FIFO scheduling.

4. A finite storage with capacity $B > 0$ messages is provided in the MMS-RS for temporary message storage.

In the real world, the message expiry and transmission times may not be exponential but exponential distributions allow mean value analysis, which indicates the performance trend of the system. The goal of this work is to develop a tractable and yet reasonably accurate model rather than trying to apply exact but intractable models that do not necessarily capture all the impact of complex system interactions. We believe our model is rich and analyzable enough to provide information that is practically important for MMS service providers.

Performance analysis of the MMS-RS can be accomplished by describing the system as a Markov chain corresponding to the system dynamics. Figure 4.5 shows the transitions among different system states where state i indicates that there are

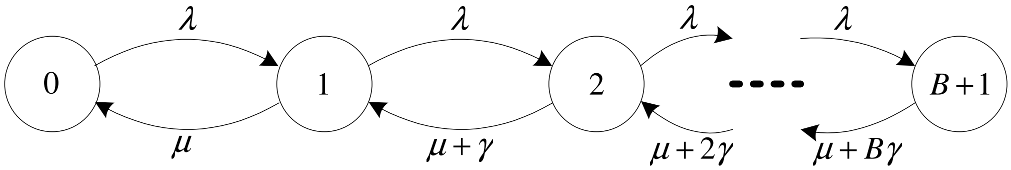


Figure 4.5: A Markov chain representation of the MMS-RS.

i messages in the system (either in temporary storage waiting for delivery or being transmitted). Let p_i denote the steady-state probability of being in state i. Using balance equations, it is clear that

$$p_i = \prod_{j=1}^{i} \left(\frac{\lambda}{\mu + (i-1)\gamma} \right) p_0, \qquad 1 \le i \le B+1 \tag{4.1}$$

where p_0 can be found using the normalizing condition $\sum_{i=0}^{B+1} p_i = 1$. In this model, the service rate at state i is $\mu + (i-1)\gamma$.

Throughout this chapter we will use the steady-state probability distribution of the system state as seen by an arriving message, *i.e.*, system state at arrival epochs. Let a_i denote the probability of having i messages in the system just before a message arrives to the system and gets accepted by the MMS-RS. Such a message never sees the system in state $(B+1)$, *i.e.*, blocking state. It can be shown that a_i is given by (for example refer to [94])

$$a_i = \frac{p_i}{1 - p_{B+1}}, \qquad 0 \le i \le B\,. \tag{4.2}$$

4.5.1 Message Delay Analysis

Define the message delay as the time between the acceptance of a message in the MMS-RS and the time its transmission starts. We are interested in finding the *actual message delay* denoted by W, and defined as the message delay experienced by a message that is successfully transmitted, *i.e.*, did not expired before transmission. In the following discussion, we use notations $F_z(t)$ and $f_z(t)$ to denote the distribution and density function of a random variable z, where

$$f_z(t) = \frac{d}{dt} F_z(t) \,. \tag{4.3}$$

Let random variable T denote the *virtual message delay* defined as the message delay experienced by a message which has infinite expiry time. Such a message is referred to as a *virtual message* and remains in the system until it is transmitted. T will, in ..., be larger than W because T does not take the finite expiry time of messages into account. Using the conditional probability formulation, the relation between the actual message delay, W, and the virtual message delay, T, can be expressed as

$$F_W(t) = \mathbb{P}\left\{W \le t\right\} = \mathbb{P}\left\{T \le t \,|\, T \le X\right\}, \tag{4.4}$$

where X is a random variable denoting the expiry time of a typical message. We assume that message expiry times are independent and exponentially distributed with the parameter γ, hence

$$f_X(t) = \gamma e^{-\gamma t} \,. \tag{4.5}$$

Assume that a virtual message arrives to the system when there are i messages in the system, *i.e.*, system is in state i. Let m_i denote this message. If $0 \le i \le B$ then m_i is accepted and the system state will change to $i+1$. If $i = 0$ then m_i will be immediately served, otherwise it must temporarily wait in the storage for i message departures (either transmission or expiry). Suppose we temporarily view the system as consisting of those i messages which are ahead of m_i. Let t_j denote the time required for the message population to decrease from j to $j-1$ ($1 \le j \le B$). Then, t_j is exponentially distributed with rate parameter $\mu + (j-1)\gamma$, *i.e.*,

$$f_{t_j}(t) = [\mu + (j-1)\gamma] e^{-[\mu+(j-1)\gamma]t} \,. \tag{4.6}$$

Let T_i denote the virtual message delay of m_i, *i.e.*, the amount of time m_i must wait before its transmission starts given that m_i is infinitely patient. Then

$$\begin{aligned} T_i &= t_1 + \cdots + t_i, \\ &= T_{i-1} + t_i, \qquad \text{for } i \ge 1 \,. \end{aligned} \tag{4.7}$$

According to our definition of virtual message delay, it is clear that $T_0 = 0$. Therefore, we have

$$f_{T_i}(t) = \begin{cases} \mu e^{-\mu t}, & i = 1 \\ f_{T_{i-1}} * f_{t_i}(t), & i > 1 \end{cases} \tag{4.8}$$

where $f_{T_{i-1}} * f_{t_i}(t)$ is the convolution of $f_{T_{i-1}}(t)$ and $f_{t_i}(t)$ which can be expressed as

$$f_{T_{i-1}} * f_{t_i}(t) = \int_0^t f_{T_{i-1}}(x) f_{t_i}(t-x)\, dx\,. \tag{4.9}$$

By solving the recursive definition (4.8) using (4.6) and (4.9), we find that

$$f_{T_i}(t) = \mu \prod_{j=1}^{i-1} \left(\frac{\mu + j\gamma}{j\gamma} \right) \left(1 - e^{-\gamma t}\right)^{i-1} e^{-\mu t}\,. \tag{4.10}$$

Using (4.7), the mean virtual message delay, $\mathbb{E}\left[T_i\right]$, is given by

$$\mathbb{E}\left[T_i\right] = \sum_{j=1}^{i} \mathbb{E}\left[t_j\right] = \sum_{j=1}^{i} \frac{1}{\mu + (j-1)\gamma}\,. \tag{4.11}$$

Having obtained the virtual message delay distribution, we now turn our attention to actual message delay. Let m_i denote a message that is accepted in the system in state i. Let β_i denote the probability that m_i does not expire before it is being transmitted, *i.e.*,

$$\beta_i = \mathbb{P}\left\{T_i \leq X\right\}\,. \tag{4.12}$$

Let β_{ij} denote the probability that m_i does not expire during t_j, the interval of time required to drive the message population from j to $j-1$. Then, β_{ij} can be obtained as follows

$$\beta_{ij} = \mathbb{P}\left\{t_j \leq X\right\} = \frac{\mu + (j-1)\gamma}{\mu + j\gamma}\,. \tag{4.13}$$

Therefore, β_i is given by

$$\beta_i = \prod_{j=1}^{i} \beta_{ij} = \frac{\mu}{\mu + i\gamma}\,. \tag{4.14}$$

Using the conditional probability given in (4.4), $F_{W_i}(t)$ is expressed as

$$\begin{aligned} F_{W_i}(t) &= \mathbb{P}\left\{T_i \leq t \,|\, T_i \leq X\right\} \\ &= \frac{\mathbb{P}\left\{T_i \leq t,\, T_i \leq X\right\}}{\mathbb{P}\left\{T_i \leq X\right\}}\,. \end{aligned} \tag{4.15}$$

From (4.12), we get

$$F_{W_i}(t) = \frac{1}{\beta_i}\int_0^t f_{T_i}(x)(1 - F_X(x))\,dx\,. \tag{4.16}$$

Equivalently,

$$\begin{aligned} f_{W_i}(t) &= \left(\frac{\mu + i\gamma}{\mu}\right) f_{T_i}(t)(1 - F_X(t)) \\ &= (\mu + i\gamma)\prod_{j=1}^{i-1}\left(\frac{\mu + j\gamma}{j\gamma}\right)\left(1 - e^{-\gamma t}\right)^{i-1} e^{-(\mu+\gamma)t}, \end{aligned} \tag{4.17}$$

where we substitute $f_{T_i}(t)$ from (4.10). From the binomial expansion of $\left(1 - e^{-\gamma t}\right)^{i-1}$ we obtain that

$$\left(1 - e^{-\gamma t}\right)^{i-1} = \sum_{j=1}^{i-1}\binom{i-1}{j-1}(-e^{-\gamma t})^{j-1}\,. \tag{4.18}$$

Using (4.17), the mean actual message delay, $\mathbb{E}\left[W_i\right]$, is given by

$$\begin{aligned} \mathbb{E}\left[W_i\right] &= \int_0^\infty t f_{W_i}(t)\,dt \\ &= (i\gamma)\left[\prod_{j=1}^{i}\left(\frac{\mu + j\gamma}{j\gamma}\right)\right]\left[\sum_{j=1}^{i}\binom{i-1}{j-1}\frac{(-1)^{j-1}}{(\mu + j\gamma)^2}\right]\,. \end{aligned} \tag{4.19}$$

We further can simplify (4.19) by comparing (4.10) and (4.17). Replace μ by $\mu + \gamma$ in (4.10) and (4.11). It is obtained that

$$(\mu + \gamma)\prod_{j=1}^{i-1}\left(\frac{\mu + (j+1)\gamma}{j\gamma}\right)\int_0^\infty t\left(1 - e^{-\gamma t}\right)^{i-1} e^{-(\mu+\gamma)t}\,dt = \sum_{j=1}^{i}\frac{1}{\mu + j\gamma}\,. \tag{4.20}$$

Therefore,

$$\begin{aligned}
\mathbb{E}\left[W_i\right] &= \int_0^{\infty} t f_{W_i}(t)\, dt \\
&= (\mu + i\gamma) \prod_{j=1}^{i-1} \left(\frac{\mu + j\gamma}{j\gamma}\right) \int_0^{\infty} t \left(1 - e^{-\gamma t}\right)^{i-1} e^{-(\mu+\gamma)t}\, dt \\
&= (\mu + i\gamma) \prod_{j=1}^{i-1} \left(\frac{\mu + j\gamma}{j\gamma}\right) \left[(\mu + \gamma) \prod_{j=1}^{i-1} \left(\frac{\mu + (j+1)\gamma}{j\gamma}\right)\right]^{-1} \sum_{j=1}^{i} \frac{1}{\mu + j\gamma} \\
&= \sum_{j=1}^{i} \frac{1}{\mu + j\gamma}\,.
\end{aligned} \tag{4.21}$$

Furthermore, the steady-state performance parameters can be computed with respect to the steady-state probability distributions given by (4.2). In particular, the average steady-state actual message delay, W, is expressed as

$$W = \sum_{i=1}^{B} q_i W_i, \tag{4.22}$$

where q_i is the steady-state probability that a non-reneging message finds i messages in the system upon arrival. A non-reneging message is a message that will receive service before expiration. Using the Bayes's theorem, q_i can be determined as follows:

$$\begin{aligned}
q_i(t) &= \lim_{\delta \to 0} \mathbb{P}\left\{N(t) = i \,|\, \text{a NRA at } t + \delta\right\} \\
&= \lim_{\delta \to 0} \frac{\mathbb{P}\left\{N(t) = i\right\} \mathbb{P}\left\{\text{a NRA at } t + \delta \,|\, N(t) = i\right\}}{\sum_{j=0}^{B} \mathbb{P}\left\{N(t) = j\right\} \mathbb{P}\left\{\text{a NRA at } t + \delta \,|\, N(t) = j\right\}} \\
&= \lim_{\delta \to 0} \frac{p_i(t)\beta_i\lambda\delta}{\sum_{j=0}^{B} p_j(t)\beta_j\lambda\delta} = \frac{p_i(t)\beta_i}{\sum_{j=0}^{B} p_j(t)\beta_j},
\end{aligned} \tag{4.23}$$

where NRA stands for "non-reneging arrival". Therefore,

$$q_i = \lim_{t \to \infty} q_i(t) = \frac{p_i\beta_i}{\sum_{j=0}^{B} p_j\beta_j} = \frac{\lambda/\mu}{1 - p_0}\beta_i p_i\,. \tag{4.24}$$

Finally, our results can be represented in terms of special functions by noting that

$$\prod_{j=1}^{i}(\mu + j\gamma) = \frac{\Gamma(1+\mu/\gamma+i)}{\Gamma(1+\mu/\gamma)}\gamma^{i}, \tag{4.25}$$

$$\sum_{j=1}^{i}\frac{1}{(\mu + j\gamma)} = \frac{1}{\gamma}\Psi(1+\mu/\gamma+i) - \frac{1}{\gamma}\Psi(1+\mu/\gamma), \tag{4.26}$$

where, Γ and Ψ denote the gamma and digamma function [95] respectively.

4.5.2 Message Loss Probability

A message is lost if upon its arrival to the MMS-RS, the temporary storage is full, or, although accepted and waiting in the storage, fails to be transmitted within its timeout period and so is removed from the storage. Therefore, the message loss probability L is given by

$$L = p_{B+1} + (1 - p_{B+1})\alpha, \tag{4.27}$$

where, α denotes the probability that a typical message accepted in the system will expire before being transmitted and is given by

$$\alpha = \sum_{i=0}^{B}(1-\beta_i)a_i = \frac{N}{1-p_{B+1}}\left(\frac{\gamma}{\lambda}\right), \tag{4.28}$$

where, N is the average number of messages in storage.

4.5.3 Average Number of Queued Messages

We are interested to find a closed-form expression for the average number of queued messages in the temporary storage. Let β denote the probability that a typical accepted message will be transmitted. Then, β is expressed as

$$\beta = \sum_{i=0}^{B}a_i\beta_i = \frac{\lambda}{\mu}\left(\frac{1-p_0}{1-p_{B+1}}\right). \tag{4.29}$$

Given that $\alpha + \beta = 1$, it is obtained that

$$N = \frac{1}{\gamma}\Big[\lambda(1-p_{B+1}) - \mu(1-p_0)\Big]. \tag{4.30}$$

4.5.4 End-to-End Performance

Message queueing may happen in three different locations before a message reaches it destination: 1) inside the originator MMS-UA, 2) in the originator MMS-RS, or 3) in the recipient MMS-RS. In previous subsections, the message delay and loss probability were analyzed for the recipient MMS-RS. In a typical interworking scenario, two MMS-RSs, namely the originator and the recipient, are involved. Given that for the originator MMS-RS, the incoming wireless link is the bottleneck not the outgoing wireline link, it can be assumed that the message loss and delay due to temporary storage in the originator MMS-RS are effectively zero. Considering the user behavior in generating multimedia messages, no queueing is expected to happen inside the user equipment, *i.e.*, MMS-UA, either. Therefore, message loss in MMS-UA is also zero assuming that message expiry times are comparable to message transmission times (otherwise all the messages expire before entering the MMS system!).

Consequently, the end-to-end message loss probability is determined by the loss probability at the recipient MMS-RS. However, for computing the end-to-end message delay, the message transmission time in the originator MMS-UA must be considered. The impact of such a delay may be well captured by the fact that we assumed message expiry time is a random variable (with exponential distribution) not a deterministic value. Therefore, to have more accurate analysis, the impact of message transmission time in the originator MMS-UA must be included in the expiry time characterization and must be added to the end-to-end message delay as well.

4.6 Simulation Results

An event-driven simulation was developed to verify the correctness of the analysis. The simulation considered a single MMS-RS in isolation. All the simulation parameters are relative to the message transmission rate μ. For the basic set of simulations, the average message size is considered to be 8 KBytes and the output link is a GPRS carrier transmitting at 10 Kbps. Simulation results are obtained by averaging over 8 independent samples. Only average points are plotted since the

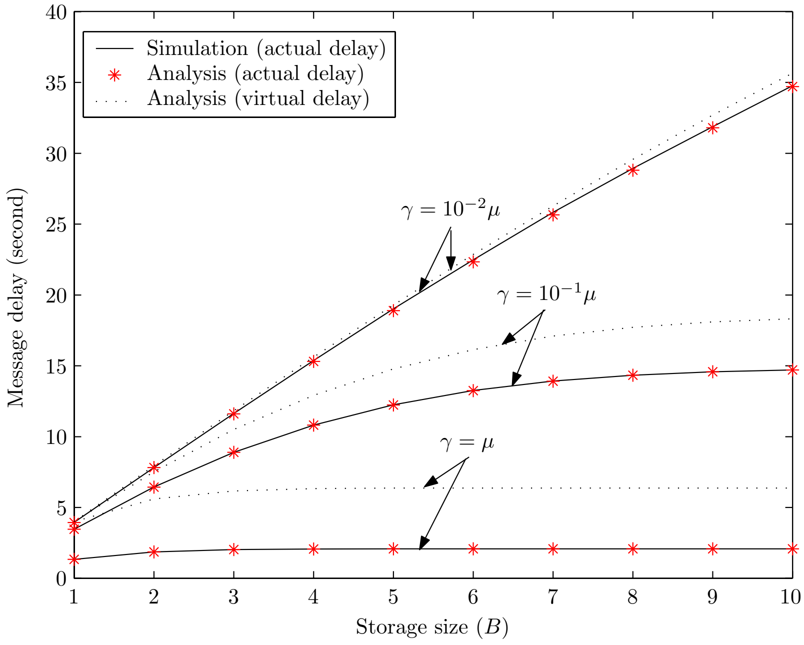


Figure 4.6: Message delay for $\lambda = \mu$.

95% confidence intervals were very close to the average value, and hence are not shown for the sake of having clear plots. The simulation length was long enough to avoid rare event problem for the simulation scenarios with small values of γ. Overall, in all the cases simulated, analytic and simulation results match with almost no discrepancy.

Effect of the expiry time

Figures 4.6 and 4.7 show the average message delay and message loss probability for different message expiry rates when the offered load is set to 1, *i.e.*, $\lambda = \mu$. As expected, message loss increases by increasing the expiry rate but message delay decreases by increasing the expiry rate. In addition to the actual message delay, the virtual message delay computed using (4.11) is depicted in Figure 4.6 as well. As the message expiry rate increases the difference between actual and virtual message delay increases too. It is observed from Figure 4.7 that as the expiry rate increases the message loss probability increases despite the fact that blocking probability has decreased. Referring to (4.27), the message loss probability depends on both

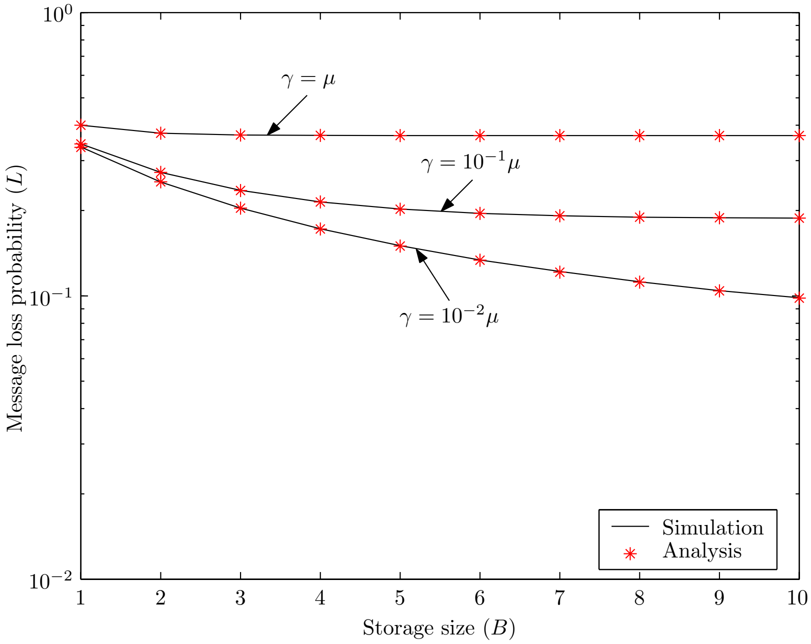


Figure 4.7: Message loss probability for $\lambda = \mu$.

blocking and expiry probability. Although, the blocking probability decreases by increasing the expiry rate, this decrease is not proportional to the increase in expiry probability. More formally, by substituting (4.30) in (4.28) and using (4.27), it can be shown that

$$\begin{aligned} L &= p_{B+1} + (1 - p_{B+1}) \left[1 - \left(\frac{\mu}{\lambda} \right) \left(\frac{1 - p_0}{1 - p_{B+1}} \right) \right], \\ &= 1 - \left(\frac{\mu}{\lambda} \right) (1 - p_0) \, . \end{aligned} \tag{4.31}$$

In this case, since $\lambda = \mu$, it is obtained that $L = p_0$. Increasing the message expiry rate results in larger p_0 which consequently means larger message loss probability.

Effect of the offered load

To investigate the effect of the offered load on the system performance, we ran the simulations for the case of having $\gamma = 10^{-1}\mu$ with different arrival rates. Three arrival rates $\lambda = 2\mu$, $\lambda = \mu$ and $\lambda = \frac{1}{2}\mu$ corresponding to loads 2, 1 and $\frac{1}{2}$, respectively, were simulated. All other system parameters are the same as before.

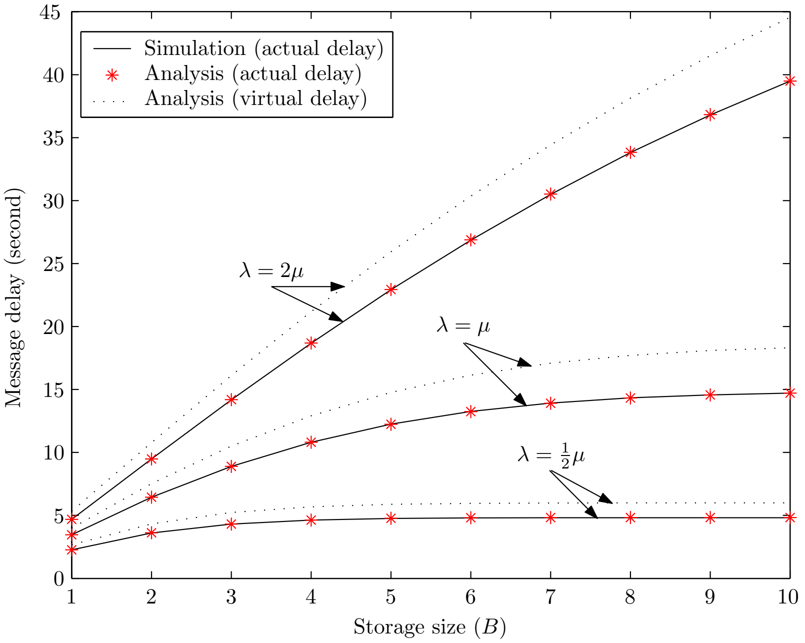


Figure 4.8: Message delay for $\gamma = 10^{-1}\mu$.

Figures 4.8 and 4.9 show the message delay and message loss probability with respect to the storage size.

As expected, both message delay and loss probability grow by increasing the offered load. As shown in the figures, the loss probability is rather insensitive to the storage size specially for high loss rates (*e.g.*, $\lambda = 2\mu$).

Effect of the service time

A GPRS carrier with transmission rate 10 Kbps is perhaps too slow to be a good candidate for building a practical MMS system. To investigate the impact of service rate, equivalently the message service time, on the system performance, we did the simulations for different service rates. Figure 4.10 show the message delay for four different service rates in terms of basic service rate $c_1 = 10$ Kbps with respect to the storage size. In these simulations, $\lambda = \mu$ and $\gamma = 10^{-1}\mu$. As the service rate increases, the message transmission time ($1/\mu$) decreases and consequently, message delay decreases too. It can be seen from the figure that as the service rate increases, the discrepancy between virtual and actual message delay decreases which can be

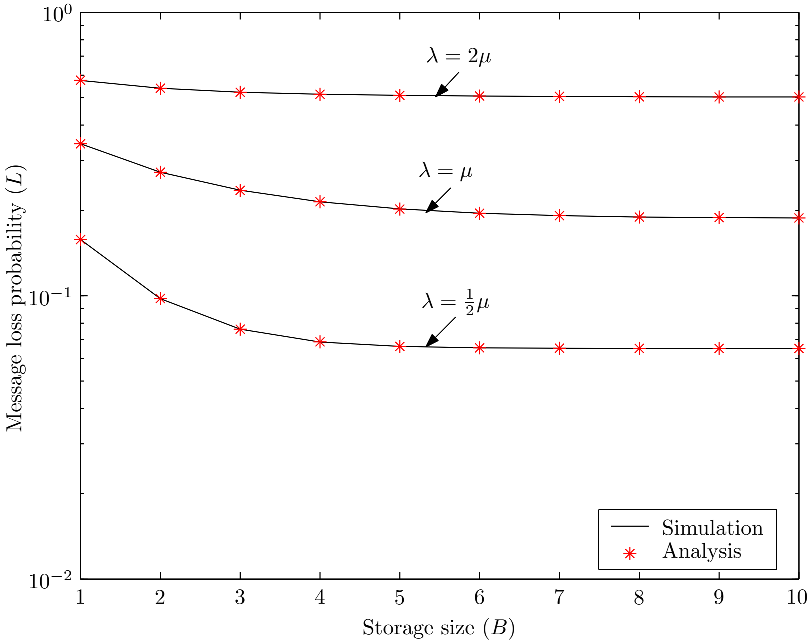


Figure 4.9: Message loss probability for $\gamma = 10^{-1}\mu$.

also verified from the analysis. Changing the service rate does not affect the loss probability as the loss probability is a function of quantities λ/μ and γ/μ which are fixed in this case.

Optimal storage size

As shown in Figures 4.7 and 4.9, message loss probability decreases by increasing the storage size up to a certain threshold. Increasing the storage size beyond that threshold will not significantly change the loss probability. However, if the storage size exceeds the threshold, message delay will continue to increase as depicted in Figures 4.6, 4.8 and 4.10. This threshold is referred to as the *optimal storage size*. An iterative approach similar to the one proposed in [88] can be used to find the optimal buffer size for a given system configuration. The iterative algorithm follows the pseudo-code represented in Figure 4.11, where ϵ is the desired precision for the convergence of message loss probability.

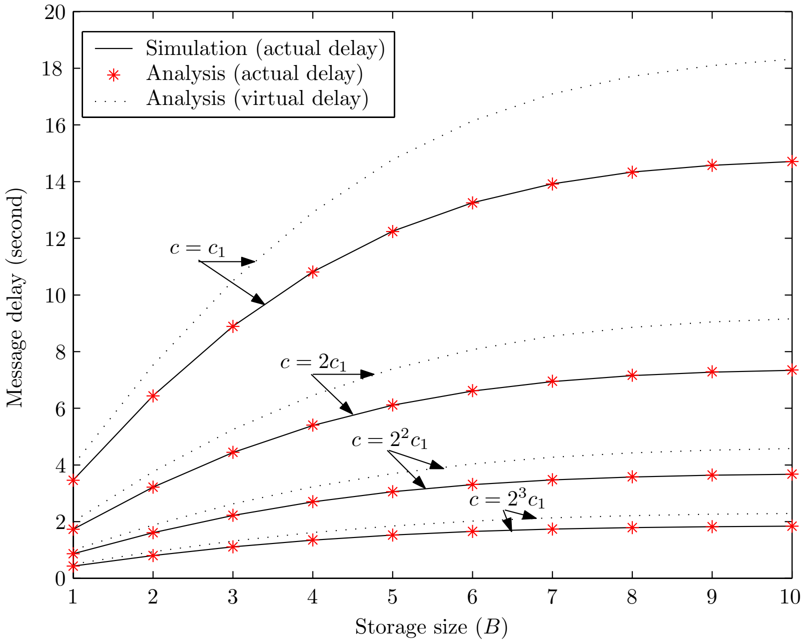


Figure 4.10: Message delay for $\lambda = \mu$ and $\gamma = 10^{-1}\mu$.

4.7 Related Work

In queueing theory terminology, this type of system is usually referred to as queue with impatient or reneging customers. The literature on queueing systems with reneging is moderate. This includes classical works such as [96–98] and recent works such as [99–102]. Among them, Barrer [96] obtained the reneging probability for deterministic patient time customers. Baccelli and Hebuterne [97] considered a queue with general patient time distribution. However, their analysis involves inverse Laplace transformations which does not provide a closed-form expression for the performance parameters. Queueing systems with state-dependent arrival/service rate have been studied in [99, 100]. References [94, 98] mostly focused on steady-state probability distributions of the queue length with reneging customers. To avoid complexity and numerical instability associated with exact analytical techniques, approximate solutions have been also investigated [101, 102].

Although queueing systems with impatient customers have been studied by several researchers, our contribution is a simpler convolution-based technique to

```
$B \leftarrow 1$;
$L \leftarrow 0$;
**repeat**
    $L' \leftarrow L$;
    $L \leftarrow$ Message loss using (4.27);
    $B \leftarrow B + 1$;
**until** $(|L - L'| > \epsilon)$;
```

Figure 4.11: Iterative algorithm for computing B.

find closed-form expressions for both the virtual and actual message delay for the Markovian system depicted in Figure 4.5. This technique avoids complicated transformations and differential equations applied in previous works by formulating the virtual message delay as a recursive convolution tailored to the MMS-RS model.

4.8 Conclusion

This chapter studied the architecture, operation and performance of a multimedia messaging system. Various components involved in the architecture and their functionalities were described. Then, a mathematical model developed to study the performance of the MMS-RS as the central component of a multimedia messaging system. The presented analysis models an MMS-RS as a finite capacity queueing system with reneging customers (multimedia messages). Using the Markovian behavior of the system, closed-form expressions describing the message delay distribution and message loss probability were presented. The analytical results were compared with those obtained from the simulation which confirmed the accuracy of the analysis. Although a Markovian system was considered in this chapter, in practice the message size and expiry time distributions may not be exponential. However, exponential distributions enable the derivation of closed-form expressions for various performance parameters which provide insight into the system behavior.

Chapter 5

Summary and Future Research

5.1 Summary

In this thesis, we studied the impact of mobility and wireless channel characteristics, *i.e.*, variability and high bit-error-rate, on the performance of wireless systems from network, transport protocol and application perspectives. A summary of the work presented in the thesis is as follows.

From network perspective, we studied the impact of mobility on call-level performance of integrated voice and data networks. We developed a distributed call admission control scheme that guarantees a pre-specified call dropping probability for voice calls while being fair to data calls. We modeled the number of calls in each cell of the network as a Gaussian process with time-dependent mean and variance. The Gaussian model is updated periodically using the information obtained from neighboring cells about their load condition. We implemented the proposed admission control algorithm in a simulated system of 19 cells. Simulation results showed that our algorithm satisfies the hard constraints on voice dropping and data blocking probabilities while maintaining a high bandwidth utilization.

From transport protocol perspective, we studied the impact of wireless channel and downlink scheduler on the performance of TCP in CDMA 1xRTT systems. We proposed a two-state TCP-aware scheduler that switches between two rates as a function of the TCP sending rate. We developed a fluid model of the steady-state TCP behavior for such a system and derived analytical expressions for TCP throughput that explicitly account for rate variability as well as the dependency between the scheduler and TCP. The model captures variations in round-trip-time, channel rate and packet error probability induced by the scheduler. Using *ns-2* simulations we showed that our model is indeed a close approximation of TCP throughput in CDMA 1xRTT systems. We then used the model to choose optimal RF layer parameters that maximize long-term TCP throughput in wireless networks.

From the application perspective, we studied the performance of wireless messaging systems. As two popular wireless applications, we studied short messaging service and multimedia messaging service. We developed a mathematical model to evaluate the performance of such systems. Using the model, we derived closed-form expressions for message loss and message delay. We also proposed an algorithm for dimensioning buffer sizes at messaging centers in order to minimize message delay.

5.2 Future Research

This section summarizes possible extensions of the research presented in this thesis.

5.2.1 Call Admission Control

- *Multiple services:* We addressed the admission control problem in integrated voice and data networks in Chapter 2. However, we assumed that there is only one type of data traffic (elastic traffic) and treated all data calls similarly. Modern cellular networks provide multiple services, *e.g.*, voice, data and multimedia, through IP Multimedia Subsystem (IMS) [103]. Different services demand different bandwidth and require different quality of service in terms of call dropping and blocking probabilities. The proposed CAC algorithm (EFGC) can readily support multiple classes of service by assigning a separate acceptance ratio to each class. However, computing these acceptance ratios in order to satisfy the desired quality of service is not trivial.

- *Elastic traffic:* Elastic data traffic is not sensitive to temporary interruptions. This property can be used to further reduce call dropping probability for voice calls. If a cell is completely full upon receiving a handoff request, then a data call can be terminated to accommodate a handoff voice call. It is even possible to reduce the allocated bandwidth to each data call in order to provide room for the handoff call. In the case of TCP flows, this can be achieved, for example, by explicitly regulating ACKs or congestion window size as proposed in [73, 104].

5.2.2 TCP Analysis in Wireless Networks

- *Multiple channel rates:* A natural extension of the system considered in Chapter 3 is to extend the results to more than two channel rates. Current cellular networks allow at least four transmission rates. However, we found that it is extremely difficult to analyze the system in this case. Using simulations is a promising approach to study TCP throughput in this case.

- *Multiple users:* It is possible to extend the single user model of Chapter 3 to multi-user systems where the channel is shared by multiple TCP sessions. In such systems we expect that optimization of the rate adaptation feature will have large multiplicative impact, translating small throughput improvements into large network capacity gains. Specifically, an improvement of 15% in the throughput of a single session may translate into the capability to support a much larger number of users in a large network.

- *Scheduling effect:* High-speed wireless data services such as CDMA EV-DO employ complicated scheduling policies to maximize radio resource utilization. The Proportional Fair scheduler of EV-DO systems [105] chooses the user that has the highest ratio of achievable rate over the average received rate. This means that a user may not be able to send and receive any data for some period of time. If this period lasts longer than TCP time-out timer, then a time-out happens that causes TCP to reduce its window size to just 1 packet. Indeed, measurements have shown that TCP suffers from frequent time-outs in EV-DO systems. We believe that the proposed TCP model can be extended to such systems as well.

5.2.3 Mobile Messaging Systems

- *Message priority:* Although the current MMS specification has provided a mechanism to specify priority for a message, the network itself does not take any action with respect to message priorities. The sole purpose of these priorities is to inform the end user about the importance of the received messages. It is imperative to utilize message priorities to create different classes of service for messaging users, *e.g.*, real-time messaging, background messaging. However, extending the results presented in Chapter 4 to multiple message priorities is a challenging problem.

- *Instant Messaging:* Instant Messaging (IM) is an IP-based application that can provide real-time written communication. Mobile IM is seen as a natural evolution of the popular SMS service. Unlike SMS that uses control channels, IM shares the communication link with other IP-based services, *e.g.*, web browsing. It is interesting to study the performance of IM in terms of message loss and delay.

Bibliography

[1] J.-Z. Sun, J. Sauvola, and D. Howie, “Features in future: 4G visions from a technical perspective,” in *Proc. IEEE Globecom*, vol. 6, San Antonio, USA, November 2001, pp. 3533–3537.

[2] U. Varshney and R. Jain, “Issues in emerging 4G wireless networks,” *IEEE Computer*, vol. 34, no. 6, pp. 94–96, June 2001.

[3] S. Y. Hui and K. H. Yeung, “Challenges in the migration to 4G mobile systems,” *IEEE Computer*, vol. 41, no. 12, pp. 54–59, December 2003.

[4] T. Zahariadis and D. Kazakos, “(R)evolution toward 4G mobile communication systems,” *IEEE Wireless Communications Magazine*, vol. 10, no. 4, pp. 6–7, August 2003.

[5] R. Knopp and P. Humblet, “Information capacity and power control in single cell multiuser communications,” in *Proc. IEEE ICC*, Seattle, USA, June 1995.

[6] P. Bender, P. Black, M. Grob, R. Padovani, N. Sindhushayana, and A. Viterbi, “CDMA/HDR: A bandwidth efficient high speed wireless data service for nomadic users,” *IEEE Communications Magazine*, vol. 38, no. 7, pp. 70–77, July 2000.

[7] P. Viswanath, D. Tse, and R. Laroia, “Opportunistic beamforming using dumb antennas,” *IEEE Transactions on Information Theory*, vol. 48, no. 6, June 2002.

[8] 3GPP TS 23.040, “Technical realization of the short message service (SMS); release 5,” v5.2.0, 2001.

[9] 3GPP TS 22.140, "Multimedia messaging service (MMS); stage 1; release 6," v6.6.0, June 2004.

[10] D. Hong and S. S. Rappaport, "Traffic model and performance analysis for cellular mobile radio telephone systems with prioritized and nonprioritized handoff procedures," *IEEE Transactions on Vehicular Technology*, vol. 35, no. 3, pp. 77–92, August 1986, see also: CEAS Tech. Rep. No. 773, College of Engineering and Applied Sciences, State University of New York, June 1999.

[11] M. Naghshineh and M. Schwartz, "Distributed call admission control in mobile/wireless networks," *IEEE Journal on Selected Areas in Communications*, vol. 14, no. 4, pp. 711–717, May 1996.

[12] D. Levine, I. Akyildiz, and M. Naghshineh, "A resource estimation and call admission algorithm for wireless multimedia networks using the shadow cluster concept," *IEEE/ACM Transactions on Networking*, vol. 5, no. 1, pp. 1–12, February 1997.

[13] S. Choi and K. G. Shin, "Predictive and adaptive bandwidth reservation for handoffs in QoS-sensitive cellular networks," in *Proc. ACM SIGCOMM'98*, vol. 27, Vancouver, Canada, October 1998, pp. 155–166.

[14] B. M. Epstein and M. Schwartz, "Predictive QoS-based admission control for multiclass traffic in cellular wireless networks," *IEEE Journal on Selected Areas in Communications*, vol. 18, no. 3, pp. 523–534, March 2000.

[15] A. Aljadhai and T. F. Znati, "Predictive mobility support for QoS provisioning in mobile wireless networks," *IEEE Journal on Selected Areas in Communications*, vol. 19, no. 10, pp. 1915–1930, October 2001.

[16] R. Ramjee, D. Towsley, and R. Nagarajan, "On optimal call admission control in cellular networks," *ACM/Kluwer Wireless Networks*, vol. 3, no. 1, pp. 29–41, March 1997.

[17] S. Wu, K. Y. M. Wong, and B. Li, "A dynamic call admission policy with precision QoS guarantee using stochastic control for mobile wireless networks,"

IEEE/ACM Transactions on Networking, vol. 10, no. 2, pp. 257–271, April 2002.

[18] B. Li, L. Yin, K. Y. M. Wong, and S. Wu, “An efficient and adaptive bandwidth allocation scheme for mobile wireless networks using an on-line local estimation technique,” *ACM/Kluwer Wireless Networks*, vol. 7, no. 2, pp. 107–116, 2001.

[19] B. Epstein and M. Schwartz, “Reservation strategies for multi-media traffic in a wireless environment,” in *Proc. IEEE VTC*, vol. 1, Chicago, USA, July 1995, pp. 165–169.

[20] J. E. Wieselthier and A. Ephremides, “Fixed- and movable-boundary channel-access schemes for integrated voice/data wireless networks,” *IEEE Transactions on Communications*, vol. 43, no. 1, pp. 64–74, January 1995.

[21] M. C. Young and Y.-R. Haung, “Bandwidth assignment paradigms for broadband integrated voice/data networks,” *Computer Communications*, vol. 21, no. 3, pp. 243–253, 1998.

[22] H.-H. Liu, J.-L. C. Wu, and W.-C. Hsieh, “Delay analysis of integrated voice and data service for GPRS,” *IEEE Communications Letters*, vol. 6, no. 8, pp. 319–321, August 2002.

[23] D.-S. Lee and C.-C. Chen, “QoS of data traffic with voice handoffs in a PCS network,” in *Proc. IEEE Globecom*, vol. 2, Taipei, Taiwan, November 2002, pp. 1534–1538.

[24] M. A. Marsan, P. Laface, and M. Meo, “Packet delay analysis in GPRS systems,” in *Proc. IEEE INFOCOM*, vol. 2, San Francisco, USA, March 2003, pp. 970–978.

[25] Y.-R. Haung, Y.-B. Lin, and J.-M. Ho, “Performance analysis for voice/data integration on a finite-buffer mobile system,” *IEEE Transactions on Vehicular Technology*, vol. 49, no. 2, pp. 367–378, March 2000.

[26] L. Yin, B. Li, Z. Zhang, and Y.-B. Lin, "Performance analysis of a dual-threshold reservation (DTR) scheme for voice/data integrated mobile wireless networks," in *Proc. IEEE WCNC*, vol. 1, Chicago, USA, September 2000, pp. 258–262.

[27] H. Wu, L. Li, B. Li, L. Yin, I. Chlamtac, and B. Li, "On handoff performance for an integrated voice/data cellular system," in *Proc. IEEE PIMRC*, vol. 5, Lisboa, Portugal, September 2002, pp. 2180–2184.

[28] C.-J. Chang, T.-T. Su, and Y.-Y. Chiang, "Analysis of a cutoff priority cellular radio system with finite queueing and reneging/dropping," *IEEE/ACM Transactions on Networking*, vol. 2, no. 2, pp. 166–175, April 1994.

[29] B. Li, L. Li, B. Li, and X.-R. Cao, "On handoff performance for an integrated voice/data cellular system," *ACM/Kluwer Wireless Networks*, vol. 9, no. 4, pp. 393–402, July 2003.

[30] M. Ghaderi and R. Boutaba, "Call admission control in mobile cellular networks: A comprehensive survey," *Wireless Communications and Mobile Computing (WCMC)*, vol. 6, no. 1, pp. 69–93, February 2006.

[31] I. Katzela and M. Naghshineh, "Channel assignment schemes for cellular mobile telecommunication systems: A comprehensive survey," *IEEE Personal Communications Magazine*, vol. 3, no. 3, pp. 10–31, June 1996.

[32] C.-J. Chang, P.-C. Huang, and T.-T. Su, "A channel borrowing scheme in a cellular radio system with guard channels and finite queues," in *Proc. IEEE ICC*, vol. 2, Dallas, USA, June 1996, pp. 1168–1172.

[33] X. Wu and K. L. Yeung, "Efficient channel borrowing strategy for multimedia wireless networks," in *Proc. IEEE Globecom*, vol. 1, Sydney, Australia, November 1998, pp. 126–131.

[34] T.-P. Chu and S. S. Rappaport, "Generalized fixed channel assignment in microcellular communication systems," *IEEE Transactions on Vehicular Technology*, vol. 43, no. 3, pp. 713–721, August 1994.

[35] J. R. Moorman and J. W. Lockwood, “Wireless call admission control using threshold access sharing,” in *Proc. IEEE Globecom*, vol. 6, San Antonio, USA, November 2001, pp. 3698–3703.

[36] M. Ghaderi, J. Capka, and R. Boutaba, “Prediction-based admission control for DiffServ wireless Internet,” in *Proc. IEEE VTC*, vol. 3, Orlando, USA, October 2003, pp. 1974–1978.

[37] Y. Shu, Z. Jin, J. Wang, and O. W. Yang, “Prediction-based admission control using FARIMA models,” in *Proc. IEEE ICC*, vol. 3, New Orleans, USA, June 2000, pp. 1325–1329.

[38] Y. Shu, Z. Jin, L. Zhang, and L. Wang, “Traffic prediction using FARIMA models,” in *Proc. IEEE ICC*, vol. 2, Vancouver, Canada, June 1999, pp. 891–895.

[39] S. Jamin, P. B. Danzig, S. J. Shenker, and L. Zhang, “A measurement-based admission control algorithm for integrated services packet networks,” *IEEE/ACM Transactions on Networking*, vol. 5, no. 1, pp. 524–540, February 1997.

[40] C. Oliveira, J. B. Kim, and T. Suda, “An adaptive bandwidth reservation scheme for high-speed multimedia wireless networks,” *IEEE Journal on Selected Areas in Communications*, vol. 16, no. 6, pp. 858–874, August 1998.

[41] A. Acampora and M. Naghshineh, “An architecture and methodology for mobile-executed handoff in cellular ATM networks,” *IEEE Journal on Selected Areas in Communications*, vol. 12, no. 8, pp. 1365–1375, October 1994.

[42] R. Guerin, H. Ahmadi, and M. Naghshineh, “Equivalent capacity and its application to bandwidth allocation in high-speed networks,” *IEEE Journal on Selected Areas in Communications*, vol. 9, no. 7, pp. 968–981, 1991.

[43] F. P. Kelly, “Notes on effective bandwidths,” in *Stochastic Networks: Theory and Applications*, F. P. Kelly, S. Zachary, and I. Ziedins, Eds. Oxford, UK: Oxford University Press, 1996, pp. 141–168.

[44] M. Schwartz, *Broadband Integrated Networks.* New Jersey, USA: Prentice Hall, 1996.

[45] Y. Fang, I. Chlamtac, and Y.-B. Lin, "Channel occupancy times and hand-off rate for mobile computing and PCS networks," *IEEE Transactions on Computers*, vol. 47, no. 6, pp. 679–692, June 1998.

[46] Y. Fang and I. Chlamtac, "Analytical generalized results for handoff probability in wireless networks," *IEEE Transactions on Communications*, vol. 50, no. 3, pp. 396–399, March 2002.

[47] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes in C: The Art of Scientific Computing*, 2nd ed. Cambridge, UK: Cambridge University Press, 1992.

[48] P. Whittle, *Prediction and Regulation by Linear Least-Square Methods*, 2nd ed. Minnesota, USA: University of Minnesota Press, 1983.

[49] P. J. Brockwell and R. A. Davis, *Time Series: Theory and Methods*, 2nd ed. New York, USA: Springer-Verlag, 1991.

[50] C. Jedrzycki and V. C. M. Leung, "Probability distribution of channel holding time in cellular telephone systems," in *Proc. IEEE VTC*, vol. 1, Atlanta, GA, May 1996, pp. 247–251.

[51] M. M. Zonoozi and P. Dassanayake, "User mobility modeling and characterization of mobility patterns," *IEEE Journal on Selected Areas in Communications*, vol. 15, no. 7, pp. 1239–1252, September 1997.

[52] R. Guerin, "Channel occupancy time distribution in a cellular radio system," *IEEE Transactions on Vehicular Technology*, vol. 35, no. 3, pp. 89–99, 1987.

[53] T. I. Association, "TIA EIA IS-2000," March 2000. [Online]. Available: www.tiaonline.org/standards/sfg/imt2k/cdma2000/

[54] G. T. Standards, "W-CDMA , IMT-2000 TIA/EIA Standard."

[55] QualComm, "1xEV: 1x EVolution, IS-856 TIA/EIA Standard." [Online]. Available: http://www.qualcomm.com/technology/1xev-do/whitepapers.html

[56] K. Mattar, A. Sridharan, H. Zang, I. Matta, and A. Bestavros, "Performance evaluation of TCP in CDMA 2000 networks," Sprint Advanced Technology Labs," Technical Report *RR06-ATL-030567*, March 2006.

[57] E. Chaponniere, S. Kandukuri, and W. Hamdy, "Effect of physical layer bandwidth variation on TCP performance in CDMA2000," in *Proc. IEEE VTC Spring*, April 2003, pp. 336–342.

[58] C. Barakat and E. Altman, "Bandwidth tradeoff between TCP and link-level FEC," *Computer Networks*, vol. 39, no. 5, pp. 133–150, 2002.

[59] B. Liu, D. L. Goeckel, and D. Towsley, "TCP-cognizant adaptive forward error correction in wireless networks," in *Proc. IEEE Globecom*, Taipei, Taiwan, November 2002.

[60] D. Barman, I. Matta, E. Altman, and R. E. Azouzi, "TCP optimization through FEC, ARQ and transmission power trade offs," in *Proc. WWIC*, Frankfurt, Germany, February 2004.

[61] F. Baccelli, R. Cruz, and A. Nucci, "CDMA channel parameters maximizing TCP throughput," in *Proc. Workshop on Information Theory and its Applications*, La Jolla, USA, February 2006.

[62] K. L. Gray and D. L. Noneaker, "The effect of adaptive-rate coding on TCP performance in wireless communications," in *Proc. EUROCOMM*, 2000.

[63] J. P. Singh, Y. Li, and N. Bambos, "Channel state awareness based transmission power adaptation for efficient TCP dynamics in wireless networks," in *Proc. IEEE ICC*, 2005.

[64] C. Casetti, M. Gerla, S. Mascolo, M. Y. Sanadidi, and R. Wang, "TCP Westwood: end-to-end congestion control for wired/wireless networks," *ACM Wireless Networks*, vol. 8, no. 5, pp. 467–479, 2002.

[65] T. Goff, J. Moronski, D. S. Phatak, and V. Gupta, "Freeze-TCP: A true end-to-end TCP enhancement mechanism for mobile environments," in *Proc. IEEE INFOCOM*, Tel Aviv, Israel, 2000, pp. 1537–1545.

[66] R. Ludwig and R. H. Katz, "The Eifel algorithm: Making TCP robust against spurious retransmissions," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 30, no. 1, pp. 30–36, 2000.

[67] R. Yavatkar and N. Bhagawat, "Improving end-to-end performance of TCP over mobile internetworks," in *Proc. IEEE Workshop on Mobile Computing Systems and Applications*, Santa Cruz, USA, 1994.

[68] H. Balakrishnan, S. Seshan, and R. H. Katz, "Improving reliable transport and handoff performance in cellular wireless networks," *ACM Wireless Networks*, vol. 1, no. 4, pp. 469–481, 1995.

[69] K. Ratnam and I. Matta, "WTCP: An efficient transmission control protocol for networks with wireless links." in *Proc. IEEE ISCC*, Athens, Greece, 1998.

[70] B. S. Bakshi, P. Krishna, N. H. Vaidya, and D. K. Pradhan, "Improving performance of TCP over wireless networks," in *Proc. International Conference on Distributed Computing Systems*, 1997.

[71] H. Balakrishnan and R. Katz, "Explicit loss notification and wireless web performance," in *Proc. IEEE Globecom*, Sydney, Australia, November 1998.

[72] M. C. Chan and R. Ramjee, "TCP/IP performance over 3G wireless links with rate and delay variation," in *Proc. ACM MOBICOM*, 2002, pp. 71–82.

[73] ——, "Improving TCP/IP performance over third generation wireless networks," in *Proc. IEEE INFOCOM*, Hong Kong, 2004.

[74] C. Barakat, E. Altman, and W. Dabbous, "On TCP performance in an heterogeneous network: A survey," *IEEE Communications Magazine*, pp. 40–46, January 2000.

[75] H. Elaarag, "Improving TCP performance over mobile networks," *ACM Comput. Surv.*, vol. 34, no. 3, pp. 357–374, 2002.

[76] A. Chockalingam, E. Altman, J. V. K. Murthy, and R. Kumar, "Cross-layer design for optimizing TCP performance," in *Proc. IEEE ICC*, Seoul, Korea, May 2005.

[77] M. Zorzi, A. Chockalingam, and R. R. Rao, "Throughput analysis of TCP on channels with memory," *IEEE J. Select. Areas Commun.*, vol. 18, no. 7, pp. 1289–1300, 2000.

[78] A. Chockalingam and G. Bao, "Performance of TCP/RLP protocol stack on correlated rayleigh fading DS-CDMA links," *IEEE Trans. Veh. Technol.*, vol. 49, no. 1, pp. 28–33, 2000.

[79] T. Klein, K. Leung, and H. Zheng, "Improved TCP performance in wireless IP networks through enhanced opportunistic scheduling algorithms," in *Proc. IEEE Globecom*, vol. 5, 2004, pp. 2744–2748.

[80] E. Altman, C. Barakat, and V. M. R. Ramos, "Analysis of AIMD protocols over paths with variable delay," in *Proc. IEEE INFOCOM*, Hong Kong, March 2004.

[81] V. Misra, W.-B. Gong, and D. Towsley, "Fluidbased analysis of a network of AQM routers supporting TCP flows with an application to RED," in *Proc. ACM SIGCOMM*, Stockholm, Sweden, August 2000, pp. 151–160.

[82] F. Baccelli and K. B. Kim, "TCP throughput analysis under transmission error and congestion losses," in *Proc. IEEE INFOCOM*, Hong Kong, March 2004.

[83] F. Baccelli, K. B. Kim, and D. D. Vlleschauwer, "Analysis of the competition between wired, DSL and wireless TCP flows in an access network," in *Proc. IEEE INFOCOM*, Miami, USA, March 2005.

[84] F. J. McWilliams and N. J. A. Sloane, *The Theory of Error-Correcting Codes.* New York, USA: North-Holland, 1977.

[85] Nokia, "MMS technology tutorial." [Online]. Available: http://www.nokia.com/support/tutorials/MMS/en/mms.html

[86] OMA, "Multimedia messaging service; architecture overview," v1.2, December 2003.

[87] 3GPP TS 26.140, "Multimedia messaging service (MMS); media formats and codecs; release 6," v6.0.0, September 2004.

[88] Y.-R. Haung, "Determining the optimal buffer size for short message transfer in a heterogeneous GPRS/UMTS network," *IEEE Transactions on Vehicular Technology*, vol. 52, no. 1, pp. 216–225, January 2003.

[89] 3GPP TS 23.140, "Multimedia messaging service (MMS); functional description; stage 2; release 6," v6.6.0, June 2004.

[90] S. Coulombe and G. Grassel, "Multimedia adaptation for the multimedia messaging service," *IEEE Communications Magazine*, vol. 42, no. 7, pp. 120–126, July 2004.

[91] WAP Forum, "Wireless application protocol architecture specification; release 2.0," July 2001.

[92] 3GPP TS 23.057, "Mobile execution environment (MExE); functional description; stage 2; release 6," v6.2.0, 2003.

[93] P. Faltstrom, "E.164 number and DNS," IETF, RFC 2822, September 2000.

[94] D. Gross and C. M. Harris, *Fundamentals of Queueing Theory*, 3rd ed. New York, USA: John Wiley & Sons, Inc., 1998.

[95] MathWorld. [Online]. Available: http://mathworld.wolfram.com

[96] D. Y. Barrer, "Queueing with impatient customers and ordered service," *Operations Research*, vol. 5, no. 5, pp. 650–656, October 1957.

[97] F. Baccelli and G. Hebuterne, "On queues with impatient customers," in *Performance'81*, F. J. Kylstra, Ed. Oxford, UK: North-Holland Publishing Company, 1981, pp. 159–179.

[98] B. V. Gnedenko and I. N. Kovalenko, *Introduction to Queueing Theory*, 2nd ed. Boston, USA: Birkhauser, 1989.

[99] A. Movaghar, “On queueing with customer impatience until the beginning of service,” *Queueing Systems*, vol. 7, no. 3, pp. 15–23, June 1998.

[100] J. Bae, S. Kim, and E. Y. Lee, “The virtual waiting time of the M/G/1 queue with impatient customers,” *Queueing Systems*, vol. 38, no. 4, pp. 485–494, August 2001.

[101] A. Brandt and M. Brandt, “Asymptotic results and a Markovian approximation for the M(n)/M(n)/s+GI system,” *Queueing Systems*, vol. 41, no. 1-2, pp. 73–94, June 2002.

[102] A. R. Ward and P. W. Glynn, “A diffusion approximation for a Markovian queue with reneging,” *Queueing Systems*, vol. 43, no. 1-2, pp. 103–128, June 2003.

[103] 3GPP TS 23.228, “IP multimedia subsystem (IMS); stage 2; release 6,” v6.9.0, March 2005.

[104] T. Bu, M. C. Chan, and R. Ramjee, “Designing wireless radio access networks for third generation cellular networks,” in *Proc. IEEE INFOCOM*, Miami, USA, 2005.

[105] A. Jalali, R. Padovani, and P. Rankaj, “Data throughput of cdma-hdr a high efficiency-high data rate personal communication wireless system,” in *Proc. IEEE VTC*, Tokyo, Japan, May 2000.