Multiple Access Protocols for Multichannel Communication Systems

by

Serena Chan

S.B. Computer Science and Engineering
Massachusetts Institute of Technology, 1999

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Master of Engineering in Electrical Engineering and Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2000

Author .....
Department of Electrical Engineering and Computer Science
May 17, 2000

Certified by .....
Professor Vincent W.S. Chan
Director, EECS Laboratory for Information and Decision Systems
Thesis Supervisor

Accepted by .....
Arthur C. Smith
Chairman, Department Committee on Graduate Theses

Multiple Access Protocols for Multichannel Communication Systems

by

Serena Chan

Submitted to the Department of Electrical Engineering and Computer Science
on May 17, 2000, in partial fulfillment of the
requirements for the degree of
Master of Engineering in Electrical Engineering and Computer Science

Abstract

System architecture design, evaluation, and optimization are key issues to developing communication systems that meet the requirements of today and expectations of the future. In this thesis, we introduce the concept of multiple access communication and the need to use efficient transmission techniques to expand both present and future wireless communication networks. We will study two areas regarding multiple access on multichannel communication systems. First, we describe fundamental multiplexing techniques that we can build upon and investigate the performance of different candidate architectures for the transmission of messages from bursty sources on multiple channels. We will consider traditional protocols such as Time Division Multiple Access (TDMA) and Slotted ALOHA (S-ALOHA) alongside a channelized architecture, which is based on the idea of multiplexing by dividing total transmission capacity into a fixed number of frequency channels. We develop mathematical models that describe the overall delay for sending large messages of a fixed length arriving from bursty sources and analyze their performances.

We will make real-world parameter assumptions in the context of wireless networks and analyze the performance to develop intuition about the effectiveness of the different architectures. Second, we will investigate channel capacity allocation among mixed traffic, i.e., multiple classes of users. We will consider a first-come first-serve (FCFS) access strategy, a nonpreemptive priority scheme, a preemptive resume priority scheme, and several channel capacity allocation schemes. We develop models that describe the overall delay for sending messages and analyze their performance. Our focus will concentrate on two classes of users. This scenario is typical of classes of users with small and large messages to transmit. present quantitative results by making real-world parameter assumptions in the context of wireless networks and analyze the performance to develop intuition about the effectiveness of each architecture.

Thesis Supervisor: Professor Vincent W.S. Chan
Title: Director, EECS Laboratory for Information and Decision Systems

Acknowledgments

This VI-A thesis is done with the support of MIT Lincoln Laboratory, a center for research operated by Massachusetts Institute of Technology, under Air Force Contract F19628-95-C-0002. Thanks to my thesis advisors Professor Vincent W.S. Chan at MIT and Steven Bernstein at MIT Lincoln Laboratory.

Contents

List of Figures

List of Tables

Chapter 1

Multiple Access Communication

1.1 Introduction

Next generation wireless networks are envisioned to support high data rates and multimedia traffic using packet oriented transport. Data calls can arise from various platform types having different mobility characteristics. Cellular communication systems are now being designed to support data transmission in addition to voice calls. Thus, the integration of voice and data combines two types of services of very different characteristics, one bursty and unscheduled, the other constant or variable rate streams. Multimedia clients may operate in different modes - (1) streaming (e.g., voice/video only), (2) bursty data only, or (3) both. The design of efficient and robust wireless media access protocols and the evaluation of their performance are key technical issues to be addressed for future wireless networks.

High capacity wireless networks can be realized by either assigning a single wideband channel or by using multiple narrow band channels. The latter approach is necessary when contiguous wide bandwidth spectrum is not available or channelization within a band is mandated. In a multichannel system, a user can transmit on any of these channels during idle times between voice calls to transmit data calls. This concept is based on different attributes of voice and data services. A delay of voice service in cellular communication network of 100ms will be recognizable and irritating to the user, but a small amount of delay for the data users is not crucial.

The future success of mobile communications depends on its ability to efficiently accommodate integrated traffic and service a variety of applications and communication sources with different quality of service requirements. It is essential to use efficient algorithms to support integrated services such that the Quality of Service (QoS) requirements of the various types of applications are met.

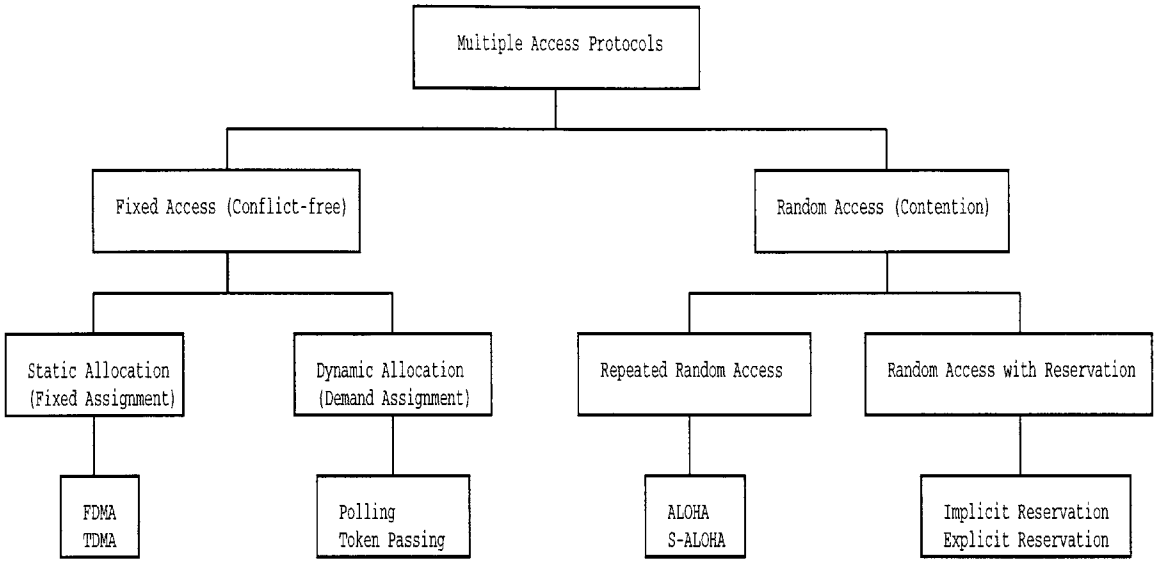


Figure 1-1: Classification of Multiple Access Protocols

1.2 Multiple Access Communication

Multiple access communication involves the sharing of a communication channel between a multiplicity of users. A difficult problem arises when each user communicates infrequently and sporadically, which is typical for bursty computer communications. Multiple access protocols are channel allocation schemes that encompass desirable performance characteristics. Multiple access protocols can be subdivided into fixed access and random protocols. Figure 1-1 illustrates this classification [34]. Fixed access (or conflict-free) protocols ensure that whenever a transmission is made on error-free channels, it will not be interfered by another transmission, and thus is successful. Conflict-free transmission can be achieved by allocating the channel to the users either statically or dynamically. Channel resources can be divided in terms of time, frequency, or some combination of time and frequency. With Time Division Multiple Access (TDMA), the channel can be divided by providing the entire frequency range (bandwidth) to a single user for a fraction of the time. With Frequency Division Multiple Access (FDMA), a fraction of the frequency range is given to every user upon call set-up until transmission is over and the channel is relinquished. Code Division Multiple Access (CDMA) is another multiple access technique which places all users on the same frequency spectrum at the same time. Each user is then identified on the channel with a unique code.

A system employing random access (or contention) protocols allows users to access the channel at any time. However, this can result in colliding transmissions where the conflicts need to be resolved. As shown in Figure 1-1, contention protocols can be categorized into two classes, repeated random protocols and random protocols with reservation. Under random protocols with reservation, a user's initial transmission

uses a random access method to gain access to the channel. Subsequent transmissions of that user are then scheduled until there is nothing more to send. Reservations are further classified as implicit and explicit reservations. The implicit reservation scheme does not use any reservation packets while a short reservation packet is used to request transmission at scheduled times in explicit reservation schemes.

1.3 Fixed Access Protocols

Fixed access protocols are designed to ensure that a transmission, whenever made, is not interfered by any other transmission and is therefore successful. Guaranteed transmission is achieved by allocating resources to users without any overlap. An important advantage of conflict-free access protocols is the ability to ensure fairness among users and the ability to control the packet delay - a feature that may be essential in real-time applications.

The two most well known protocols in this class are the Frequency Division Multiple Access (FDMA) in which a fraction of the frequency bandwidth is allocated to every user all the time, and the Time Division Multiple Access (TDMA) in which the entire bandwidth is used by each user for a fraction of the time. In principle, no overhead in the form of control messages is incurred in both the FDMA and TDMA protocols. However, due to the static and fixed assignment, parts of the channel resources might be idle even though some other users have data to transmit.

1.3.1 Frequency Division Multiple Access

With Frequency Division Multiple Access (FDMA), the entire available capacity of a channel C is subdivided into K frequency bands each with rate $\frac{C}{K}$ to serve a single user. Figure 1-2 illustrates the implementation of FDMA on a channel with equal capacity subchannels. Note that if users have uneven long term demands, it is possible to divide the frequency range unevenly, i.e., proportional to the demands. The main advantage of FDMA is its simplicity. It does not require any coordination or synchronization among the users since each can use its own frequency band without interference. This protocol, however, wastes channel resources especially when one user is idle. Other users cannot utilized an idle user's share of the channel. For example, if a user is using a communication system for bursty computer communications at 1% duty cycle, the utility (efficiency) of the resources will be as low as 1%. An additional disadvantage to the use of FDMA is the lack of flexibility or scalability when it comes to adding additional users to the network.

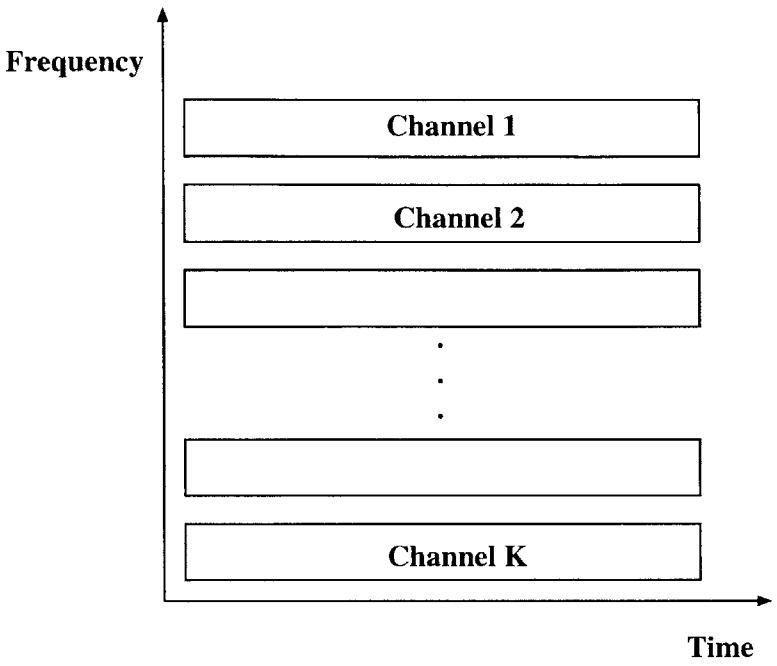


Figure 1-2: FDMA Channel Allocation

The expected delay for the transmission of a message using FDMA is

$$\overline{X}_{FDMA} = \frac{NL}{C}, \tag{1.1}$$

assuming a channel of capacity C [bits/sec], message length of L [bits], N users and ignoring queueing delay, i.e. an increasing arrival rate of messages. By applying queueing theory, we can include the queueing delay to determine the total expected wait time. For simplicity, we will use a M/D/1 queue, assuming fixed-length messages, Poisson message arrivals and an infinite user buffer. The expected service rate for this case is $\mu = \frac{1}{\overline{X}_{FDMA}}$. We also assume that the users are given a predetermined frequency and transmit their messages only on that channel. Subsequently, the total expected transmission delay is

$$\begin{aligned} D_{FDMA} &= Q_{M/D/1}(\frac{\lambda}{K}, \frac{1}{\overline{X}_{FDMA}}) + \overline{X}_{FDMA} + 2T_{PD} \\ &= \frac{\lambda \overline{X^2}_{FDMA}}{2(1 - \lambda \overline{X}_{FDMA})} + \overline{X}_{FDMA} + 2T_{PD} \\ &= \frac{\lambda N^2 L^2}{C^2(2 - 2\frac{\lambda NL}{C})} + \frac{NL}{C} + 2T_{PD}, \end{aligned} \tag{1.2}$$

where λ is the total message arrival rate of N users and T_{PD} is the propagation delay. The propagation delay is defined as the time between the last transmitted bit at the source and the time of the last received bit at the destination. The round-

trip propagation delay is thus $2T_{PD}$. See Appendix A for more information about queueing theory and the $Q_{M/D/1}$ formula.

1.3.2 Time Division Multiple Access

In Time Division Multiple Access (TDMA), the time axis is divided into time slots, pre-assigned to the different users. Each user is allowed to transmit freely during its assigned slot. During the allotted time slot the entire system resources are devoted to that user. The slot assignments follow a predetermined pattern that repeats itself periodically; each such period is called a cycle or a frame, as shown in Figure 1-3. Thus each frame consists of a sequence of slots: slot 1, slot 2, ..., slot N. A user occupies every Nth slot. The first user occupies slot 1, $N+1$, $2N+1$, ...; the second user occupies slots 2, $N+2$, $2N+2$, ...; and so on.

To calculate the expected delay of sending a message using TDMA, we begin by assuming that each user has messages of length L [bits] to send on a channel with capacity C [bits/sec]. Messages are divided into packets of length l_p [bits]. A ceiling function is used to ensure that enough packets are used to transmit the entire message. Within a frame of length $T_f = N\tau$ [sec], a slot of duration $\tau = \frac{l_p}{C}$ [sec] is given to each user. Thus, the expected transmission time for a single user to send a message of length L [bits] is

$$\overline{X}_{TDMA} = (\lceil \frac{L}{l_p} \rceil - 1)T_f + \tau. \tag{1.3}$$

Taking queueing delay into account, we can determine the total expected wait time for transmitting a message. For simplicity, we will use a M/D/1 queue, assuming fixed-length messages, Poisson message arrivals and an infinite user buffer. With an assumption of fixed-length messages, Poisson message arrivals and an infinite use buffer, we can model this case with a M/D/1 queue. The expected service rate for

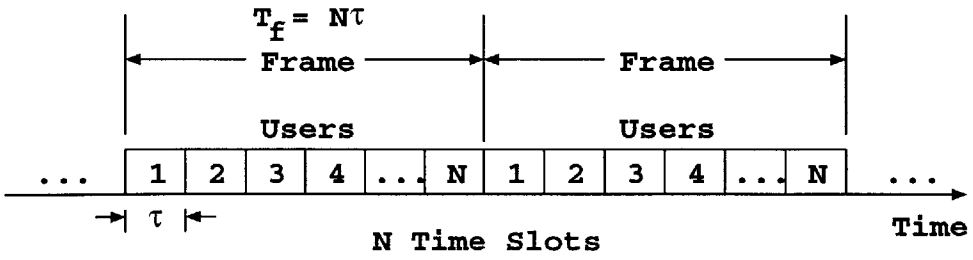


Figure 1-3: TDMA Channel Allocation

this case is $\mu = \frac{1}{\overline{X}_{TDMA}}$. Subsequently, the total expected transmission delay is

$$\begin{aligned} D_{TDMA} &= \frac{T_f}{2} + Q_{M/D/1}(\lambda, \frac{1}{\overline{X}_{TDMA}}) + \overline{X}_{TDMA} + 2T_{PD} \\ &= \frac{N\tau}{2} + \frac{\lambda \overline{X^2}_{TDMA}}{2(1 - \lambda \overline{X}_{TDMA})} + \overline{X}_{TDMA} + 2T_{PD} \\ &= \frac{Nl_p}{2} + \frac{\lambda(\frac{\lceil \frac{L}{l_p} \rceil - 1)Nl_p}{C} + \frac{l_p}{C})^2}{2 - 2\lambda(\frac{(\lceil \frac{L}{l_p} \rceil - 1)Nl_p}{C} + \frac{l_p}{C})} + \frac{(\lceil \frac{L}{l_p} - 1)Nl_p}{C} + \frac{l_p}{C} + 2T_{PD}, \quad (1.4) \end{aligned}$$

where λ is the total message arrival rate of N users and T_{PD} is the propagation delay. The $\frac{T_f}{2}$ term is the expected amount of time it takes to reach the allotted time slot. The round-trip propagation delay is denoted as $2T_{PD}$.

1.4 Code Division Multiple Access

Code Division Multiple Access (CDMA) places all users onto the same frequency spectrum at the same time, as shown in Figure 1-4. Each user is identified on the channel with a unique code. This code is used at the transmitting site to encode the traffic and it may also be used to spread it across the frequency spectrum. At the receiver, the code is used to extract the user's data.

In concept, CDMA is intended to provide more capacity than FDMA or TDMA as well as allow a graceful degradation of the channel performance as more users enter a cell and use the spectrum. In actual practice, CDMA is quite complex and some of the concepts, while attractive on paper, are difficult to implement. Without closed-form equations for analysis, CDMA access is not further discussed in this thesis.

1.5 Random Access Protocols

An alternative approach to the use of static allocation schemes described above is the use of dynamic allocations based on user demand. Users can make reservations that announce their intent to transmit. Reservation schemes can be implemented with the use of either time division multiplexing (or round-robin ordering) to make the reservations for channel use or collision resolution to resolve conflicts that arise when users transmit requests at the same time.

Contention (random access) schemes differ from conflict-free (fixed access) schemes

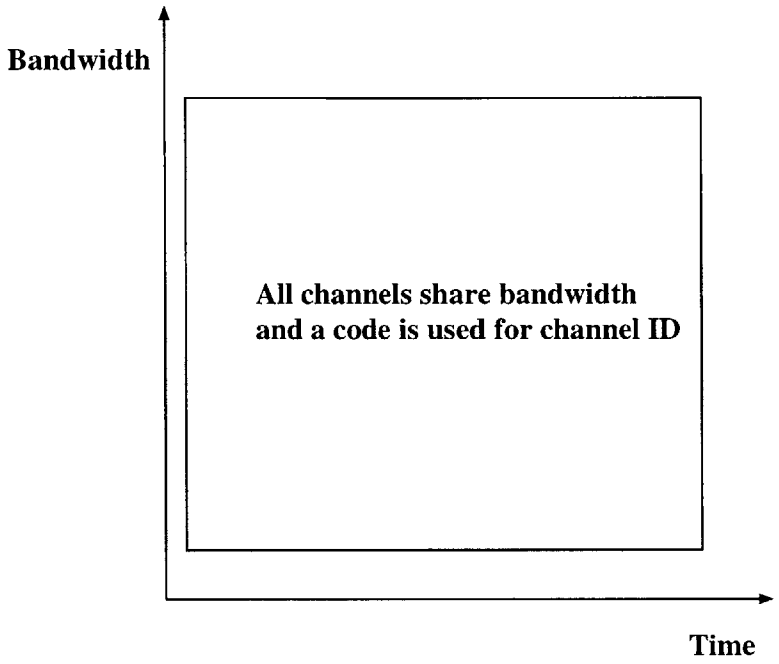


Figure 1-4: CDMA Channel Allocation

in principle since a transmission from a user is not guaranteed to be successful. To guarantee that all messages are eventually transmitted successfully, the protocol must dictate a way to resolve conflicts once they occur. Another difference between contention schemes and conflict-free schemes is the handling of idle users. We have noted in the previous sections that idle users consume a portion of the channel resources when using fixed access protocols. This wastefulness becomes impractical when there is a large number of potential users in the system. In contrast, idle users do not transmit when using contention schemes and thus do not consume any portion of the channel resources. With bursty users, if the probability of interference is small, taking the chance of having to resolve the interference compensates for the resources that have to be expanded to ensure freedom of conflicts.

1.5.1 ALOHA

The pure ALOHA system was proposed at the University of Hawaii in 1970 [1]. It provided radio communication between a central computer and various data terminals. Unslotted ALOHA is the most uncoordinated protocol. Users are allowed to send packets whenever they have anything to transmit, regardless of what other users may be doing. It is then assumed that a user can learn whether the packet has been successfully received at its destination. There are two ways the sender can learn this: (1) if all users are able to observe all transmissions, the sender can determine whether its packet has collided with any others; (2) an acknowledgment packet can be sent

back from the destination. A unsuccessful packet needs to be retransmitted. In order to avoid re-collision, the two colliders must each wait a random amount of time before retransmitting.

To develop the mathematics associated with the pure ALOHA random access protocol, we begin by assuming that the number of packets generated and retransmitted in the network are Poisson distributed, with a mean generation rate of λ packets per second. Every packet will have the same fixed duration of τ [sec], where $\tau = \frac{l_p}{C}$. Subsequently, the mean offered channel traffic G in packets per time slot is

$$G = \tau\lambda. \tag{1.5}$$

The probability that n packets are generated is given by

$$Pr[n,t] = \frac{(Gt)^n}{n!}e^{-Gt}. \tag{1.6}$$

A newly arrived packet is transmitted immediately. Our assumptions include an infinite number of users and memoryless packet generation per user. If a packet needs to be retransmitted, we assume it goes back into the pool of potential arrivals, so that the class of new packets and retransmitted packets still obey a Poisson distribution. We are aware that this assumption does not clearly reflect the dynamics or variations in retransmission rate. An individual user's packets actually tend to arrive in bursts. Even though our assumptions are not realistic, they are used to make the analysis simpler.

Consider a packet (new or old) scheduled for transmission at some time t as shown in Figure 1-5. This packet will be successful if no other packet is scheduled for transmission in the interval $(t-T, t+T)$. This period of duration $2T$ is called the vulnerable period. The probability of success P_{succ}, is the probability that no packet is scheduled in an interval of length $2T$. Since the scheduling points correspond to a Poisson process, for a newly generated packet, we have

$$\begin{aligned} P_{succ} &= Pr[n=0, t=2] \\ &= e^{-2G}. \end{aligned} \tag{1.7}$$

Since the mean offered channel traffic in packets per time slot is G, the rate of successfully transmitted packets is

$$\begin{aligned} S &= GP_{succ} \\ &= Ge^{-2G}, \end{aligned} \tag{1.8}$$

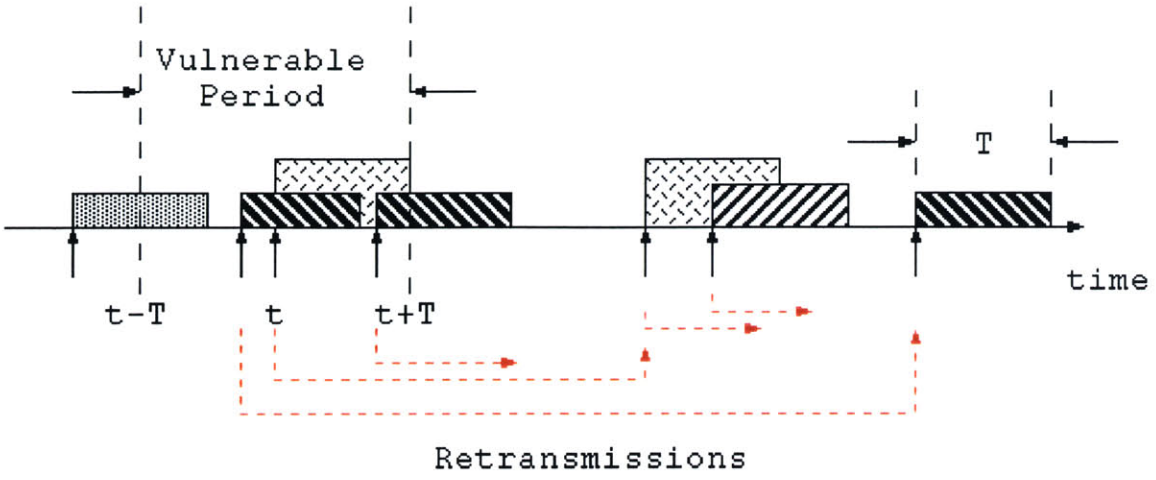


Figure 1-5: Packet Timing for Pure ALOHA

where S indicates throughput. From Equation 1.8, we can see that the maximum achievable throughput is

$$S_{max} = \frac{1}{2e} \approx 0.184. \tag{1.9}$$

The low throughput is the price we pay for a simple protocol that allows for superior delay performance when traffic is light. To calculate the expected packet delay, we need to consider three factors: transmission delay τ, delay due to collisions, and the round-trip propagation delay $2T_{PD}$. Although the transmission time and the propagation delays are fixed, delays due to collisions may vary depending on the resolution scheme. In our case, a user must wait a random uniformly distributed delay $U([0, H])$ before retransmitting. From the analysis expressed in [13], the expected number of required retries r is

$$E[r] \approx e^{2G} - 1 \tag{1.10}$$

with an expected delay T_c per collision of

$$E[T_c] = \frac{(H+1)\tau}{2} + 2T_{PD} \text{ for } H \gg 1. \tag{1.11}$$

The total expected delay for the successful transmission of a newly generation packet therefore is

$$\begin{aligned} D_{ALOHA} &= \tau + E[r]E[T_c] + 2T_{PD} \\ &= \frac{l_p}{C} + [e^{2G} - 1][\frac{(H+1)l_p}{2C} + 2T_{PD}] + 2T_{PD} \text{ for } H \gg 1, \end{aligned} \tag{1.12}$$

where $\tau = \frac{l_p}{C}$.

A plot of the pure ALOHA protocol in terms of its throughput S versus the offered load G is shown in Figure 1-6. Under light traffic conditions, we can see that ALOHA communication is very attractive. In most cases, a sender gains immediate use of the full data-carrying capacity of the channel. However, under heavy traffic conditions, the attainable efficiency is pretty low and there are potential instability problems. If the offered load increases beyond the value of $S_{max} = \frac{1}{2e}$, the system will continue to drift into higher load and lower throughput, as seen in the figure.

1.5.2 Slotted ALOHA

Slotted ALOHA (S-ALOHA) introduces one degree of coordination to the pure ALOHA scheme in that the channel is divided into time slots and all packets are sent entirely within a slot. This time synchronization is to improve the poor performance of pure ALOHA by reducing packet collisions. By restricting a user to transmit at the beginning of the next time slot, if a collision occurs, there is no partial overlap of the colliding packets. Since packets completely overlap, the vulnerable period for the probability of a packet transmission is τ.

To begin our analysis of the S-ALOHA protocol, packets are presumed to be generated by the infinite set of all users at a total finite rate of G packets per time slot. The number of arrivals in a slot is assumed to obey a Poisson distribution

$$P[n] = \frac{G^n}{n!} e^{-G}. \tag{1.13}$$

A newly arrived packet is transmitted in the first available slot. By assuming an infinite number of users and memoryless packet generation per user, we do not have to worry about a particular user having two packets to send at the same time because the probability that a newly generated packet belongs to one of the finite set of busy users goes to zero. If a packet needs to be retransmitted, we assume it goes back into the pool of potential arrivals, so that the class of new packets and retransmitted packets still obeys a Poisson distribution. We are aware that this assumption does not clearly reflect the dynamics or variations in retransmission rate. An individual user's packets actually tend to arrive in bursts. Even though our assumptions are not realistic, they are used to make the analysis simpler.

Let G be the total Poisson arrival rate of new packets plus retransmitted packets per time slot. There is a successful transmission in a slot if and only if exactly one transmission occurs. Let S be the fraction of successful slots, also known as the

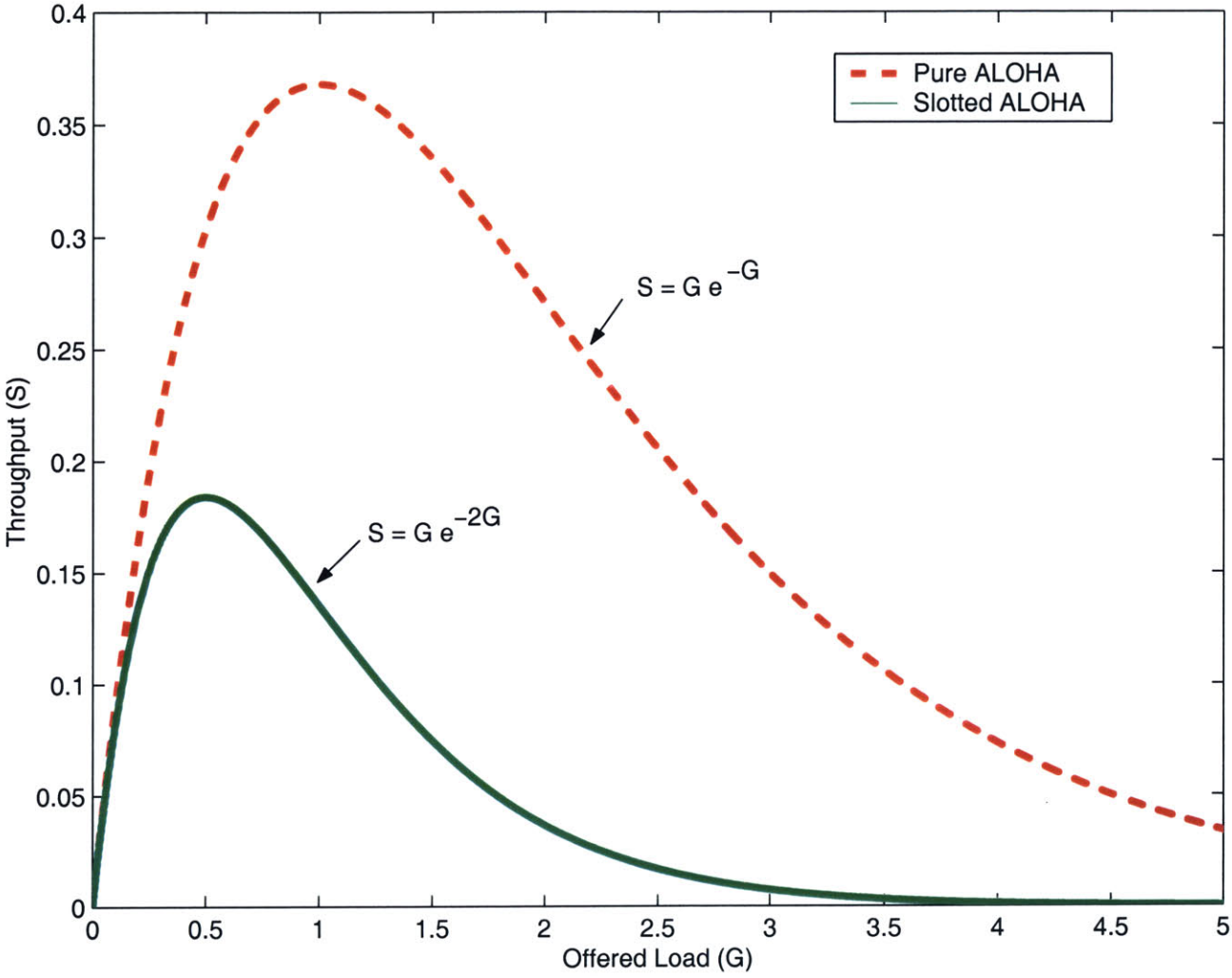


Figure 1-6: Throughput of Pure ALOHA and Slotted ALOHA

throughput per slot

$$S = Ge^{-G}, \tag{1.14}$$

which has a maximum achievable throughput of

$$S_{max} = \frac{1}{e} \approx 0.368. \tag{1.15}$$

A plot of the S-ALOHA protocol in terms of its throughput S versus the offered load G is shown in comparison to the pure ALOHA protocol in Figure 1-6. Notice that S-ALOHA shows an improvement in throughput but still suffers from instability.

The expected transmission time experienced per newly generated packet for S-ALOHA follows directly from the development for pure ALOHA. The new expected number of retransmission for S-ALOHA is

$$E[r] \approx e^G - 1. \tag{1.16}$$

The expected delay T_c per collision must now also consider an additional expected delay of waiting for the next available time slot, as shown by

$$\begin{aligned} E[T_c] &= \frac{\tau}{2} + \frac{(H+1)\tau}{2} + 2T_{PD} \\ &= \frac{(H+2)\tau}{2} + 2T_{PD} \text{ for } H \gg 1. \end{aligned} \tag{1.17}$$

The average time to gain access to a slot is an additional $\frac{\tau}{2}$ and then its transmission time is τ with a round-trip propagation delay of $2T_{PD}$. By summing all three main delay components, the total expected delay for a newly generated packet with S-ALOHA is thus

$$\begin{aligned} D_{S-ALOHA} &= \frac{\tau}{2} + \tau + E[r]E[T_c] + 2T_{PD} \\ &\approx \frac{3l_p}{2C} + [e^G - 1][\frac{(H+2)\tau}{2} + 2T_{PD}] + 2T_{PD} \text{ for } H \gg 1, \end{aligned} \tag{1.18}$$

where $\tau = \frac{l_p}{C}$. The term $\frac{\tau}{2}$ indicates the expected amount of time to access a slot while the term τ is the slot transmission time.

1.6 Multichannel Systems

Under certain implementation scenarios, the multichannel approach can prove to provide advantages over single channel systems, both in terms of throughput and delay

performances, and from the viewpoint of reliability and system management. Multichannel local area networks (LAN) have been simulated in [24] and [25] with CSMA and CSMA/CD protocols. With carrier sense multiple access (CSMA), a packet is not allowed to transmit if the channel is sense to be busy. If two users star to transmit almost simultaneously, they will shortly detect a collision in process and both cease transmitting. This technique is called CSMA/Collision Detection (CSMA/CD). Multichannel LAN's with nonpersistent CSMA/CD protocols can provide significant advantages in terms of the average packet delay and enormous gains in the packet delay variance with respect to the case of a single channel LAN for equal total data rate. The availability of several parallel channels allows for rescheduling delays much shorter than in single channel LAN's to be used, due to the possibility of separating retransmission attempts by transmitting over different channels. Remarkable performance improvements can be obtained when transmitting multipacket messages, if parallel transmission are allowed; when the average number of packets per message is not smaller than the number of channels, the average and variance of the message delay obtained with the multichannel option monotonically decreases for an increasing number of channels.

Multichannel systems of the same bandwidth can offer cost reduction, high reliability, fault tolerance, flexibility, and scalability. The subdivision of the channel into lower speed channels allows the use of simple technologies in the design of network interfaces, thus reducing the cost of implementing multichannel systems. The modular structure of multichannel systems allow for gradual growth with the addition of new channels depending on load demands, thereby extending the useful life of these systems. Existing networks can be upgraded with the addition of new equipment instead of replacing the entire system. High reliability and fault tolerance are achieved using multichannel systems due to the redundant system architecture. When a station transmits unsuccessfully due to a faulty channel, the problem can be solved if the next(re)transmission occurs on another channel.

1.7 Evaluation Metrics

There are many issues that must be considered with multiple access techniques, including: (1) throughput - the percentage of the capacity used to successfully transmit data over the channel, (2) delay - the time it takes for a packet to successfully reach the intended receiver, and (3) stability - i.e., whether an increasing number of users are attempting to access the channel but a decreasing number are actually succeeding. A system is kth order stable if the first k moments of the delay of a randomly chosen packet are finite. For queueing systems, stability is defined as the requirement that under steady state, the average total delay is finite [40].

It should be clear that a desirable multiaccessing system is one that is at least first-

order stable, has high average throughput, and has low accessing delay. The following subsections briefly describe issues of delay, throughput, queue length, capacity, and loss probability.

1.7.1 Delay

The average total delay per packet is defined as the sum of the average waiting time in the buffer and the average time spent from the first attempted transmission to the final successful one [40]. It is normally not possible to reduce the transmission time of a packet. There are additional sources of delay, such as queueing time or communications scheduling, that contribute to the overall delay, which we try to minimize. While delay depends on the parameters of the system and the distribution of traffic, we can manipulate some of these system parameters to reduce the overall delay incurred by the traffic. Evaluation of the mean delay, or waiting time, is often through plotting it as a function of traffic intensity, throughput, and offered load. The units of delay are usually normalized with respect to the ideal capacity.

1.7.2 Throughput

Throughput is defined as the amount of traffic that is successfully transmitted to its intended destination. The throughput is equivalent to the offered load, which is the amount of traffic actually transmitted, under the ideal situation of an error free channel. The maximum throughput is the same as the system capacity when there is no waste in the system.

1.7.3 Queue Length

The queue length is another important parameter to evaluate. Queues in a network will form at points of congestion given that there are waiting facilities within the communication system. The required length of a buffer can be estimated if the mean queue length can be predicted. Of course, the true nature of traffic is unpredictable and may vary from the mean. Still, the queue length is parameter of interest that could affect other system parameters.

1.7.4 Capacity

The capacity is defined as a measure for the amount of traffic that the system or communication link can handle.

1.7.5 Loss Probability

Loss probability is defined as the chance that traffic is lost. The loss of traffic can result from packet collisions, packets arrive at a full buffer, or packets arriving at a system with no buffer.

1.8 Introduction to the Following Chapters

The topic of multiple access communication has been introduced in the previous sections. We have described two fundamental classes of multiplexing techniques. Fixed access protocols allow a user's transmission on the communication channel to be isolated from all other users while random access protocols allow many users to contend for access with a possibility of packet collisions. Random access techniques also allow for the efficient transmission of messages from bursty sources, typical of data communications. However, fixed access techniques offer high capacity utilization.

We use the topics introduced in this chapter to develop architectures with the benefits of fixed access and random access techniques. This is in response to the need to use efficient transmission techniques to expand both present and future wireless communication networks. We will analyze the performance of many uplink architectures in Chapter 2. We are particularly interested in evaluating various multiple access techniques by plotting the average delay versus throughput. While the system parameters have been designed to focus on wireless networks, the analysis that is developed is applicable to many other communication systems.

In Chapter 3, we look at prioritization and distribution of channel capacity for different classes of users. We examine traditional priority schemes for multiple classes of users. We develop methodology for optimizing the distribution of channel capacity given to each user class based on certain time metrics and system constraints. Finally, in Chapter 4, we provide more conclusions and insights as to the architectural considerations when designing future multichannel communication systems.

Chapter 2

Large Message Transmissions

2.1 Introduction

System architecture design, evaluation, and optimization are key issues to developing communication systems that meet the requirements of today and expectations of the future. The design of next-generation communication systems, influenced by present-day systems, need to consider future communication services. The future of wireless promises to provide an enormous range of mobile services to users via a range of mobile terminals that enable the use of the cellular telephone in almost any location, indoor or outdoor. Third-generation mobile communication systems are being designed to support a large variety of services, most of which are not known yet. The air interface must be able to handle variable and asymmetric bit rates, up to 2 Mbps, with different quality of service requirements (bit error probability and delay) such as multimedia services with bandwidth on demand [34]. Effective packet access protocols are also needed to cope with bursty real time and non-real time data.

In this chapter, we will evaluate the performance of several uplink architectures for a wireless communication system[1]. We will consider traditional protocols such as TDMA and S-ALOHA alongside a channelized architecture, where the total channel capacity is divided into a fixed number of equal-capacity subchannels. We will develop the overall delay for sending large messages of a fixed length arriving from bursty sources, a typical scenario of users needing to send large files to a destination. This scenario is typical of a user needing to send a large file. We will make real-world parameter assumptions in the context of wireless networks and analyze the performance to develop intuition about the effectiveness of the different architectures.

[1]This chapter is based on work in [16].

2.2 Scheduling Algorithms

In this section, we describe two scheduling algorithms, namely First-Come First-Serve and Round Robin. Our objective is to find a scheduling algorithm that exhibits good delay performance for fixed-sized messages. The majority of scheduling algorithms consider messages based on message characteristics such as length, the amount of time that the message has been in the system, and the amount of service so far received by the message. With static scheduling, message priority does not change with time. Dynamic scheduling algorithms however can change a message's priority valued based on the amount of time that the message has spent in the system or the amount of service already received by the message. Additionally, scheduling algorithms can be non-preemptive or preemptive, where message transmissions can be interrupted by messages of higher priority.

2.2.1 First-Come First-Serve

A simple non-preemptive static scheduling algorithm is the first-come first-serve (FCFS) algorithm. Messages are served in order of arrival. This simple algorithm has the shortcoming of being unfair to short messages, but we need not worry about the fairness issue in this chapter as we are studying fixed-length messages. FCFS can be simply modeled with queueing systems described in Appendix A. The average message delay for a system using FCFS scheduling with fixed-length messages, i.e., M/D/1 queueing system, given in closed-form is

$$D_{FCFS} = \frac{\lambda \overline{X^2}}{2(1-\rho)} + \overline{X}, \tag{2.1}$$

where $\overline{X}$ is the message service time (e.g., transmission time), λ is the message arrival rate and ρ is the channel utilization. Channel utilization is defined as

$$\rho = \frac{\lambda}{\mu}, \tag{2.2}$$

where μ is the service rate. Recall that the service rate is defined as the inverse of the service time. The service time is the length of the packet in bits divided by the transmission bit rate, $\overline{X} = \frac{L}{C}$. For fixed service times, $\overline{X^2} = \overline{X}^2$. Equation 2.1 therefore describes the average message delay in terms of the first and second

moments of the message transmission times and can be rewritten as

$$\begin{aligned} D_{FCFS} &= \frac{\lambda \overline{X^2}}{2(1-\rho)} + \overline{X} \\ &= \frac{\lambda L^2}{C^2(2-2\frac{\lambda L}{C})} + \frac{L}{C}. \end{aligned} \tag{2.3}$$

The significant observation to be made in this case is that FCFS offers no fairness or discrimination among jobs on the basis of their required service time $\overline{X}$.

2.2.2 Round Robin

A round robin algorithm can overcome the shortcoming of FCFS. In this scheme, messages are generally divided into smaller packet sizes and are served one packet at a time in a round-robin fashion, rotating among the messages. Short messages no longer need to wait for the complete transmission of a long message. However, round robin scheduling can result in large delays when there are many messages in the system. The average message delay for round robin scheduling can be described in the closed-form

$$\begin{aligned} D_{RR} &= \frac{\overline{X}}{1-\rho} + \overline{X} \\ &= \frac{L}{C(2-2\frac{\lambda L}{C})} + \frac{L}{C}, \end{aligned} \tag{2.4}$$

where $\overline{X}$ is the message transmission time and ρ is the channel utilization. The fairness of round robin scheduling can be immediately seen from Equation 2.4. A job twice as long will spend on the average twice as long in the system, thus the discrimination is linear. For more information about the properties of round-robin scheduling, refer to [20]. Since we will be addressing fixed-length message transfers, round robin scheduling will not be further discussed.

2.3 System Model

Evaluating throughput performance and time delay is important to designing the proper network architecture for different communication scenarios. In this section we will study several wireless multichannel systems with different access protocols. We will compare these systems to their single channel system counterparts. We develop mathematical models that describe the overall delay for sending large messages of

a fixed length arriving from bursty sources and analyze their performances. We will make real-world parameter assumptions in the context of wireless networks and analyze the performance to develop intuition about the effectiveness of the different architectures.

2.3.1 Message Arrival Model

To begin our analysis, messages are presumed to be generated by an infinite set of users at a composite finite rate of λ packets per time slots. Messages are assumed to be a fixed length of L [bits]. If the user population is finite, the binomial arrival model would be used. However, the results would approach those of the Poisson model for a large set of users [35].

2.4 Architecture and Delay Models

We will develop equations for average delays for transmitting messages on six different uplink architectures. Denote

$$\begin{aligned} \lambda &= \text{Message arrival rate} \\ N &= \text{Number of users in system} \\ T_f &= \text{Frame length} \\ \overline{X} &= \text{Message service time} = \frac{L}{C} \\ T_{PD} &= \text{Propagation delay} \end{aligned}$$

Table 2.1 lists the section reference to each of the six different uplink architectures.

Architecture Model		Section
Fixed Access		
	TDMA 1 Channel	2.4.1.1
	TDMA 10 Channels	2.4.1.2
Random Access		
	S-ALOHA 1 Channel	2.4.2.1
	S-ALOHA 10 Channels Case 1	2.4.2.2
	S-ALOHA 10 Channels Case 2	2.4.2.3
Channelized Architecture		
	TDMA Reservation + 9 Channels	2.4.3.1.1
	TDMA Reservation + 1 Channel	2.4.3.1.2
	S-ALOHA Reservation + 9 Channels	2.4.3.2.1
	S-ALOHA Reservation + 1 Channel	2.4.3.2.2

Table 2.1: Uplink Architectures

2.4.1 Fixed Access

2.4.1.1 TDMA 1 Channel

This system implements the TDMA for all users on a single channel with total capacity C [bits/sec]. With a total channel capacity of C [bits/sec], and a large message of L [bits] and packet size $l_p = L$ [bits], the overall expected transmission delay becomes

$$\begin{aligned} T_{TDMA_{Case1}} &= \frac{T_f}{2} + Q_{M/D/1}(\lambda, \frac{1}{\overline{X}_{TDMA}}) + \overline{X}_{TDMA} + 2T_{PD} \\ &= \frac{1}{2}\frac{NL}{C} + \frac{\lambda L^2}{C^2(2 - 2\frac{\lambda L}{C})} + \frac{L}{C} + 2T_{PD}. \end{aligned} \tag{2.5}$$

2.4.1.2 TDMA 10 Channels

This system implements the TDMA for all users on a multichannel system. Each of the K channels has equal capacity of $\frac{C}{K}$ [bits/sec]. The user population is divided into K equally sized groups with each group accessing one channel. The transmission time of the message on a data channel thus becomes

$$\overline{X}_{TDMA,Case2} = \frac{LK}{C}. \tag{2.6}$$

With a total channel capacity of C [bits/sec], and a large message of L bits and packet size $l_p = L$, the overall expected transmission delay becomes

$$\begin{aligned} T_{TDMA_{Case2}} &= \frac{T_f}{2} + Q_{M/D/k}(\lambda, \frac{1}{\overline{X}_{TDMA,Case2}}, \frac{1}{\overline{X^2}_{TDMA,Case2}}, K) + \overline{X}_{TDMA,Case2} + 2T_{PD} \\ &= \frac{1}{2}\frac{NLK}{C} + \frac{\lambda^K \overline{X^2}(\overline{X})^{K-1}}{2(K-1)!(K-\lambda\overline{X})^2(\sum_{n=0}^{K-1}\frac{(\lambda\overline{X})^n}{n!} + \frac{(\lambda\overline{X})^K}{(K-1)!(K-\lambda\overline{X})})} + \frac{LK}{C} + 2T_{PD}. \end{aligned} \tag{2.7}$$

2.4.2 Random Access

2.4.2.1 S-ALOHA 1 Channel

This system implements the S-ALOHA protocol for all users on a single channel with total capacity C [bits/sec]. As in Section 1.5.2, assuming an infinite user population with a total channel capacity of C [bits/sec], and a large message of L [bits] and packet size $l_p = L$ [bits], the overall expected transmission delay is

$$\begin{aligned} T_{S-ALOHA_{Case1}} &= \frac{\tau}{2} + \tau + E[r]E[T_c] + 2T_{PD} \\ &\approx \frac{3}{2}\frac{l_p}{C} + [e^G - 1][\frac{(H+2)\tau}{2} + 2T_{PD}] + 2T_{PD} \text{ for } H \gg 1, \end{aligned} \tag{2.8}$$

where $\tau = \frac{l_p}{C}$.

2.4.2.2 S-ALOHA 10 Channels Case 1

This system implements the S-ALOHA protocol for all users on a multichannel system. Each of the K channels has equal capacity of $\frac{C}{K}$ [bits/sec]. The user population is divided into K equally sized groups with each group accessing one channel. With a total channel capacity of C [bits/sec], and a large message of L [bits] and packet size $l_p = L$ [bits], the overall expected transmission delay becomes

$$\begin{aligned} T_{S-ALOHA_{Case2}} &= \frac{\tau}{2} + \tau + E[r]E[T_c] + 2T_{PD} \\ &\approx \frac{3}{2}\frac{Kl_p}{C} + [e^G - 1][\frac{(H+2)\tau}{2} + 2T_{PD}] + 2T_{PD} \text{ for } H \gg 1, \end{aligned} \tag{2.9}$$

where $\tau = K\frac{l_p}{C}$.

2.4.2.3 S-ALOHA 10 Channels Case 2

This system implements the S-ALOHA protocol for all users on a multichannel system. Each of the K channels has equal capacity of $\frac{C}{K}$ [bits/sec]. Every user and for every packet transmission (or retransmission) selects randomly, uniformly, and independently of the past the channel over which this specific packet is to be transmitted. Both the S-ALOHA 10 Channel cases yield the same average delay since the message arrival processes are Poisson. In order to understand this, see Appendix B.1. For valid comparisons with TDMA 10 Channels, we only evaluate S-ALOHA 10 Channels Case 1.

2.4.3 Channelized Uplink

The channelized uplink architecture offers channel resources to multiple users by first dividing the transmission medium into K equal-capacity channels. These channels can then be assigned to users on a need basis, i.e. demand assignment. One channel is designated as the reservation channel. The remaining channel resources, used for transmitting data, are given to the users in a FCFS discipline upon availability. Access to the data channels is controlled by the scheduler. The reservation channel, on the other hand, can use any multiple access scheme. We will focus on TDMA and S-ALOHA techniques for the reservation channel.

2.4.3.1 Fixed Access Reservation

We assume a multichannel system with K equal capacity channels of $\frac{C}{K}$ [bits/sec]. One of these channels is dedicated for reservation packets while the remaining channel resources are used for transmitting data messages. This system used a fixed access protocol such as TDMA on the reservation channel, where each user is assigned a time slot in a reservation frame. We assume that there are N users, each user's reservation packet is of length l_r [bits] and the channel capacity of the reservation channel is $C_r = \frac{C}{K}$ [bits/sec]. The channel is assumed to be error free. A user who wants access to the data channels must first send a reservation packet within his given time slot. The expected setup time for sending the reservation packet is thus

$$\begin{aligned} D_{SU,TDMA}(\lambda, l_r, N, C) &= D_{TDMA}(\lambda, l_r, N, \frac{C}{K}) \\ &= \frac{1}{2}\frac{NKl_r}{C} + \frac{\lambda K^2 (l_r)^2}{C^2(2 - 2\frac{\lambda K l_r}{C})} + \frac{Kl_r}{C} + 2T_{PD}. \end{aligned} \quad (2.10)$$

2.4.3.1.1 TDMA Reservation + 9 Channels

This system implements the channelized uplink architecture with the TDMA protocol for accessing the reservation channel. In this case, we assume that there are $K = 10$ equal capacity subchannels. Since one channel is used for reservation messages, there are $K - 1 = 9$ data channels remaining, each with capacity $\frac{C}{K}$ [bits/sec]. Access to these data channels is controlled by the scheduler and the earliest available channel is given to the users in a FCFS discipline. The transmission time of the actual message, once the queue is available is simply

$$\overline{X}_{TDMA,FCFS_{Case1}} = \frac{LK}{C}. \tag{2.11}$$

The service time is the length of the packet in bits divided by the transmission bit rate. When the packet length is random, the service time is also random. If the packets are a fixed length, then the service time is deterministic. Thus, the expected wait for an available channel appears as an M/D/k queue with $(K-1)$ channels each operating at $\frac{C}{K}$ [bits/sec] to handle message arrivals of rate λ.

By combining the appropriate elements described above, the overall expected message delay for either reservation discipline is

$$\begin{aligned} T_{TDMA,FCFS_{Case1}} &= D_{SU,TDMA} + Q_{M/D/k}(\lambda, \frac{1}{\overline{X}_{TDMA,FCFS_{Case1}}}, K-1) \\ &\quad + \overline{X}_{TDMA,FCFS_{Case1}} + 2T_{PD} \\ &= \frac{1}{2}\frac{NKl_r}{C} + \frac{\lambda K^2 (l_r)^2}{C^2(2 - 2\frac{\lambda K l_r}{C})} + \frac{Kl_r}{C} + 2T_{PD} \\ &\quad + \frac{\lambda^K \overline{X^2}(\overline{X})^{K-1}}{2(K-1)!(K-\lambda\overline{X})^2(\sum_{n=0}^{K-1}\frac{(\lambda\overline{X})^n}{n!} + \frac{(\lambda\overline{X})^K}{(K-1)!(K-\lambda\overline{X})})} \\ &\quad + \frac{LK}{C} + 2T_{PD}. \end{aligned} \tag{2.12}$$

2.4.3.1.2 TDMA Reservation + 1 Channel

In this case, we use a subchannel of capacity $\frac{C}{K}$ [bits/sec], where $K = 10$, for transmitting reservation messages. The remaining capacity of $\frac{K-1}{K}C$ [bits/sec] is used as one subchannel for data transmission. Access to the data channel is controlled by the scheduler and given to the users in a FCFS discipline. The transmission time of the

message on the data channel is

$$\overline{X}_{TDMA,FCFS_{Case2}} = \frac{LK}{(K-1)C}. \tag{2.13}$$

The service time is the length of the packet in bits divided by the transmission bit rate. When the packet length is random, the service time is also random. If the packets are a fixed length, then the service time is deterministic. Thus, the expected wait for an available channel appears as an M/D/1 queue with 1 channel operating at $\frac{K-1}{K}C$ [bits/sec] to handle message arrivals of rate λ.

By combining the appropriate elements described above, the overall expected message delay for either reservation discipline is

$$\begin{aligned} T_{TDMA,FCFS_{Case2}} &= D_{SU,TDMA} + Q_{M/D/1}(\lambda, \frac{1}{\overline{X}_{TDMA,FCFS_{Case2}}}) \\ &\quad + \overline{X}_{TDMA,FCFS_{Case2}} + 2T_{PD} \\ &= \frac{1}{2}\frac{NKl_r}{C} + \frac{\lambda K^2 (l_r)^2}{C^2(2 - 2\frac{\lambda K l_r}{C})} + \frac{Kl_r}{C} + 2T_{PD} \\ &\quad + \frac{\lambda L^2 K^2}{(K-1)^2 C^2 (2 - 2\frac{\lambda LK}{(K-1)C})} + \frac{LK}{(K-1)C} + 2T_{PD}. \end{aligned} \tag{2.14}$$

2.4.3.2 Random Access Reservation

Packet contention in the reservation channel can be seen with the implementation of a random access protocol. As seen in the previous chapter, S-ALOHA has superior performance compared to pure ALOHA. Thus, we will continue our analysis considering only S-ALOHA as the random access protocol. We assume a multichannel system with K equal capacity channels of $\frac{C}{K}$ [bits/sec]. Again we assume that the channels are error-free. One of these channels is dedicated for reservation packets while the remaining channel resources are used for transmitting data messages. After sending a reservation packet, users gain access to a data channel on a FCFS basis.

The random access reservation with S-ALOHA has an expected setup time of

$$\begin{aligned} D_{SU,S-ALOHA}(l_r, C, K, G) &= T_{S-ALOHA}(l_r, C, K, G) \\ &= \frac{\tau}{2} + \tau + E[r]E[T_c] + 2T_{PD} \\ &\approx \frac{3Kl_r}{2C} + [e^G - 1][\frac{(H+2)\frac{Kl_r}{C}}{2} + 2T_{PD}] + 2T_{PD}, \end{aligned} \tag{2.15}$$

where $\tau = \frac{Kl_r}{C}$. Unlike using a fixed access scheme on the reservation channel, a user using random access may need to resend its reservation packet if packet collisions occur.

2.4.3.2.1 S-ALOHA Reservation + 9 Channels

This system implements the channelized uplink architecture with the S-ALOHA protocol for accessing the reservation channel. We assume a multichannel system with K equal capacity channels of $\frac{C}{K}$ [bits/sec], where $K = 10$. One of these channels is dedicated for reservation packets while the other nine are used for transmitting data messages. After sending a reservation packet, users gain access to a data channel on a FCFS basis. The transmission time of the actual message, once the queue is available is simply

$$\overline{X}_{S-ALOHA,FCFS_{Case1}} = \frac{LK}{C}. \tag{2.16}$$

The service time is the length of the packet in bits divided by the transmission bit rate. When the packet length is random, the service time is also random. If the packets are a fixed length, then the service time is deterministic. Thus, the expected wait for an available channel appears as an M/D/k queue with $(K-1)$ channels each operating at $\frac{C}{K}$ [bits/sec] to handle message arrivals of rate λ.

By combining the appropriate elements described above, the overall expected

message delay for either reservation discipline is

$$\begin{aligned} T_{S-ALOHA,FCFS_{Case1}} &= D_{SU,S-ALOHA} + Q_{M/D/k}(\lambda, \frac{1}{\overline{X}_{S-ALOHA,FCFS_{Case1}}}, K-1) \\ &\quad + \overline{X}_{S-ALOHA,FCFS_{Case1}} + 2T_{PD} \\ &\approx \frac{3Kl_r}{2C} + [e^G - 1][\frac{(H+2)\frac{Kl_r}{C}}{2} + 2T_{PD}] + 2T_{PD} \\ &\quad + \frac{\lambda^K \overline{X^2} (\overline{X})^{K-1}}{2(K-1)!(K-\lambda\overline{X})^2(\sum_{n=0}^{K-1} \frac{(\lambda\overline{X})^n}{n!} + \frac{(\lambda\overline{X})^K}{(K-1)!(K-\lambda\overline{X})})} \\ &\quad + \frac{LK}{C} + 2T_{PD}. \end{aligned} \tag{2.17}$$

2.4.3.2.2 S-ALOHA Reservation + 1 Channel

In this case, we use a subchannel of capacity $\frac{C}{K}$ [bits/sec], where $K = 10$, for transmitting reservation messages. The remaining capacity of $\frac{K-1}{K}C$ [bits/sec] is used as one subchannel for data transmission. Access to the data channel is controlled by the scheduler and given to the users in a FCFS discipline. The transmission time of the actual message, once the queue is available is simply

$$\overline{X}_{S-ALOHA,FCFS_{Case2}} = \frac{LK}{(K-1)C}. \tag{2.18}$$

The service time is the length of the packet in bits divided by the transmission bit rate. When the packet length is random, the service time is also random. If the packets are a fixed length, then the service time is deterministic. Thus, the expected wait for an available channel appears as an M/D/1 queue with 1 channel each operating at $\frac{K-1}{K}C$ [bits/sec] to handle message arrivals of rate λ.

By combining the appropriate elements described above, the overall expected

message delay for either reservation discipline is

$$\begin{aligned} T_{S-ALOHA,FCFS_{Case2}} &= D_{SU,S-ALOHA} + Q_{M/D/k}(\lambda, \frac{1}{\overline{X}_{S-ALOHA,FCFS_{Case2}}}, K-1) \\ &\quad + \overline{X}_{S-ALOHA,FCFS_{Case2}} + 2T_{PD} \\ &\approx \frac{3Kl_r}{2C} + [e^G - 1][\frac{(H+2)\frac{Kl_r}{C}}{2} + 2T_{PD}] + 2T_{PD} \\ &\quad + \frac{\lambda L^2 K^2}{(K-1)^2 C^2 (2 - 2\frac{\lambda LK}{(K-1)C})} + \frac{LK}{(K-1)C} + 2T_{PD}. \end{aligned} \quad (2.19)$$

2.5 Analysis

2.5.1 Model Parameters

Now that we have developed the analytical expressions of several system models, it is necessary to evaluate these uplink architectures with a few examples of message delay performance. We will make real-world parameter assumptions in the context of wireless networks. Table 2.2 lists the parameters used in our model.

Parameter	Description	Units
λ	Composite message arrival time	[msgs/sec]
K	Number of channels	[channels]
N	Number of users in system	[users]
T_f	Frame length	[sec]
C	Total uplink channel capacity	[bits/sec]
l_r	Reservation request packet length	[bits]
L	Message length	[bits]
l_p	Packet length	[bits]
T_{PD}	Propagation delay	[sec]

Table 2.2: Model Parameters for Evaluating Large Message Transfer Performance.

2.5.1.1 Reservation Packet Length (l_r)

We assume that the length of the reservation packet, l_r is 1000 [bits]. Similar to the Internet Protocol (IP) header packet, the reservation packet contains source and

destination addresses, data size, and even message priority.

2.5.1.2 Message Length (L)

We assume that the transmitted data are fixed-length messages with a size of 1 [Mb].

2.5.1.3 Packet Length (l_p)

Large messages may be divided into smaller packets for transmission. To simplify our analysis, we consider the transmission of one large message. Thus, the packet length l_p is equal to the entire message of L [bits].

2.5.1.4 Propagation Delay (T_{PD})

The propagation delay, T_{PD}, is set to 5 [μsec], thus the round-trip propagation delay is 10 [μsec]. This propagation delay is typical for distances of 1.5 [km] in wireless local area network (LAN) systems.

2.5.1.5 Channelized Architecture Configuration

In our analysis, we will divide the total channel capacity evenly into $K = 10$ sub-channels. One channel is designated for reservation packets. The remaining channel resources are used for data transmissions. We will analyze two different scenarios. In Case 1, we have 9 channels, each of capacity $\frac{C}{10}$ [bits/sec] available for data transmission. In Case 2, we have 1 channel with capacity $\frac{9}{10}C$ [bits/sec] available for data transmission.

2.5.2 Delay Performance Figures

Since we have defined the auxiliary parameters of our model with values typical of wireless communication systems, we must consider the remaining parameters of total uplink capacity, user population, and message arrival rate. The values we will use are summarized in Table 2.3. Note that we specify the user population in all cases. This is only necessary for the fixed access schemes. Our mathematical models for the random access schemes assume an infinite population.

Capacity (C [bits/sec])	Users (N)	λ_{max} [msgs/sec]	Figure
10k	10	0.01	2-1
100k	100	0.1	2-2
1M	1000	1	2-3
10k	10	0.01	2-4
100k	100	0.1	2-5
1M	1000	1	2-6
10k	10	0.01	2-7
100k	100	0.1	2-8
1M	1000	1	2-9

Table 2.3: Table of Large Message Transmissions Performance Figures.

In the following delay performance plots, the relationship between utilization ρ and message arrival rate λ is defined as

$$\rho = \lambda \frac{L}{C}. \tag{2.20}$$

where L is the message length and C is the total channel capacity.

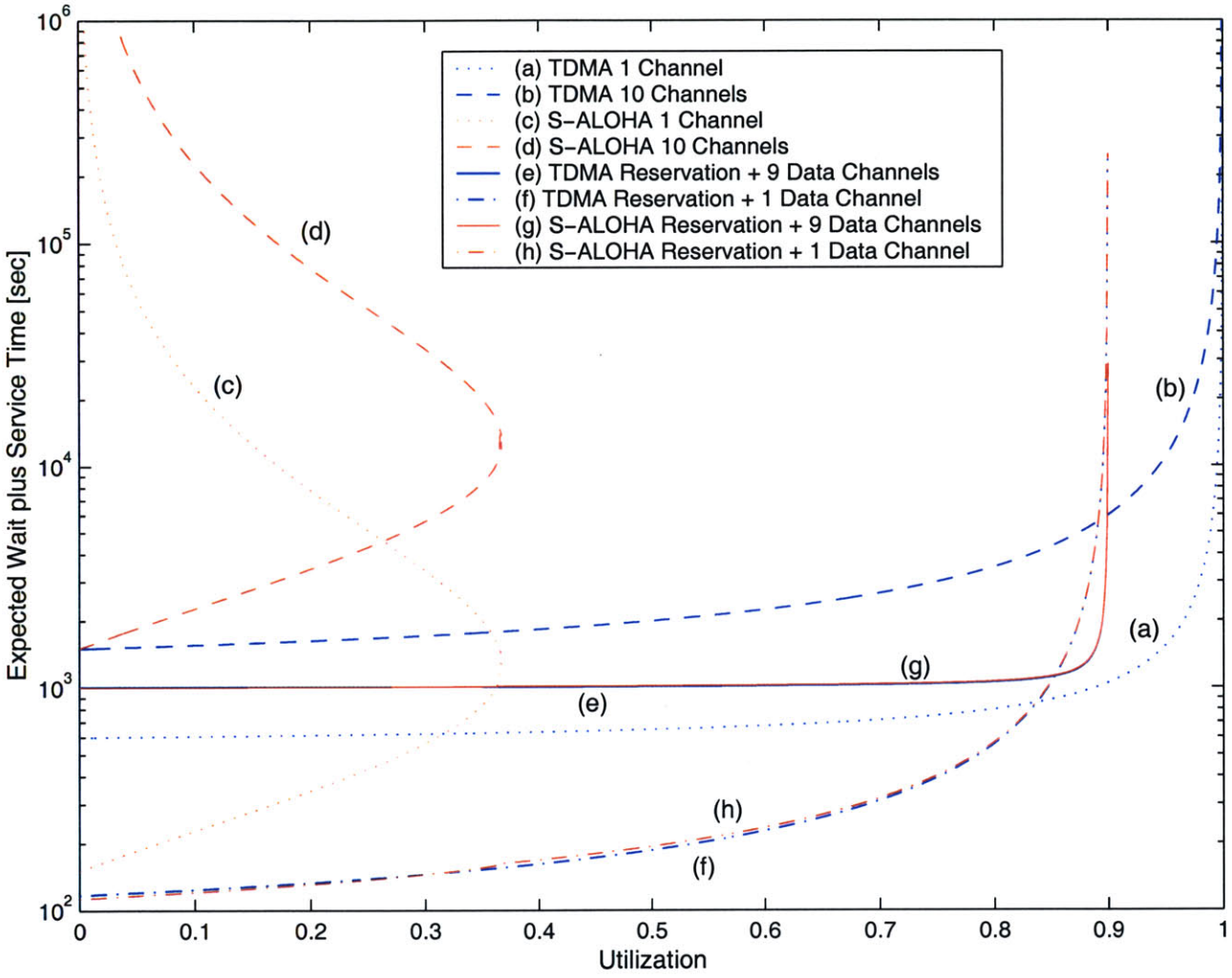


Figure 2-1: Performance Analysis 1.
L = 1 [Mb], N = 10, C = 10 [kbits/sec] with $0 \leq \lambda \leq 0.01$ [msgs/sec]

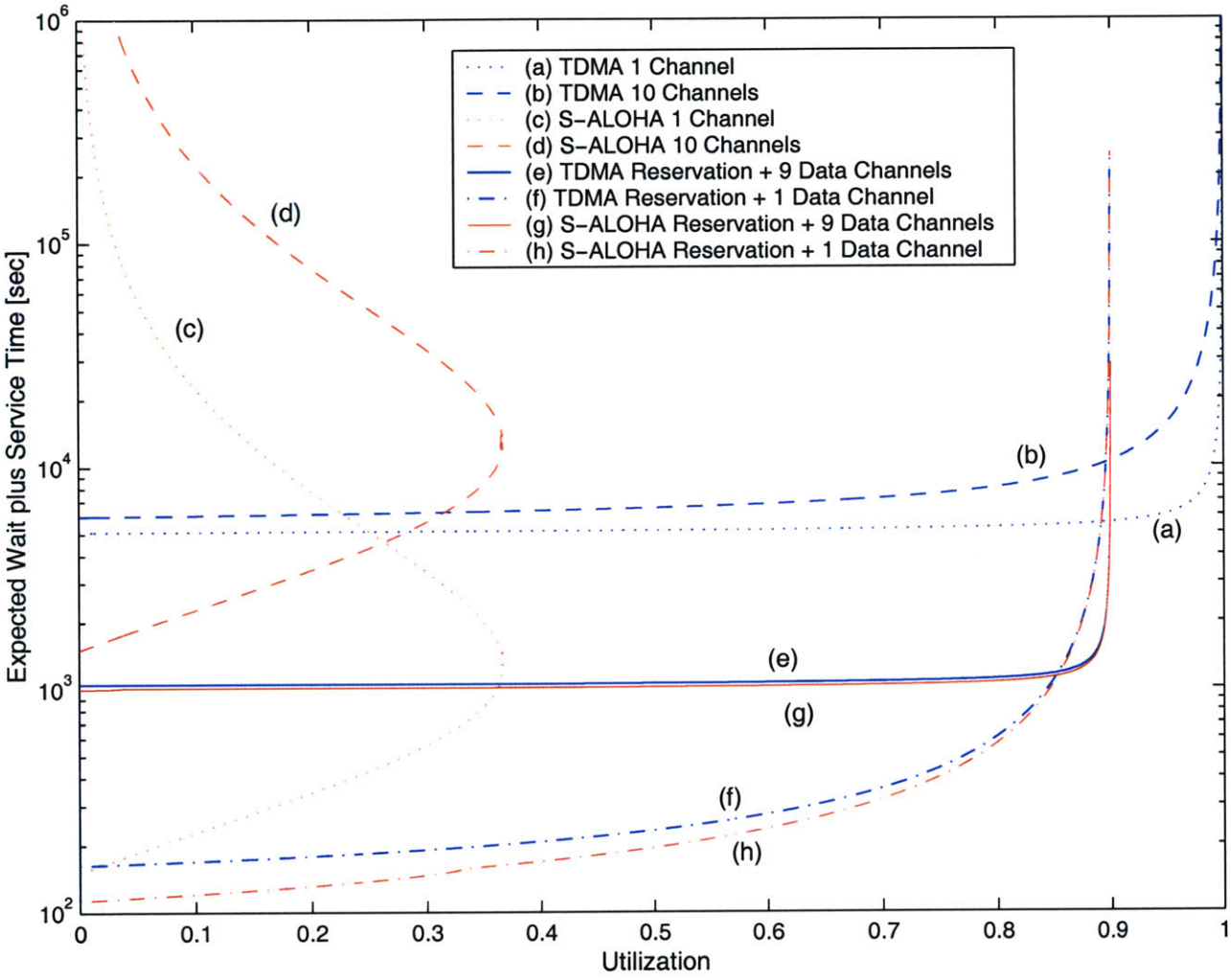


Figure 2-2: Performance Analysis 2.
L = 1 [Mb], N = 100, C = 10 [kbits/sec] with $0 \leq \lambda \leq 0.01$ [msgs/sec]

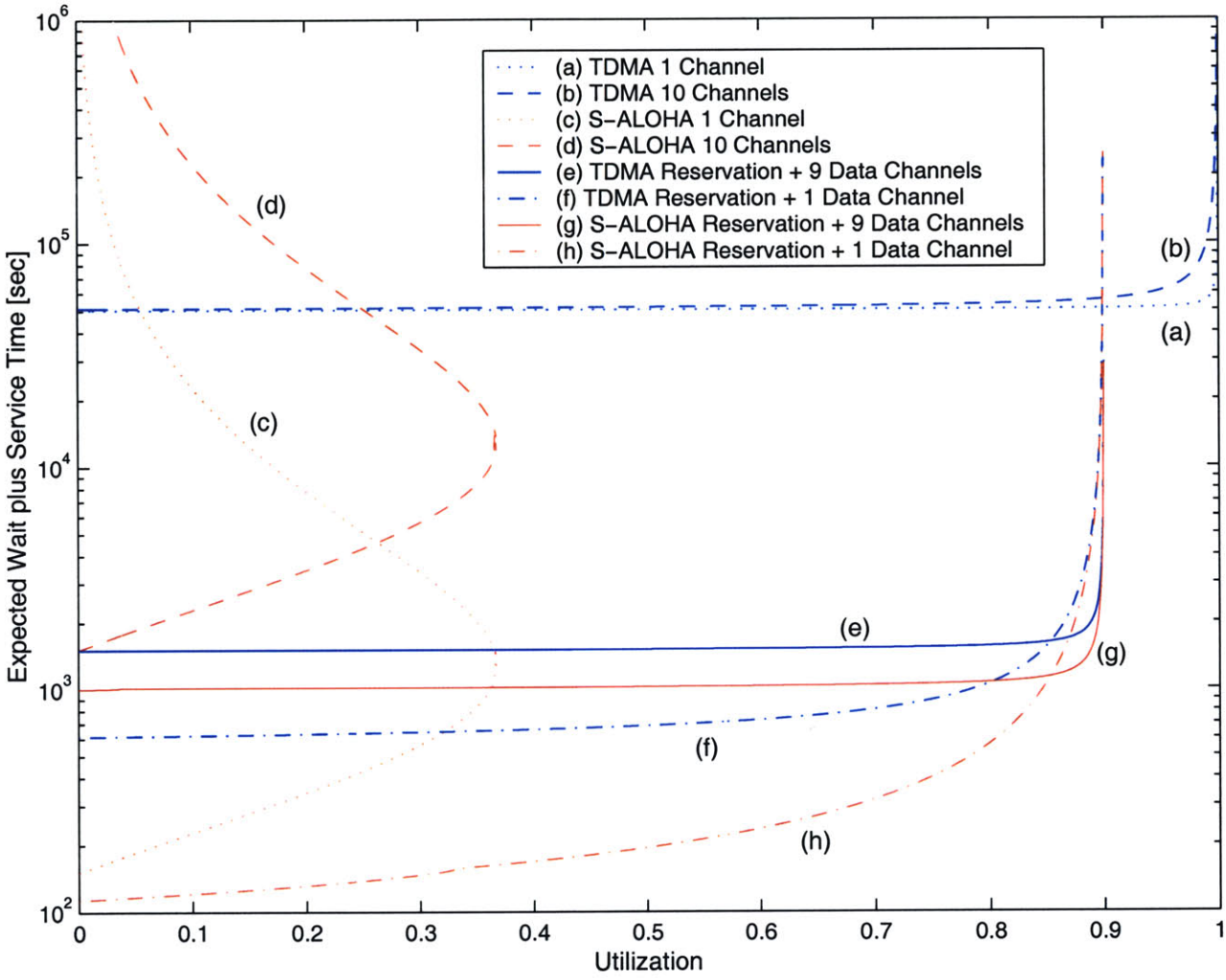


Figure 2-3: Performance Analysis 3.
L = 1 [Mb], N = 1000, C = 10 [kbits/sec] with $0 \leq \lambda \leq 0.01$ [msgs/sec]

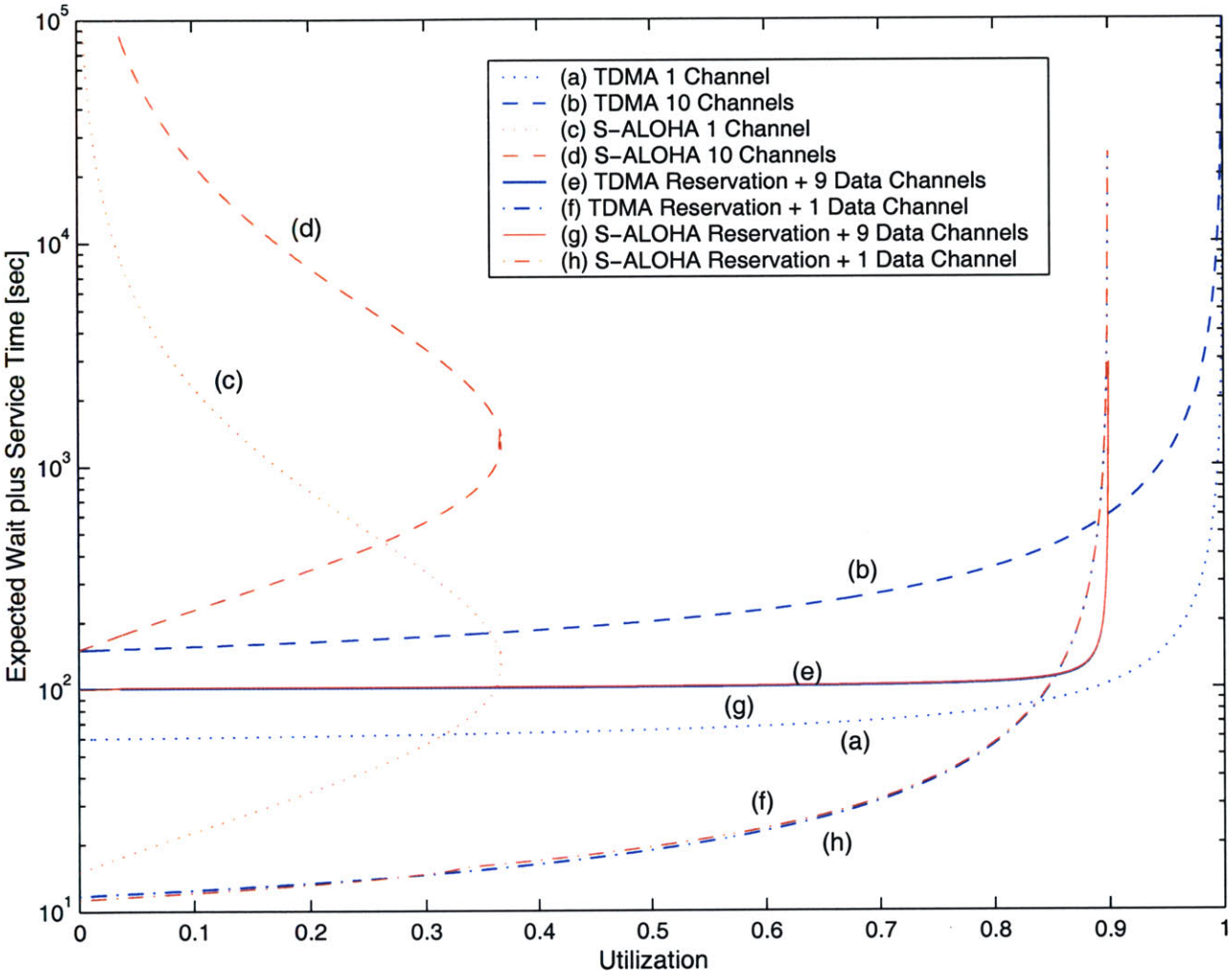


Figure 2-4: Performance Analysis 4.
L = 1 [Mb], N = 10, C = 100 [kbits/sec] with $0 \leq \lambda \leq 0.1$ [msgs/sec]

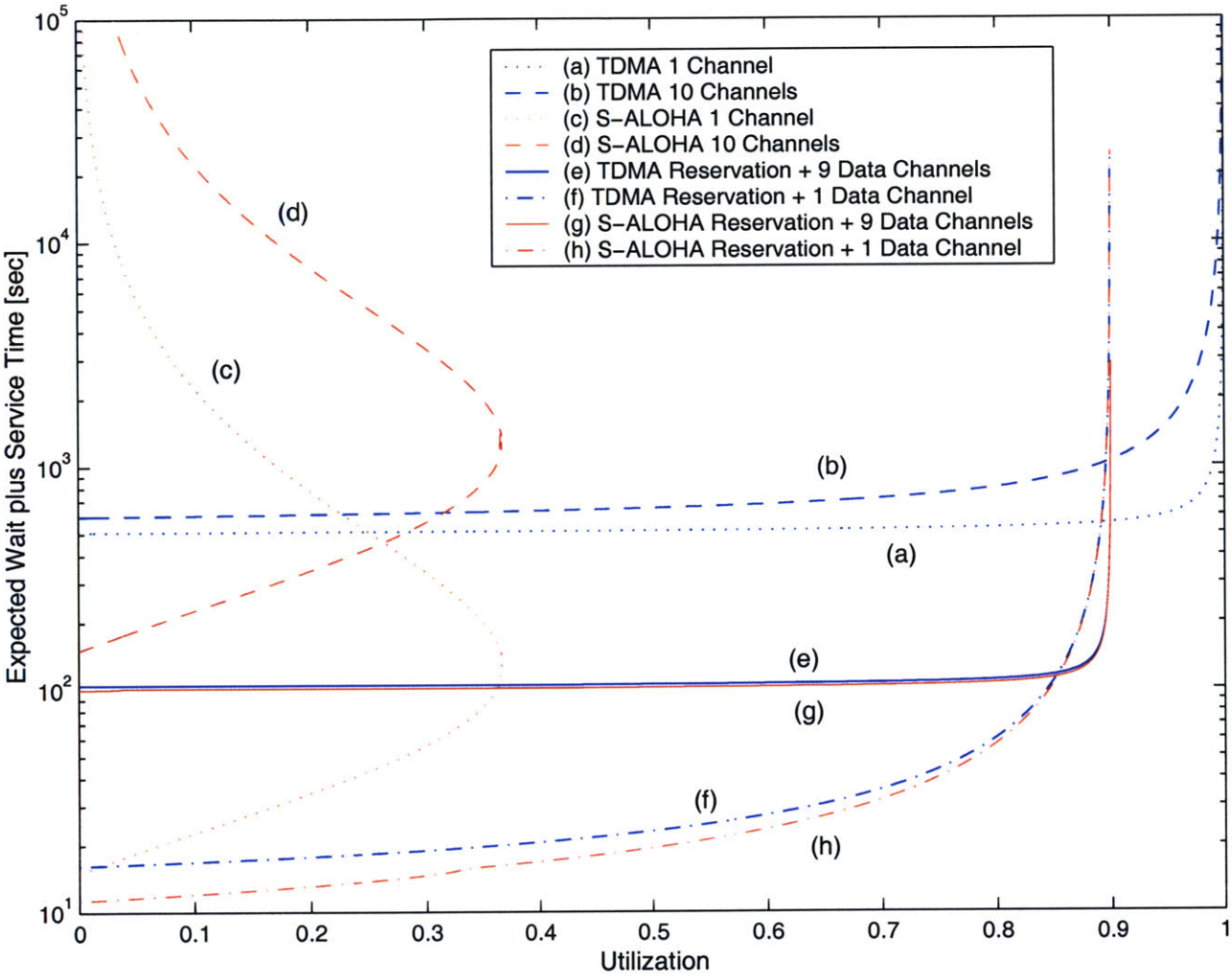


Figure 2-5: Performance Analysis 5.
L = 1 [Mb], N = 100, C = 100 [kbits/sec] with $0 \leq \lambda \leq 0.1$ [msgs/sec]

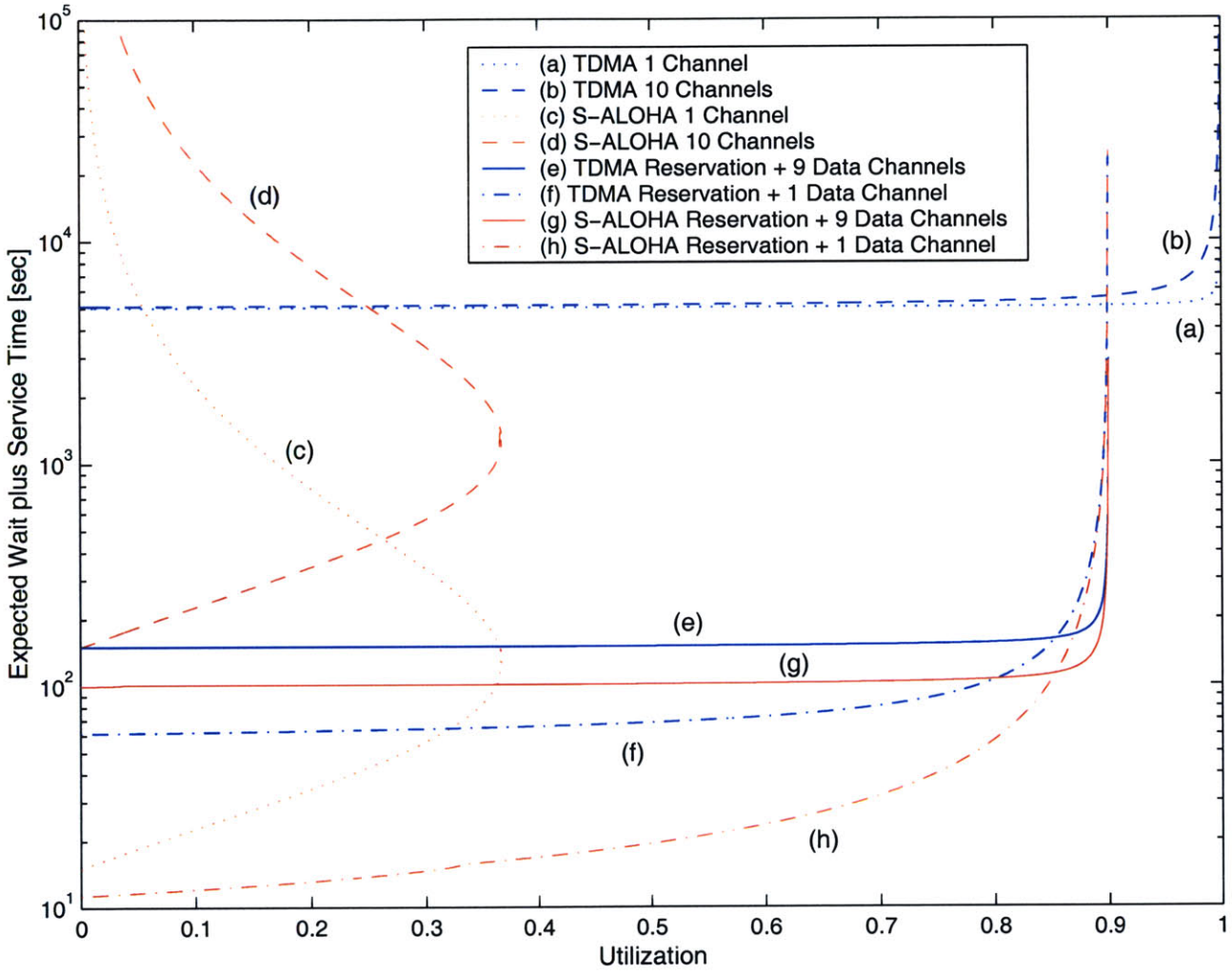


Figure 2-6: Performance Analysis 6.
L = 1 [Mb], N = 1000, C = 100 [kbits/sec] with $0 \leq \lambda \leq 0.1$ [msgs/sec]

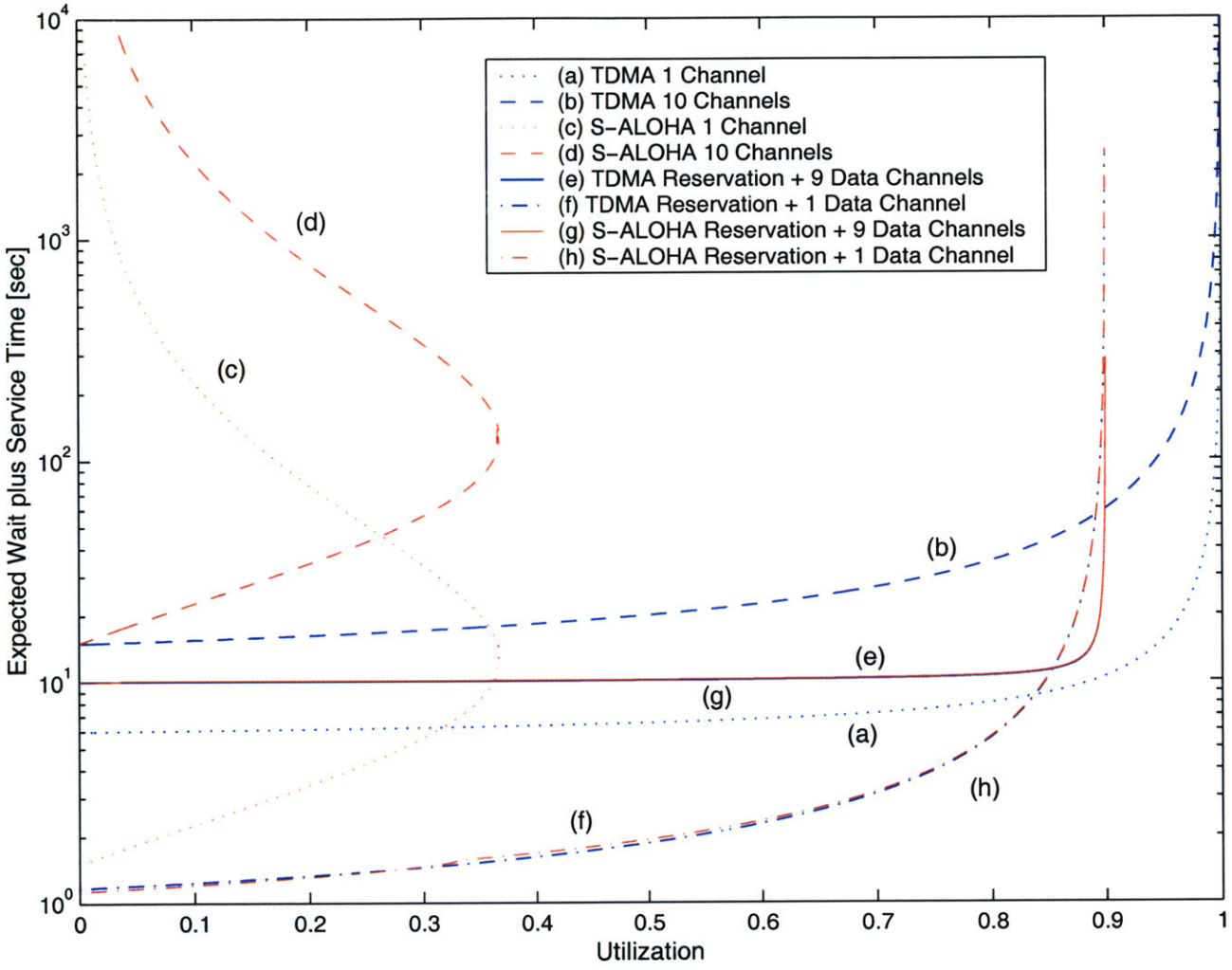


Figure 2-7: Performance Analysis 7.
L = 1 [Mb], N = 10, C = 1 [Mbits/sec] with $0 \leq \lambda \leq 1$ [msgs/sec]

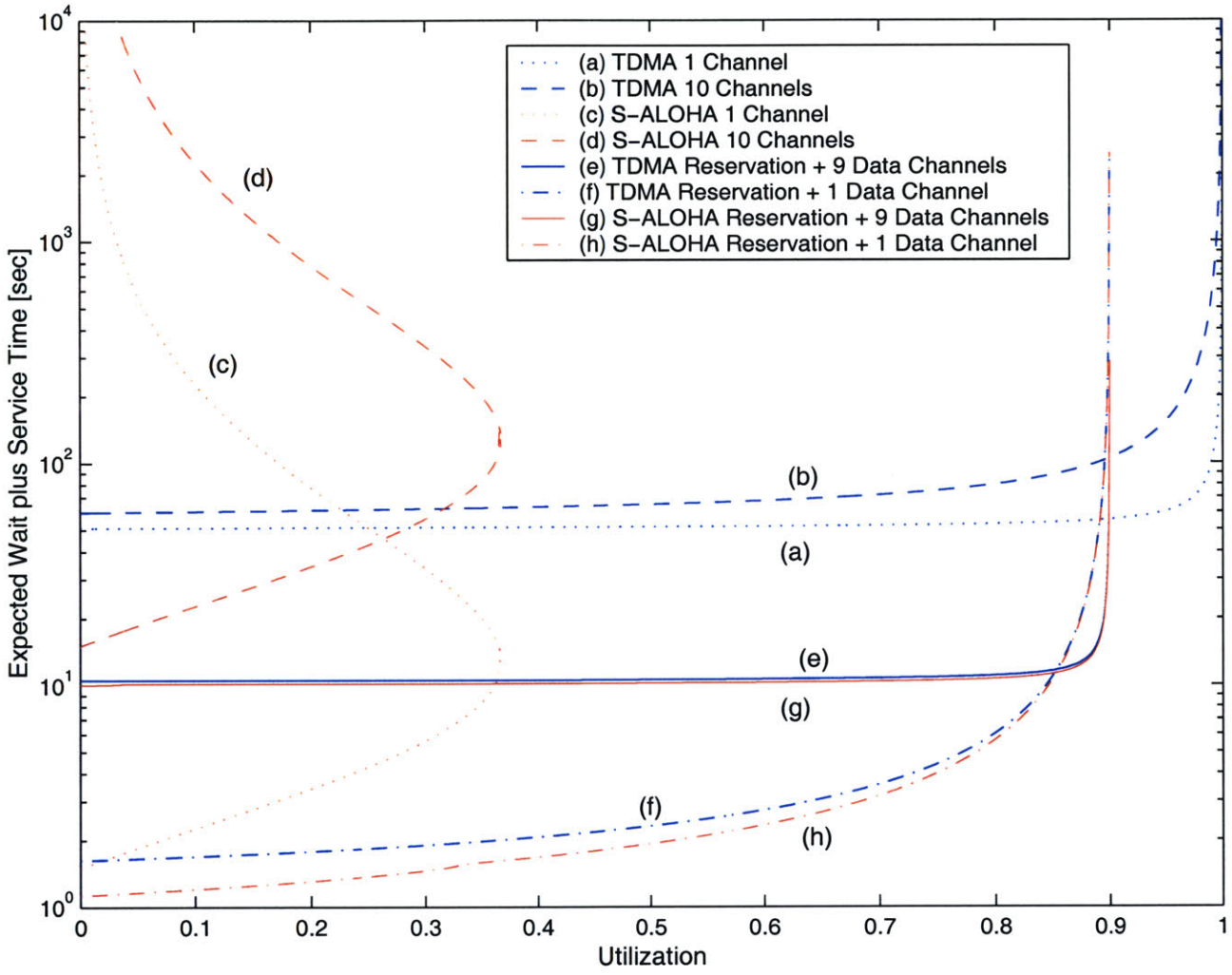


Figure 2-8: Performance Analysis 8.
L = 1 [Mb], N = 100, C = 1 [Mbits/sec] with $0 \leq \lambda \leq 1$ [msgs/sec]

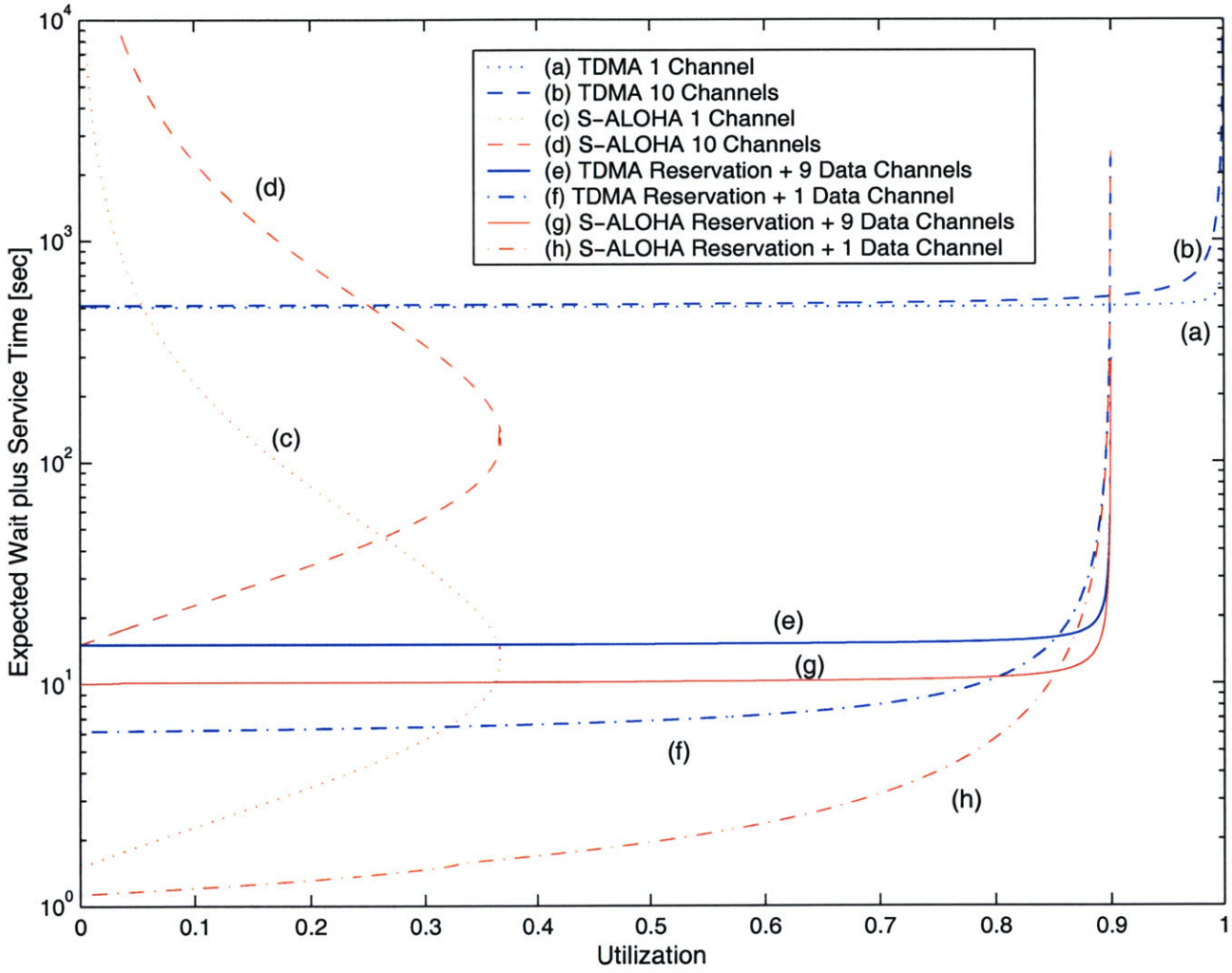


Figure 2-9: Performance Analysis 9.
L = 1 [Mb], N = 100, C = 1 [Mbits/sec] with $0 \leq \lambda \leq 1$ [msgs/sec]

2.6 Interpretation of Results

We will summarize the results that can be observed by studying the delay performance charts and comparing the various uplink architectures.

2.6.1 Fixed Access

TDMA results in the best capacity utilization under high traffic conditions. TDMA however wastes channel resources when traffic is low. When users are idle, their time slots are unused. The TDMA techniques shown, (a) and (b), generally have greater expected delays than any of the channelized architectures, (e)-(h). The difference is at least one order of magnitude.

2.6.1.1 TDMA 1 Channel

$$\begin{aligned} T_{TDMA_{Case1}} &= \frac{1}{2}\frac{NL}{C} + \frac{\lambda L^2}{C^2(2-2\frac{\lambda L}{C})} + \frac{L}{C} + 2T_{PD} \\ &= \frac{3}{2}\frac{L}{C} + \frac{\lambda L^2}{C^2(2-2\rho)} + 2T_{PD}. \end{aligned} \tag{2.21}$$

where $\rho = \frac{\lambda L}{C}$. Equation 2.21 is valid for $\lambda L < C$. The delay performance plots show that there is an vertical asymptote of $\rho = \frac{\lambda L}{C} = 1$. When the arrival rate of messages λ is very small, the delay due to queueing is insignificant. As λL approaches C, the queuing delay dominates the other delay components in the equation.

2.6.1.2 TDMA 10 Channels

$$\begin{aligned} T_{TDMA_{Case2}} &= \frac{1}{2}\frac{NLK}{C} + \frac{\lambda^K \overline{X^2}(\overline{X})^{K-1}}{2(K-1)!(K-\lambda\overline{X})^2(\sum_{n=0}^{K-1}\frac{(\lambda\overline{X})^n}{n!} + \frac{(\lambda\overline{X})^K}{(K-1)!(K-\lambda\overline{X})})} \\ &\quad + \frac{LK}{C} + 2T_{PD}. \end{aligned}$$

This case has the same characteristics as seen with TDMA 1 Channel. However, the average delay of a message transfer has increased since the channel capacity has been reduced thus increasing the service time by a factor of K.

2.6.2 Random Access

Contrast to TDMA, S-ALOHA results in the best expected delay performance under light traffic conditions. Since there are no fixed slots for users, S-ALOHA is a better technique for coping with traffic from bursty sources. However S-ALOHA systems suffer from instability and has a maximum achievable utilization of $\frac{1}{e}$.

2.6.2.1 S-ALOHA 1 Channel

$$T_{S-ALOHA_{Case1}} \approx \frac{3}{2}\frac{l_p}{C} + [e^G - 1][\frac{(H+2)\tau}{2} + 2T_{PD}] + 2T_{PD} \text{ for } H \gg 1.$$

Utilization reaches a maximum value of $\frac{1}{e}$ which is the maximum throughput of the S-ALOHA protocol.

2.6.2.2 S-ALOHA 10 Channels

$$T_{S-ALOHA_{Case2}} \approx \frac{3}{2}\frac{Kl_p}{C} + [e^G - 1][\frac{(H+2)\tau}{2} + 2T_{PD}] + 2T_{PD} \text{ for } H \gg 1.$$

This case has the same characteristics as with S-ALOHA 1 Channel. However, the average delay of a message transfer has increased since the channel capacity has been reduced thus increasing the service time by a factor of K.

2.6.3 Channelized Architecture

2.6.3.1 Fixed Access Reservation

The expected delays of the channelized architectures asymptotically reach 90% utilization which is consistent with our intuition and our mathematical models which state that 10% of channel resources are allocated for reservation packets with the remaining 90% allocated for data transmissions. Also note that the channelized architectures, (e) and (g), result in a constant expected delay for values up to approximately 70% channel utilization.

2.6.3.1.1 TDMA Reservation + 9 Channels

$$T_{TDMA,FCFS_{Case1}} = \frac{1}{2}\frac{NKl_r}{C} + \frac{\lambda K^2(l_r)^2}{C^2(2-2\frac{\lambda Kl_r}{C})} + \frac{Kl_r}{C} + 2T_{PD}$$
$$+\frac{\lambda^K \overline{X^2}(\overline{X})^{K-1}}{2(K-1)!(K-\lambda\overline{X})^2(\sum_{n=0}^{K-1}\frac{(\lambda\overline{X})^n}{n!} + \frac{(\lambda\overline{X})^K}{(K-1)!(K-\lambda\overline{X})})}$$
$$+\frac{LK}{C} + 2T_{PD}.$$

In this case, we allocate 1 subchannel for reservations using TDMA and 9 subchannels, each of capacity $\frac{1}{10}C$ [bits/sec] for data transmissions.

2.6.3.1.2 TDMA Reservation + 1 Channel

$$T_{TDMA,FCFS_{Case2}} = \frac{1}{2}\frac{NKl_r}{C} + \frac{\lambda K^2(l_r)^2}{C^2(2-2\frac{\lambda Kl_r}{C})} + \frac{Kl_r}{C} + 2T_{PD}$$
$$+\frac{\lambda L^2K^2}{(K-1)^2C^2(2-2\frac{\lambda LK}{(K-1)C})} + \frac{LK}{(K-1)C} + 2T_{PD}.$$

In this case, we allocate 1 subchannel for reservations using TDMA and 1 subchannel, with capacity $\frac{9}{10}C$ [bits/sec] for data transmission. This case has the same characteristics as with TDMA Reservation + 9 Channels. However, the average delay of a message transfer has decreased since the channel capacity for data transmission has been increased while the average time to gain access to a channel remains the same.

2.6.3.2 Random Access Reservation

The performance of the channelized architecture implementing a random access protocol on the reservation channel shows an improvement compared to the channelized architecture with fixed access reservation. This improvement becomes more evident as the number of users N increased. Note however that the population is assumed to be infinite in the S-ALOHA cases and is only stable when utilization is $\leq \frac{1}{e}$. In the TDMA cases, as N increases, the total frame length on the reservation channel also increases. A user's expected wait for his time slot becomes greater which can significantly add to his overall expected delay.

2.6.3.2.1 S-ALOHA Reservation + 9 Channels

$$T_{S-ALOHA,FCFS_{Case1}} \approx \frac{3Kl_r}{2C} + [e^G - 1][\frac{(H+2)\frac{Kl_r}{C}}{2} + 2T_{PD}] + 2T_{PD}$$
$$+\frac{\lambda^K \overline{X^2}(\overline{X})^{K-1}}{2(K-1)!(K-\lambda\overline{X})^2(\sum_{n=0}^{K-1}\frac{(\lambda\overline{X})^n}{n!} + \frac{(\lambda\overline{X})^K}{(K-1)!(K-\lambda\overline{X})})}$$
$$+\frac{LK}{C} + 2T_{PD} \text{ for } H \gg 1.$$

In this case, we allocate 1 subchannel for reservations using S-ALOHA and 9 subchannels, each of capacity $\frac{1}{10}C$ [bits/sec] for data transmissions.

2.6.3.2.2 S-ALOHA Reservation + 1 Channel

$$T_{S-ALOHA,FCFS_{Case2}} \approx \frac{3Kl_r}{2C} + [e^G - 1][\frac{(H+2)\frac{Kl_r}{C}}{2} + 2T_{PD}] + 2T_{PD}$$
$$+\frac{\lambda L^2K^2}{(K-1)^2C^2(2-2\frac{\lambda LK}{(K-1)C})} + \frac{LK}{(K-1)C}$$
$$+2T_{PD} \text{ for } H \gg 1.$$

In this case, we allocate 1 subchannel for reservations using S-ALOHA and 1 subchannel, with capacity $\frac{9}{10}C$ [bits/sec] for data transmission. This case has the same characteristics as seen with S-ALOHA Reservation + 9 Channels. However, the average delay of a message transfer has decreased since the channel capacity for data transmission has been increased while the average time to gain access to a channel remains the same. Notice in Figures 2-1, 2-4, and 2-7, that (h), S-ALOHA Reservation + 1 Channel, starts out with a lower delay than (f), TDMA Reservation + 1 Channel. The delay for accessing the reservation channel significantly increases until it reaches a maximum utilization of $\frac{1}{e}$, as seen by the S-ALOHA curves (c) and (d). The TDMA cases are seen as better in these figures because we have a system with $N = 10$ users whereas the S-ALOHA scenarios assume an infinite user population. The TDMA cases are corrected in the later figures as we increase the user population to mimic an infinite population.

2.7 Multichannel Architecture

In our analysis, we have divided the uplink channel into $K = 10$ equal capacity subchannels. One channel is dedicated for reservations while the remaining channel

resources are used for data transmissions. If we increase the number of subchannels, we can obtain higher utilization rates but at the cost of higher delay times. This phenomenon can be observed in Figure 2-10, where we investigated the channelized architecture with fixed access reservation with 1, 5, 10, 20, and 50 channels. In each multichannel scenario, the total capacity is divided evenly. As K increases, the overall delay significantly increases due to the increased delay in transmitting over a channel of smaller capacity. Note that it is possible to divide the total channel capacity unevenly, i.e., proportional to the demands of the system users, but that case is not considered here.

In Multichannel Case 1, we have K channels, each of capacity $\frac{C}{K}$ [bits/sec] with $K-1$ channels available for data transmission. In Multichannel Case 2, we have one channel with capacity $\frac{C}{K}$ [bits/sec] for reservations and the remaining capacity of $\frac{K-1}{K}C$ [bits/sec] available for data transmission. Comparing Figure 2-11 to Figure 2-10, we see that the Case 2 scenarios have a better delay performance. The amount of time to gain access to a data channel remains the same but the performance of the message transmission is better due to the increased capacity of the data channel.

With the ability of creating multichannel systems, we turn to the problem of optimizing the number of channels for data transmission. Appendix B.2 provides a simple way of looking at the problem of optimally dividing channel capacity. The result indicates that in order to minimize the overall delay, one should design a system where the optimal number of channels K is one.

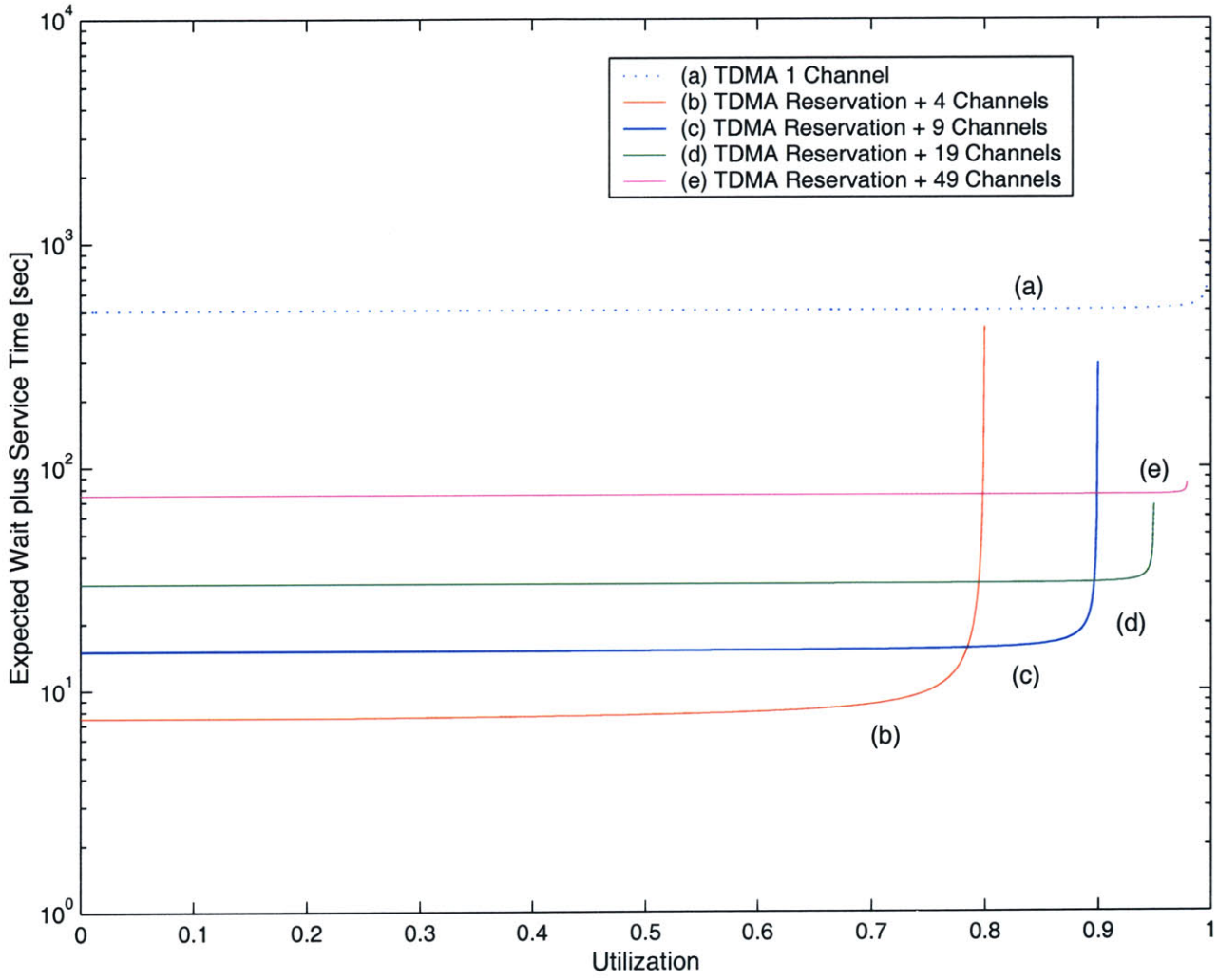


Figure 2-10: Multichannel Performance Analysis Case 1.
L = 1 [Mb], N = 1000, C = 1 [Mbits/sec] with $0 \leq \lambda \leq 1$ [msgs/sec]

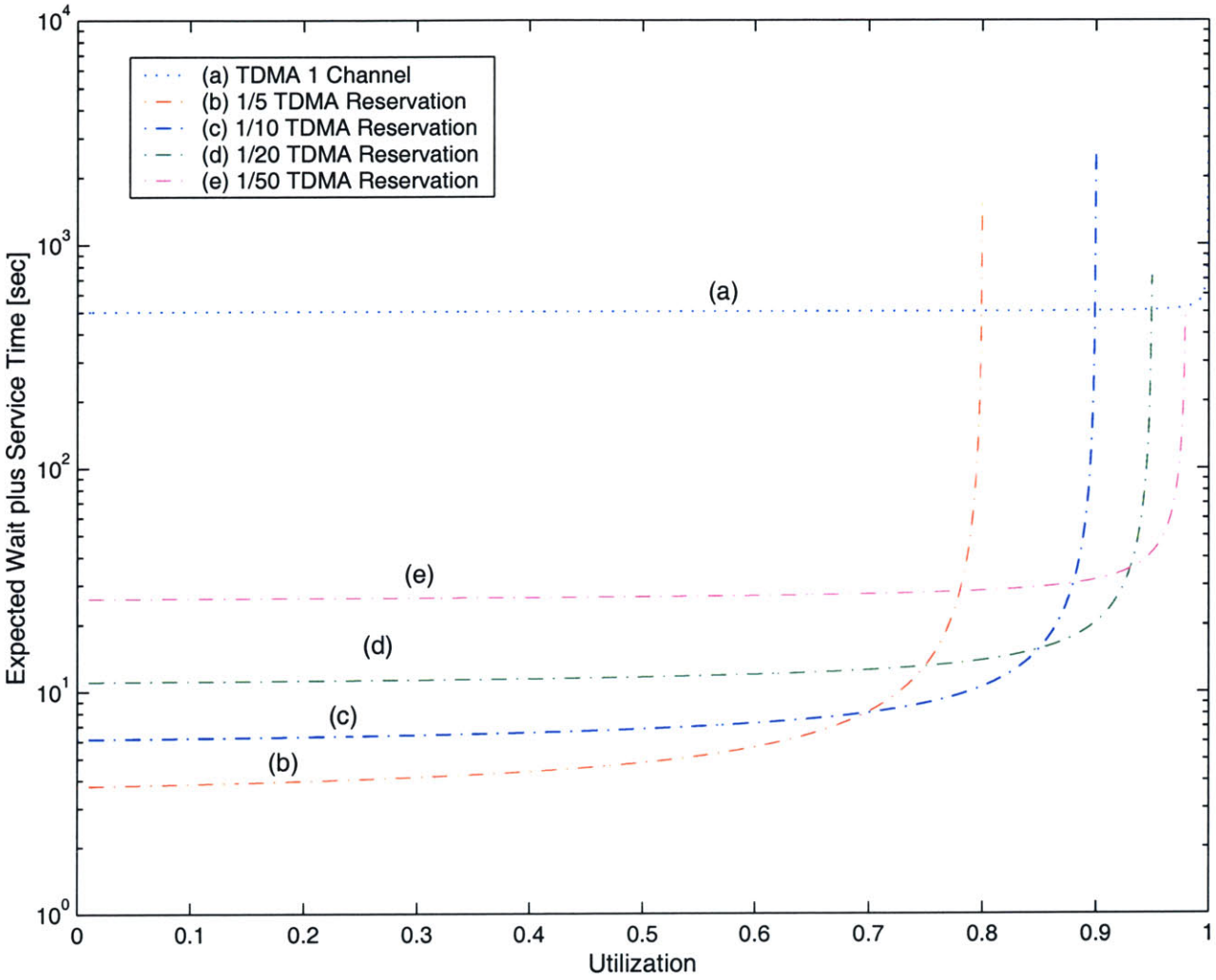


Figure 2-11: Multichannel Performance Analysis Case 2.
L = 1 [Mb], N = 1000, C = 1 [Mbits/sec] with $0 \leq \lambda \leq 1$ [msgs/sec]

2.8 Summary

We have introduced the topic of scheduling algorithms and channel architectures in this chapter. We then narrowed the list of scheduling algorithms and channel architectures to model, compare, and analyze. Modeling allows us to make assumptions to create a model that is simple yet efficiently true to the real system so that the answers provided by the model have some credibility. The delay performance of each the uplink architecture for different scenarios were plotted. The figures showed that random access algorithms are more suited for handling traffic from bursty sources, but they have low utilization. We see that higher utilization is achievable with the combination of a fixed access multiple access scheme. Multiple access schemes attempting to improve both utilization and delay performance must somehow incorporate the appropriate characteristics of fixed access and random access techniques. And while Kleinrock's proof in Appendix B.2 shows that the optimal number of channels is one, there may be situations were multichannel systems are appropriate for reasons emphasized in the previous chapter, Section 1.6.

Chapter 3

Channel Capacity Allocation for Mixed Traffic

3.1 Introduction

The next generation communication systems promise to provide a wide range of services to users, including high quality voice, variable rate data, full motion video, high resolution image, etc. In order to guarantee each service its Quality of Service (QoS), we must identify resource allocation and sharing schemes capable of statistically multiplexing services with considerably different characteristics. High- and low-rate users with different QoS requirements will coexist in the network, and thus, effective resource management has to guarantee required quality for all users.

In priority queueing systems [4], [12], and [20], users are distinguished into types and are served according to the priority of their type. It is well recognized that to support various traffics efficiently on packet networks a system has to be developed to satisfy demands of all classes of traffic according to the QoS. The scarcity of channel capacity to support emerging multimedia wireless applications requires efficient strategies for using the available bandwidth.

In this chapter, we will investigate the performance of different access schemes for multiple classes of users. We will consider a first-come first-serve (FCFS) access strategy, a nonpreemptive priority scheme, a preemptive priority scheme, and a channel capacity allocation scheme. We develop models that describe the overall delay for sending messages and analyze the performance. Our focus will concentrate on two classes of users. This scenario is typical of classes of users with small and large messages to transmit. We present quantitative results by making real-world parameter assumptions in the context of wireless communications, allowing the development of

intuition about the performance of the different architectures.

3.2 Priority Queueing

In priority queueing systems [4], [12], and [20], users are distinguished into types and are served according to the priority of their type. Consider the M/G/1 queueing system with the difference that arriving users are divided into n different priority classes with class 1 having the highest priority, class 2 having the second highest, and so on. The arrival rate and the first two moments of service time of each user class k are denoted $\lambda_k, \overline{X_k} = \frac{1}{\mu_k}$, and $\overline{X_k^2}$, respectively. The arrival processes of all users are assumed independent, Poisson, and independent of the service times.

3.2.1 No Priority (FCFS)

The simplest scheme for access to channel resources is a first-come first-serve (FCFS) ordering [4], [12], and [20], thus there is no priority rule in place. The model is an M/G/1 system with a total arrival rate of

$$\lambda = \lambda_1 + \lambda_2 + ... + \lambda_n. \tag{3.1}$$

The expected service time is then

$$\overline{X} = \frac{\lambda_1}{\lambda}\overline{X_1} + \frac{\lambda_2}{\lambda}\overline{X_2} + ... + \frac{\lambda_k}{\lambda}\overline{X_n}. \tag{3.2}$$

which follows since the combination of independent Poisson processes is itself a Poisson process whose rate is the sum of the rates of the component processes.

Consider the situation where there are two types of users, who arrive according to independent Poisson processes with respective rates λ_A and λ_B and have service distributions $\overline{X}_A$ and $\overline{X}_B$. For simplicity, we assume that the service times are identical and constant for all customers in a class, so that we have $\overline{X^2} = \overline{X}^2$. The M/D/1 queueing delay that a message undergoes is given by the following expression

$$\begin{aligned} Q &= \frac{\sum_{i=1}^{k} \lambda_i \overline{X^2}_i}{2(1 - \rho_1 - ... - \rho_k)} \\ &= \frac{\lambda_A \overline{X_A^2} + \lambda_B \overline{X_B^2}}{2(1 - \rho_A - \rho_B)}, \end{aligned} \tag{3.3}$$

where $\rho_k = \frac{\lambda_k}{\mu_k}$. Thus under a simple FCFS scheme, the total expected delay for a

user is

$$
\begin{aligned}
T &= \frac{\lambda_A}{\lambda_A+\lambda_B}T_A + \frac{\lambda_B}{\lambda_A+\lambda_B}T_B \\
&= \frac{\lambda_A}{\lambda_A+\lambda_B}(\overline{X}_A + Q) + \frac{\lambda_B}{\lambda_A+\lambda_B}(\overline{X}_B + Q) \\
&= \frac{1}{2}\frac{2\lambda_A L_A C - {\lambda_A}^2 {L_A}^2 - 4\lambda_A L_A \lambda_B L_B + \lambda_A \lambda_B {L_B}^2 + 2\lambda_B L_B C - {\lambda_B}^2 {L_B}^2 + \lambda_B \lambda_A {L_A}^2}{C(\lambda_A+\lambda_B)(C-\lambda_A L_A - \lambda_B L_B)} \\
&= \frac{1}{2}\frac{2xC - x^2 - 4xy + y\lambda_A L_B + 2yC - y^2 + x\lambda_B L_A}{\lambda C(C-x-y)}, \qquad (3.4)
\end{aligned}
$$

where $x = \lambda_A\overline{X_A}$, $y = \lambda_B\overline{X_B}$, and $\lambda = \lambda_A+\lambda_B$. T_A and T_B denote the time that Class A and Class B wait to gain access to channel resources and transmit their messages, respectively.

3.2.2 Nonpreemptive Priority

The nonpreemptive priority technique [4], [12], and [20] allows a user to be served without being interrupted even if a user of higher priority arrives in the meantime. Each priority class is separated into different queues. When the channel become available, the first user of the highest nonempty priority queue enters service.

We will begin by denoting the following terms:

$$
\begin{aligned}
Q_k &= \text{Average queueing time for priority } k, \\
\rho_k &= \frac{\lambda_k}{\mu_k} = \text{System utilization for priority } k, \\
T_k &= \text{Average delay for priority } k.
\end{aligned}
$$

The average queueing delay and total expected delay per user in each class is defined as

$$
\begin{aligned}
Q_k &= \frac{\sum_{i=1}^{n}\lambda_i\overline{X_i^2}}{2(1-\rho_1-\ldots-\rho_{k-1})(1-\rho_1-\ldots-\rho_k)}, \qquad (3.5) \\
T_k &= \frac{1}{\mu_k} + Q_k. \qquad (3.6)
\end{aligned}
$$

We assume that the overall system utilization is less than 1, that is,

$$\rho_1 + \rho_2 + ... + \rho_n < 1. \tag{3.7}$$

If this assumption is not satisfied, there may be a priority class whose average delay is infinite.

For the complete derivations of these equations, please refer to [4]. Notice that we can change the average delay a user experiences by classifying it to the appropriate priority class. The average delay generally decreases when users with short service times are given higher priority. For example, consider the copy machine waiting lines, where priority is often given to people who have a few copies to make.

Consider the situation where there are two types of users, who arrive according to independent Poisson processes with respective rates λ_A and λ_B and have service distributions $\overline{X}_A$ and $\overline{X}_B$. For simplicity, we assume that the service times are identical and constant for all customers in a class, so that we have $\overline{X^2} = \overline{X}^2$. Under a nonpreemptive priority scheme, we can determine the average time that it takes for a user of each class to gain access to the channel and transmit its message by

$$\overline{X}_B = \frac{L_B}{C}, \tag{3.8}$$

$$Q_B = \frac{1}{2}\frac{\lambda_A {L_A}^2 + \lambda_B {L_B}^2}{C(C - \lambda_B L_B)}, \tag{3.9}$$

$$\begin{aligned} T_B &= \overline{X}_B + Q_B \\ &= \frac{1}{2}\frac{2L_BC - \lambda_B {L_B}^2 + \lambda_A {L_A}^2}{C(C - \lambda_B L_B)} \\ &= \frac{1}{2}\frac{2L_BC - yL_B + xL_A}{C(C - y)}, \end{aligned} \tag{3.10}$$

$$\overline{X}_A = \frac{L_A}{C}, \tag{3.11}$$

$$Q_A = \frac{1}{2}\frac{\lambda_A {L_A}^2 + \lambda_B {L_B}^2}{(C - \lambda_B L_B)(C - \lambda_A L_A - \lambda_B L_B)}, \tag{3.12}$$

$$
\begin{aligned}
T_A &= \overline{X}_A + Q_A \\
&= \frac{1}{2}\frac{2L_AC^2 - C\lambda_A{L_A}^2 - 4L_AC\lambda_BL_B + 2\lambda_BL_B\lambda_A{L_A}^2 + 2L_A{\lambda_B}^2{L_B}^2 + C\lambda_B{L_B}^2}{C(C-\lambda_BL_B)(C-\lambda_AL_A\lambda_BL_B)} \\
&= \frac{1}{2}\frac{2L_AC^2 - xCL_A - 4yL_AC + 2xyL_A + 2y^2L_A + yCL_B}{C(C-y)(C-xy)}, \qquad (3.13)
\end{aligned}
$$

where $x = \lambda_A L_A$ and $y = \lambda_B L_B$.

3.2.3 Preemptive Resume Priority

The preemptive resume priority scheme [4], [12], and [20] allows an arriving higher-priority user to interrupt a current user's service. Service is resumed from the point of interruption once all users of higher priority have been served. Thus higher-priority users do not have to wait for lower-priority classes, a feature of the nonpreemptive priority scheme, evident in Q_k in Equation 3.5. The total expected delay under preemptive resume priority is

$$T_k = \overline{X}_k + Q_u + Q_w, \qquad (3.14)$$

where the average waiting time corresponding to service of users of priority 1 to k who are present in the queue when a user arrives is

$$Q_u = \frac{\sum_{i=1}^{k} \lambda_i \overline{X_i^2}}{2(1-\rho_1-\ldots-\rho_k)}, \qquad (3.15)$$

and the average waiting time corresponding to service times of users of priority 1 to $k-1$ who arrive while the user is waiting for service is

$$Q_w = \sum_{i=1}^{k-1} \frac{1}{\mu_i}\lambda_i T_k = \sum_{i=1}^{k-1} \rho_i T_k, \qquad (3.16)$$

for $k > 1$, and is zero for $k = 1$. Subsequently the final result is, for $k = 1$,

$$T_1 = \frac{\frac{1}{\mu_1}(1-\rho_1) + \frac{1}{2}\lambda_1\overline{X_1^2}}{1-\rho_1}, \qquad (3.17)$$

and for $k > 1$,

$$T_k = \frac{\frac{1}{\mu_k}(1-\rho_1-\ldots-\rho_k) + \frac{1}{2}\sum_{i=1}^{k}\lambda_i\overline{X_i^2}}{(1-\rho_1-\ldots-\rho_{k-1})(1-\rho_1-\ldots-\rho_k)}. \qquad (3.18)$$

For the derivations of these equations, please refer to [4]. For an example that approximates a preemptive resume priority, consider a transmission link serving several Poisson packet streams of different priorities. The packets of each stream are subdivided into many small subpackets. In the absence of packets of higher priority, they are continuously transmitted on the line. Otherwise, the transmission of the subpackets of a given packet is interrupted when a packet of higher priority arrives and is resumed when no subpackets of higher priority packets are left in the system. Preemptive resume priority is not easy to implement in a broadcast system, e.g., satellite or wireless communication system, because we cannot preempt a message that is already in flight or transmitting and coordination is often difficult.

Consider the situation where there are two types of users, who arrive according to independent Poisson processes with respective rates λ_A and λ_B and have service distributions $\overline{X}_A$ and $\overline{X}_B$. For simplicity, we assume that the service times are identical and constant for all customers in a class, so that we have $\overline{X^2} = \overline{X}^2$. Under a preemptive resume priority scheme, we can determine the average time that it takes for a user of each class to gain access to the channel and transmit its message by

$$\begin{aligned} T_B &= \frac{\overline{X}_B(1-\lambda_B\overline{X}_B)+\frac{1}{2}\lambda_B\overline{X_B^2}}{1-\lambda_B\overline{X}_B} \\ &= \frac{1}{2}\frac{L_B(2C-\lambda_B L_B)}{C(C-\lambda_B L_B)} \\ &= \frac{1}{2}\frac{L_B(2C-y)}{C(C-y)}, \end{aligned} \tag{3.19}$$

$$\begin{aligned} T_A &= \frac{\overline{X}_A(1-\lambda_B\overline{X}_B-\lambda_A\overline{X}_A)+\frac{1}{2}(\lambda_B\overline{X_B^2}+\lambda_A\overline{X_A^2})}{(1-\lambda_B\overline{X}_B)(1-\lambda_B\overline{X}_B-\lambda_A\overline{X}_A)} \\ &= \frac{1}{2}\frac{2L_AC-\lambda_A{L_A}^2-2L_A\lambda_BL_B+\lambda_B{L_B}^2}{(C-\lambda_BL_B)(C-\lambda_AL_A-\lambda_BL_B)} \\ &= \frac{1}{2}\frac{2L_AC-xL_A-2yL_A+yL_B}{(C-y)(C-x-y)}, \end{aligned} \tag{3.20}$$

where $x = \lambda_A L_A$ and $y = \lambda_B L_B$.

3.3 Channel Capacity Allocation

3.3.1 System Model

Evaluating the effectiveness of channel capacity allocation is an important performance evaluation. In this section we determine the division of total channel capacity for large message transfers with different classes of users. We begin by stating our model and necessary assumptions to simplify the analysis. Next we develop the analytical expressions to evaluate the system with typical real-world values. Following the presentation of results we summarize the important characteristics of the various channel capacity allocation schemes.

3.3.2 Model Parameters

Consider the situation where there are two types of user, who arrive according to independent Poisson processes with respective rates λ_A and λ_B and have service distributions $\overline{X}_A$ and $\overline{X}_B$. For simplicity, we assume that the service times are identical and constant for all customers in a class, so that we have $\overline{X^2} = \overline{X}^2$.

We will make real-world parameter assumptions in the context of wireless networks and analyze the performance to develop intuition about the effectiveness of each channel capacity allocation scheme. Table 3.1 lists the parameters used in our model.

Parameter	Description	Units
λ_a	Message arrival time for Class A	[msgs/sec]
λ_b	Message arrival time for Class B	[msgs/sec]
C	Total uplink channel capacity	[bits/sec]
C_a	Capacity allocated to Class A	[bits/sec]
C_b	Capacity allocated to Class B	[bits/sec]
L_a	Length of packet for Class A	[bits]
L_b	Length of packet for Class B	[bits]
θ	Division of channel capacity for Class A	$[0 \leq \theta \leq 1]$
$1-\theta$	Division of channel capacity for Class B	$[0 \leq 1-\theta \leq 1]$

Table 3.1: Model Parameters for Channel Capacity Allocation for Mixed Traffic.

3.3.2.1 Channel Capacity Allocation (θ)

A portion of the total channel capacity, equivalent to θC, is allocated to Class A users. The value of θ thus ranges from 0 to 1, $(0 \leq \theta \leq 1)$. Figure 3-1 is an illustration of the system model that is described. The remaining channel capacity of $(1 - \theta)C$ is allocated to Class B users.

While Figure 3-1 illustrates one channel of total capacity C, the following analysis of determining θ can also be applied to discrete multiple channels. The value of θ will then provide an approximate measure of dividing the multiple channels among the different classes of users. This practice can be seen at the supermarket, where there are special checkout counters for customers with few items. The question is then to determine how many of these checkout counters should be implemented.

3.4 Analysis

We will consider a communication system with static channel capacity allocation where the value of θ remains fixed in a system. With two classes of users, we will separate the total channel capacity into θC and $(1 - \theta)C$ as shown in Figure 3-1. To understand the relationship between Class A and Class B users under varying values of θ, we provide the following graph in Figure 3-2. The system has a total channel capacity of 1 [Kb/sec] and the average message length for Class A is 1 [Kb] and for Class B is 1 [Mb]. We plot the delays of Class A versus the delays of Class B as θ varies from 0 to 1. As expected, when $\theta = 0$, full channel capacity is given to Class B users and the average delay for their message transmission is at its minimum. When $\theta = 1$, full channel capacity is given to Class A users and the average delay for their message transmission is at its minimum.

We now develop a general function involving the normalized total delay times for Class A and Class B users. The total channel capacity is divided among Class A and

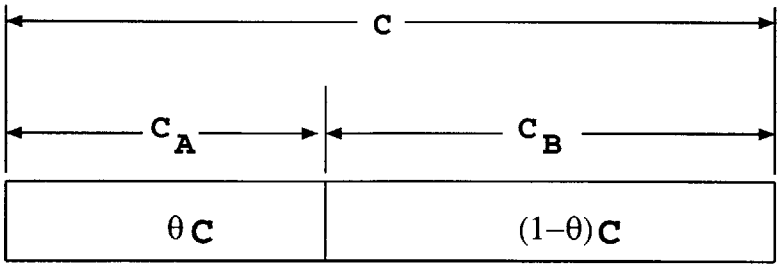


Figure 3-1: Channel Capacity Allocation.

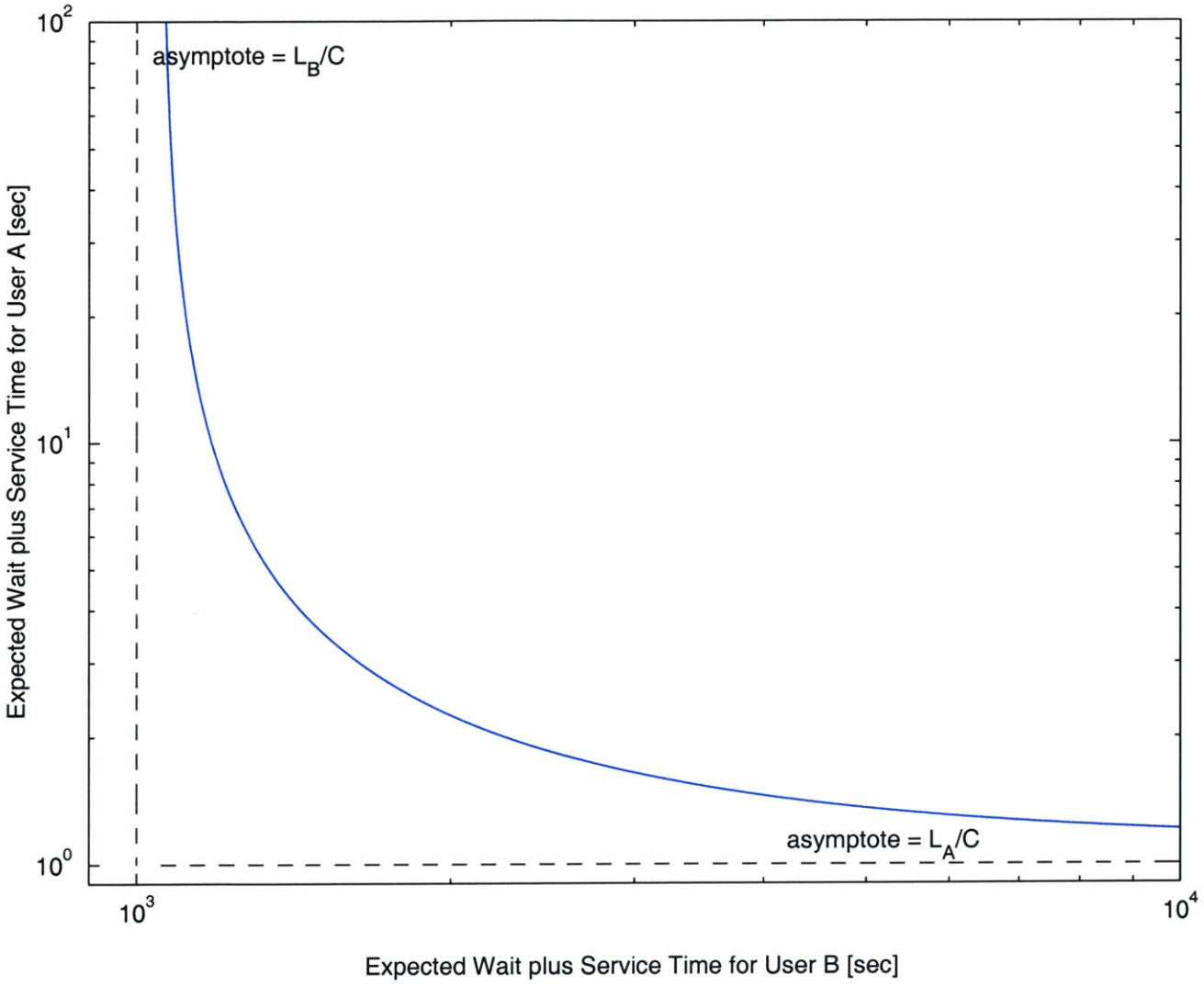


Figure 3-2: Delay Relationship between Class A and Class B Users.
$L_a = 1$ [Kb], $L_b = 1$ [Mb], C = 1 [Kb/sec]

Class B users according to the following ratios

$$C_A = \theta C, \tag{3.21}$$

$$C_B = (1-\theta)C. \tag{3.22}$$

The transmission time for packets of length L_A and L_B and the second moments are calculated as follows

$$\overline{X_A} = \frac{L_A}{C_A} = \frac{L_A}{\theta C}, \tag{3.23}$$

$$\overline{X_B} = \frac{L_B}{C_B} = \frac{L_B}{(1-\theta)C}, \tag{3.24}$$

$$\overline{X_A^2} = \overline{X_A}^2 = \frac{{L_A}^2}{\theta^2 C^2}, \tag{3.25}$$

$$\overline{X_B^2} = \overline{X_B}^2 = \frac{{L_B}^2}{(1-\theta)^2 C^2}. \tag{3.26}$$

The queueing delay for each class according to the M/D/1 Poisson process is

$$\begin{aligned} Q_A &= \frac{\lambda_A {L_A}^2}{\theta^2 C^2 (2 - 2\frac{\lambda_A L_A}{\theta C})} \\ &= \frac{1}{2}\frac{\lambda_A {L_A}^2}{\theta C(\theta C - \lambda_A L_A)}, \end{aligned} \tag{3.27}$$

$$\begin{aligned} Q_B &= \frac{\lambda_B {L_B}^2}{(1-\theta)^2 C^2 (2 - 2\frac{\lambda_B L_B}{(1-\theta)C})} \\ &= \frac{1}{2}\frac{\lambda_B {L_B}^2}{C(\lambda_B L_B + \theta C - C)(\theta - 1)}. \end{aligned} \tag{3.28}$$

Thus the total service time for each class is its transmission time plus its queueing delay is

$$T_A = \overline{X}_A + Q_A, \tag{3.29}$$

$$T_B = \overline{X}_B + Q_B. \tag{3.30}$$

Each sum is then normalized by dividing the total service time by its transmission

time,

$$
\begin{aligned}
T_{A_N} &= \frac{\overline{X_A} + Q_A}{\frac{L_A}{C_A}} \\
&= \frac{\frac{L_A}{\theta C} + \frac{\lambda_A {L_A}^2}{\theta^2 C^2 (2 - 2\frac{\lambda_A L_A}{\theta C})}\theta C}{L_A} \\
&= \frac{1}{2}\frac{2\theta C - \lambda_A L_A}{\theta C - \lambda_A L_A}, \qquad (3.31)
\end{aligned}
$$

$$
\begin{aligned}
T_{B_N} &= \frac{\overline{X_B} + Q_B}{\frac{L_B}{C_B}} \\
&= \frac{(\frac{L_B}{(1-\theta)C} + \frac{\lambda_B {L_B}^2}{(1-\theta)^2 C^2 (2 - 2\frac{\lambda_B L_B}{(1-\theta)C})})(1-\theta)C}{L_B} \\
&= \frac{1}{2}\frac{\lambda_B L_B + 2\theta C - 2C}{\lambda_B L_B + \theta C - C}. \qquad (3.32)
\end{aligned}
$$

Our objective function can be written as $G(T_{A_N}, T_{B_N})$. Now we would like to determine the value θ^* that minimizes some objective function involving $G(T_{A_N}, T_{B_N})$. The value θ^* provides the best reasonable service for the two classes of users by allocating the appropriate amount of channel capacity.

3.4.1 Case 1

One simple objective function is equating the delays for both classes of users, $G(T_{A_N}, T_{B_N}) = T_{A_N} = T_{B_N}$. The solution can be determined analytically in the following manner

$$
\begin{aligned}
G(T_{A_N}, T_{B_N}) &= T_{A_N} = T_{B_N} \qquad (3.33) \\
\frac{1}{2}\frac{2\theta C - \lambda_A L_A}{\theta C - \lambda_A L_A} &= \frac{1}{2}\frac{\lambda_B L_B + 2\theta C - 2C}{\lambda_B L_B + \theta C - C} \\
\theta^* &= \frac{\lambda_A L_A}{\lambda_A L_A + \lambda_B L_B}. \qquad (3.34)
\end{aligned}
$$

where θ^* is the optimum value of θ. If we use the following substitutions

$$x = \lambda_A L_A, \tag{3.35}$$

$$y = \lambda_B L_B. \tag{3.36}$$

we can simplify θ^* to be

$$\theta^* = \frac{x}{x+y}. \tag{3.37}$$

Without loss of generality, we can assume that $x > y$ and define the following relationship between x and y

$$\gamma = \frac{y}{x}, \tag{3.38}$$

such that $\gamma \in [0, 1]$. This will all us to rewrite θ^* as

$$\theta^* = \frac{1}{1+\gamma}. \tag{3.39}$$

Notice that when $\gamma = 1$, i.e, when the message sizes and the arrival rates of both classes are equal, θ^* is $\frac{1}{2}$. Figure 3-3 illustrates this relationship.

Now if we plug θ^* back into the equations for the normalized delays for each class, we get the following

$$\begin{aligned}
T_{A_N}(\theta^*) &= \frac{\overline{X_A} + Q_A}{\frac{L_A}{C_A}} \\
&= \frac{1}{2}\frac{2C - \lambda_B L_B - \lambda_A L_A}{C - \lambda_B L_B - \lambda_A L_A} \\
&= \frac{1}{2}\frac{2C - y - x}{C - y - x} \\
&= \frac{2C - 2y - 2x + y + x}{2C - 2y - 2x} \\
&= 1 + \frac{1}{2}\frac{y + x}{C - y - x} \\
&= 1 + \frac{1}{2}\frac{1 + \gamma}{\frac{C}{x} - (1 + \gamma)}.
\end{aligned} \tag{3.40}$$

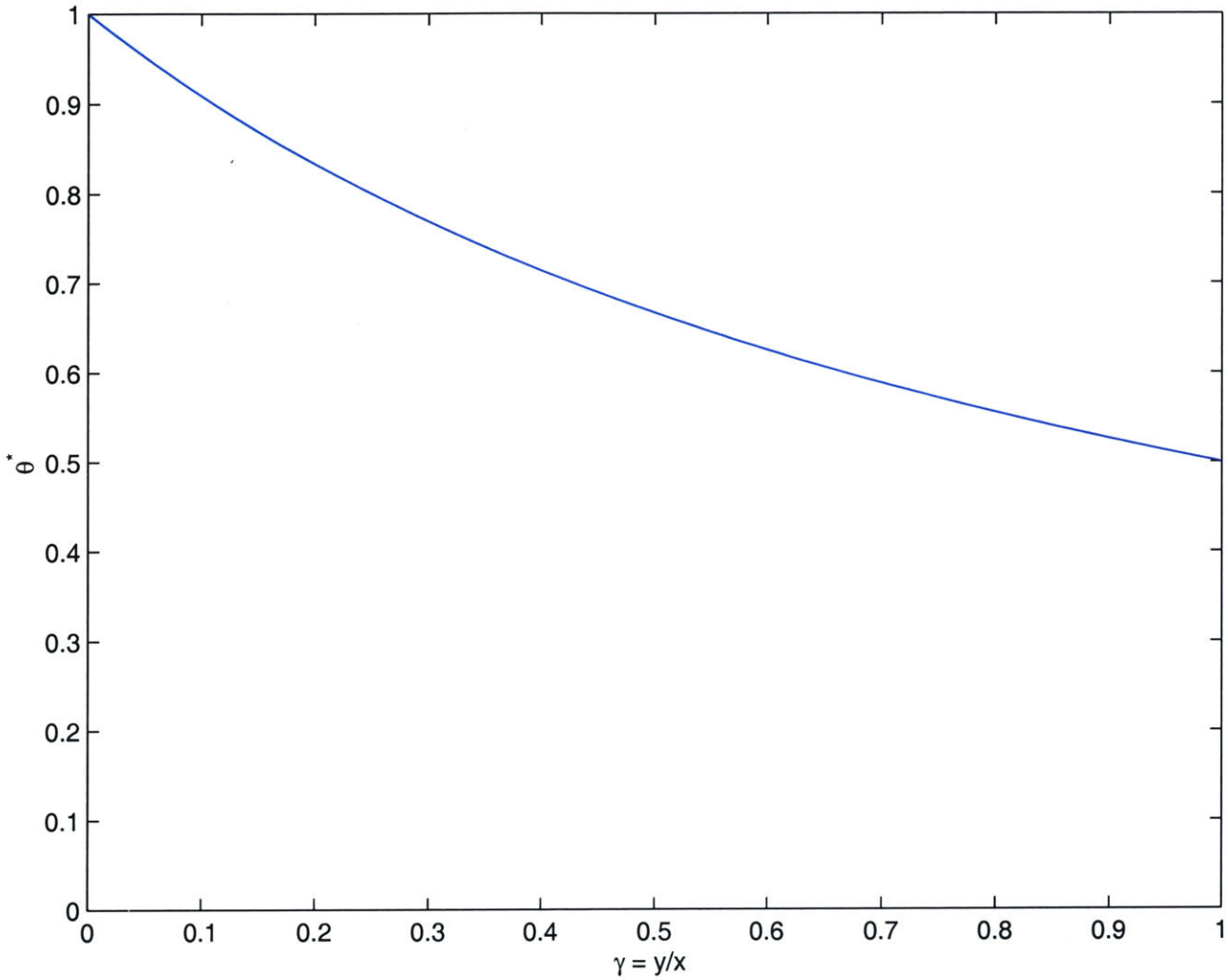


Figure 3-3: Relationship between γ and θ^* (I)
where $y < x$ and $\theta^* = \frac{x}{x+y}$.

$$
\begin{aligned}
T_{B_N}(\theta^*) &= \frac{\overline{X_B} + Q_B}{\frac{L_B}{C_B}} \\
&= \frac{1}{2}\frac{2C - \lambda_B L_B - \lambda_A L_A}{C - \lambda_B L_B - \lambda_A L_A} \\
&= \frac{1}{2}\frac{2C - y - x}{C - y - x} \\
&= 1 + \frac{1}{2}\frac{1+\gamma}{\frac{C}{x} - (1+\gamma)}. \qquad (3.41)
\end{aligned}
$$

If we plug θ^* back into the equations for the non-normalized delays for each class, we get the following

$$
\begin{aligned}
T_A(\theta^*) &= \overline{X}_A + Q_A \\
&= \frac{1}{2}\frac{(\lambda_B L_B + \lambda_A L_A)(2C - \lambda_B L_B - \lambda_A L_A)}{\lambda_A C(C - \lambda_B L_B - \lambda_A L_A)} \\
&= \frac{1}{2}\frac{(y+x)(2C - y - x)}{\lambda_A C(C - y - x)} \\
&= \frac{1}{2}\frac{1+\gamma}{\lambda_A \frac{C}{x}}\left[2 + \frac{1+\gamma}{\frac{C}{x} - (1+\gamma)}\right], \qquad (3.42)
\end{aligned}
$$

$$
\begin{aligned}
T_B(\theta^*) &= \overline{X}_B + Q_B \\
&= \frac{1}{2}\frac{(\lambda_B L_B + \lambda_A + L_A)(2C - \lambda_B L_B - \lambda_A L_A)}{\lambda_B C(C - \lambda_B L_B - \lambda_A L_A)} \\
&= \frac{1}{2}\frac{(y+x)(2C - y - x)}{\lambda_B C(C - y - x)} \\
&= \frac{1}{2}\frac{1+\gamma}{\lambda_B \frac{C}{x}}\left[2 + \frac{1+\gamma}{\frac{C}{x} - (1+\gamma)}\right]. \qquad (3.43)
\end{aligned}
$$

It is unreasonable to equate the normalized expected delays because class priorities may vary. Thus we look at optimizing θ for other relationships between Class A and Class B users. In each of the following cases, the beginning steps in the analysis remains the same, i.e., system setup parameters for C_A, C_B, $\overline{X_A}$, $\overline{X_B}$, $\overline{X_A^2}$, $\overline{X_B^2}$, Q_A, Q_B, T_A, and T_B. We define an objective function $G(T_{A_N}, T_{B_N})$ to describe the

relationship. We solve for the optimum value of θ in each of the following cases by first differentiating the relational equation with respect to θ, equating it to zero, and solving for θ.

3.4.2 Case 2

In this case, we look at optimizing θ when we sum the normalized expected delays for transmitting messages from both Class A and Class B users. With the substitution of Equations 3.36 and 3.36, we obtain the following relation

$$G(T_{A_N}, T_{B_N}) = T_{A_N} + T_{B_N} \tag{3.44}$$

$$\begin{aligned} f &= \frac{1}{2}\frac{2\theta C - \lambda_A L_A}{\theta C - \lambda_A L_A} + \frac{1}{2}\frac{\lambda_B L_B + 2\theta C - 2C}{\lambda_B L_B + \theta C - C} \\ &= \frac{4\theta C^2 - 4C^2\theta^2 - 3\theta Cy + (-3C + 3\theta C)x + 2yx}{2\theta C^2 - 2C^2\theta^2 - 2\theta Cy + (-2C + 2\theta C)x + 2yx} \\ &= \frac{1}{2}\frac{-4\theta C^2 + 4\theta^2 C^2 - 2yx - 3\theta Cx + 3Cx + 3\theta Cy}{-\theta C^2 + \theta^2 C^2 - yx - \theta Cx + Cx + \theta Cy}. \end{aligned} \tag{3.45}$$

The differentiation of f with respect to θ is

$$\begin{aligned} \frac{\partial f}{\partial \theta} &= \frac{4C^2 - 8\theta C^2 - 3Cy + 3Cx}{2\theta C^2 - 2C^2\theta^2 - 2\theta Cy + (-2C + 2\theta C)x + 2yx} \\ &\quad - \frac{(4\theta C^2 - 4C^2\theta^2 - 3\theta Cy + (-3C + 3\theta C)x + 2yx)(2C^2 - 4\theta C^2 - 2Cy + 2Cx)}{(2\theta C^2 - 2C^2\theta^2 - 2\theta Cy + (-2C + 2\theta C)x + 2yx)} \\ &= \frac{1}{2}\frac{C(2\theta C^2 x + yx^2 - y^2 x - C^2 x - \theta^2 C^2 x + \theta^2 C^2 y + 2Cyx - 4\theta Cyx)}{(-\theta C^2 + C^2\theta^2 - yx + Cx - \theta Cx + \theta Cy)^2}. \end{aligned} \tag{3.46}$$

We then set Equation 3.46 to be zero and solve for θ. The two solutions for θ are

$$\begin{aligned} \theta^* &= \frac{1}{2}\frac{-2Cx + 4yx \pm 2\sqrt{-2Cx^2y + 2y^2x^2 - 2Cy^2x + y^3x + yC^2x + yx^3}}{(y - x)C} \\ &= \frac{2yx - Cx \pm \sqrt{xy(C - y - x)^2}}{(y - x)C}, \end{aligned} \tag{3.47}$$

where

$$\theta_1^* = \frac{2yx - Cx + \sqrt{xy(C-y-x)^2}}{(y-x)C}, \tag{3.48}$$

$$\theta_2^* = \frac{2yx - Cx - \sqrt{xy(C-y-x)^2}}{(y-x)C}. \tag{3.49}$$

Now we are going to manipulate the solutions so that we may obtain a more intuitive equation about the relationship between x, y, and θ and to determine which root is an admissible solution. We will focus on θ_1^* first. The numerator can be reduced as follows

$$\begin{aligned}
numerator_1 &= 2yx - Cx + \sqrt{xy(C-y-x)^2} \\
&= x(2y-C) + [C-(x+y)]\sqrt{xy} \\
&= C(\sqrt{xy} - x) + 2xy - (x+y)\sqrt{xy} \\
&= C(\sqrt{x})(\sqrt{y} - \sqrt{x}) + [2\sqrt{xy} - (x+y)]\sqrt{xy} \\
&= C(\sqrt{x})(\sqrt{y} - \sqrt{x}) - [x - 2\sqrt{xy} + y]\sqrt{xy} \\
&= C(\sqrt{x})(\sqrt{y} - \sqrt{x}) - (\sqrt{y} - \sqrt{x})^2\sqrt{xy} \\
&= \sqrt{x}(\sqrt{y} - \sqrt{x})[C - \sqrt{y}(\sqrt{y} - \sqrt{x})].
\end{aligned} \tag{3.50}$$

The denominator can be expanded to the following

$$\begin{aligned}
denominator &= (y-x)C \\
&= C(\sqrt{y} + \sqrt{x})(\sqrt{y} - \sqrt{x}).
\end{aligned} \tag{3.51}$$

Thus θ_1^* becomes

$$\begin{aligned}
\theta_1^* &= \frac{\sqrt{x}(\sqrt{y} - \sqrt{x})[C - \sqrt{y}(\sqrt{y} - \sqrt{x})]}{C(\sqrt{y} + \sqrt{x})(\sqrt{y} - \sqrt{x})} \\
&= \frac{\sqrt{x}[C - \sqrt{y}(\sqrt{y} - \sqrt{x})]}{C(\sqrt{y} + \sqrt{x})} \\
&= \frac{\sqrt{x}}{\sqrt{y} + \sqrt{x}} - \frac{\sqrt{xy}}{C}\left[\frac{\sqrt{y} - \sqrt{x}}{\sqrt{y} + \sqrt{x}}\right].
\end{aligned} \tag{3.52}$$

Using the relationship defined in Equation 3.38, we can simplify θ^* as

$$\begin{aligned}
\theta_1^* &= \frac{\sqrt{x}}{\sqrt{y}+\sqrt{x}} - \frac{\sqrt{xy}}{C}\left[\frac{\sqrt{y}-\sqrt{x}}{\sqrt{y}+\sqrt{x}}\right] \\
&= \frac{1}{1+\sqrt{\gamma}} - \frac{\sqrt{xy}}{C}\left[\frac{\sqrt{\gamma}-1}{\sqrt{\gamma}+1}\right] \\
&= \frac{1}{1+\sqrt{\gamma}} + \frac{\sqrt{xy}}{C}\left[\frac{1-\sqrt{\gamma}}{1+\sqrt{\gamma}}\right] \\
&= \frac{1}{1+\sqrt{\gamma}} + \frac{x\sqrt{\gamma}}{C}\left[\frac{1-\sqrt{\gamma}}{1+\sqrt{\gamma}}\right] \\
&= \frac{1}{1+\sqrt{\gamma}}\left[1+\frac{x}{C}\sqrt{\gamma}(1-\sqrt{\gamma})\right]. \qquad (3.53)
\end{aligned}$$

A plot of the relationship between γ and θ_1^* is plotted in Figure 3-4. In this graph, $\frac{x}{C}$ is given as a constant and γ increases as $y = (\lambda_B L_B)$ increases. In the limiting case where the message size and arrivals are the same for both classes of users, i.e., $x = y$, the second term of Equation 3.53 drops out since it goes to zero because $\gamma = 1$ and the final result is

$$\begin{aligned}
\theta_1^*(\gamma = 1) &= \frac{1}{1+\sqrt{\gamma}}\left[1+\frac{x}{C}\sqrt{\gamma}(1-\sqrt{\gamma})\right] \\
&= \frac{1}{2}. \qquad (3.54)
\end{aligned}$$

Now if we plug θ^* back into the equations for the normalized delays for each class, we get the following

$$\begin{aligned}
T_{A_N}(\theta^*) &= \frac{\overline{X_A} + Q_A}{\frac{L_A}{C_A}} \\
&= \frac{1}{2}\frac{2C + x\sqrt{\gamma} - 2y - x}{C - y - x} \\
&= \frac{2C - 2y - 2x + x\sqrt{\gamma} + x}{2C - 2y - 2x} \\
&= 1 + \frac{1}{2}\frac{x\sqrt{\gamma} + x}{C - y - x} \\
&= 1 + \frac{1}{2}\frac{1+\sqrt{\gamma}}{\frac{C}{x} - (1+\gamma)}, \qquad (3.55)
\end{aligned}$$

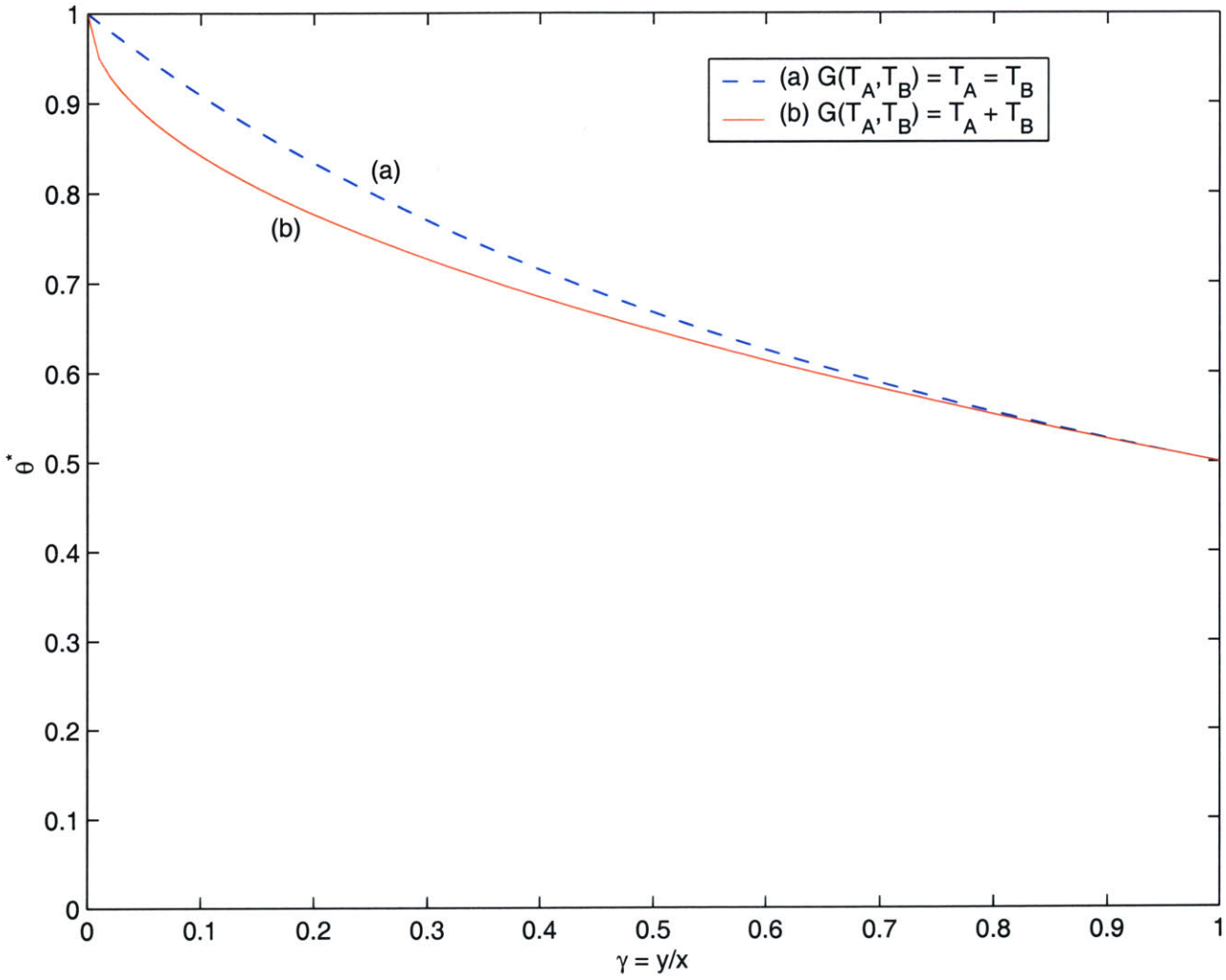


Figure 3-4: Relationship between γ and θ^* (II)
where $y < x$, (a)$\theta^* = \frac{x}{x+y}$ and (b) $\theta^* = \frac{1}{1+\sqrt{\gamma}} + \frac{x\sqrt{\gamma}}{C}(\frac{1-\sqrt{\gamma}}{1+\sqrt{\gamma}})$.

$$
\begin{aligned}
T_{B_N}(\theta^*) &= \frac{\overline{X_B} + Q_B}{\frac{L_B}{C_B}} \\
&= \frac{1}{2}\frac{2C\sqrt{\gamma} - 2x\sqrt{\gamma} + y - y\sqrt{\gamma}}{C\sqrt{\gamma} - x\sqrt{\gamma} - y\sqrt{\gamma}} \\
&= \frac{2C\sqrt{\gamma} - 2x\sqrt{\gamma} - 2y\sqrt{\gamma} + y\sqrt{\gamma} + y}{2C\sqrt{\gamma} - 2x\sqrt{\gamma} - 2y\sqrt{\gamma}} \\
&= 1 + \frac{y + y\sqrt{\gamma}}{2C\sqrt{\gamma} - x\sqrt{\gamma} - y\sqrt{\gamma}} \\
&= 1 + \frac{1}{2}\frac{1 + \sqrt{\gamma}}{\frac{C}{y}\sqrt{\gamma} - \frac{1}{\sqrt{\gamma}} - \sqrt{\gamma}} \qquad (3.56)
\end{aligned}
$$

.

If we plug θ^* back into the equations for the non-normalized delays for each class, we get the following

$$
\begin{aligned}
T_A(\theta^*) &= \overline{X}_A + Q_A \\
&= \frac{1}{2}\frac{L_A(1+\sqrt{\gamma})(2C + x\sqrt{\gamma} - 2y - x)}{(C + x\sqrt{\gamma} - y)(C - y - x)} \\
&= \frac{1}{2}\frac{L_A(1+\gamma)}{C + x\sqrt{\gamma} - y}\left[2 + \frac{1+\sqrt{\gamma}}{\frac{C}{x} - (1+\gamma)}\right], \qquad (3.57)
\end{aligned}
$$

$$
\begin{aligned}
T_B(\theta^*) &= \overline{X}_B + Q_B \\
&= \frac{1}{2}\frac{L_B(1+\sqrt{\gamma})(2C\sqrt{\gamma} - 2x\sqrt{\gamma} + y - y\sqrt{\gamma})}{(C - x - x\sqrt{\gamma})(C\sqrt{\gamma} - x\sqrt{\gamma} - y\sqrt{\gamma})} \\
&= \frac{1}{2}\frac{L_B(1+\sqrt{\gamma})}{C - x - x\sqrt{\gamma}}\left[2 + \frac{1+\sqrt{\gamma}}{\frac{C}{y}\sqrt{\gamma} - \frac{1}{\sqrt{\gamma}} - \sqrt{\gamma}}\right]. \qquad (3.58)
\end{aligned}
$$

Now we must check on the second root. The numerator can be reduced as follows:

$$\begin{aligned} numerator_2 &= 2yx - Cx - \sqrt{xy(C-y-x)^2} \\ &= x(2y-C) - [C-(x+y)]\sqrt{xy} \\ &= -C(x+\sqrt{xy}) + 2xy + \sqrt{xy}(x+y) \\ &= -C(\sqrt{x})(\sqrt{x}+\sqrt{y}) + [x+2\sqrt{xy}+y]\sqrt{xy} \\ &= -C(\sqrt{x})(\sqrt{x}+\sqrt{y}) + (\sqrt{x}+\sqrt{y})^2\sqrt{xy} \\ &= \sqrt{x}(\sqrt{x}+\sqrt{y})[\sqrt{y}(\sqrt{x}+\sqrt{y}) - C]. \end{aligned} \tag{3.59}$$

The denominator remains the same, as shown in Equation 3.51. Thus θ^* becomes:

$$\begin{aligned} \theta_2^* &= \frac{\sqrt{x}(\sqrt{x}+\sqrt{y})[\sqrt{y}(\sqrt{x}+\sqrt{y}) - C]}{C(\sqrt{y}+\sqrt{x})(\sqrt{y}-\sqrt{x})} \\ &= \frac{\sqrt{x}(\sqrt{x}+\sqrt{y}) - C)}{C(\sqrt{y}-\sqrt{x})}. \end{aligned} \tag{3.60}$$

As y approaches x, Equation 3.60 blows up and approaches infinity. This solution is thus inadmissible since by definition, $\theta \in [0,1]$.

Note that the value of θ^* is a minimum and not a maximum. It can be shown that θ^* is a minimum if the second derivative of the function is positive. Another way to verify this is the realization that when $\theta = 0$ or $\theta = 1$, one of the expected delays is infinite. The endpoints of our function are thus greater than any value of the function when $0 < \theta < 1$. Since there is only one inflection point in our curve, the value of θ^* is a minimum. This is still true for Case 3 but not for Case 4 since the function is no longer quadratic.

3.4.3 Case 3

In this case, we look at optimizing θ when we sum the normalized expected delays for transmitting messages from both Class A and Class B users with an adjustment factor on Class B users. The weight η indicates the importance of Class B users. Notice that when $\eta = 1$, we have Case 2: $G(T_{A_N}, T_{B_N}) = T_{A_N} + T_{B_N}$. Thus the

relationship is shown as

$$G(T_{A_N}, T_{B_N}) = T_{A_N} + \eta T_{B_N} \tag{3.61}$$

$$\begin{aligned} f &= \frac{1}{2}\frac{2\theta C - \lambda_A L_A}{\theta C - \lambda_A L_A} + \eta[\frac{1}{2}\frac{\lambda_B L_B + 2\theta C - 2C}{\lambda_B L_B + \theta C - C}] \\ &= \frac{(-\theta C - 2\eta\theta C + C + 2\eta C)x - 2\theta C^2 + 2C^2\theta^2}{(2C - 2\theta C)x + 2C^2\theta^2 - 2\theta C^2 - 2xy + 2\theta Cy} \\ &\quad \frac{+2\eta C^2\theta^2 - 2\eta\theta C^2 + (-1-\eta)xy + (2\theta C + \eta\theta C)y}{(2C - 2\theta C)x + 2C^2\theta^2 - 2\theta C^2 - 2xy + 2\theta Cy} \\ &= \frac{1}{2}\{\frac{2\theta C^2 - 2C^2\theta^2 + 2\eta\theta C^2 - 2\eta C^2\theta^2 + xy}{\theta C^2 - C^2\theta^2 + xy - \theta Cy - Cx + \theta Cx} \\ &\quad + \frac{\eta xy - 2\theta Cy - \eta\theta Cy + \theta Cy + \theta Cx + 2\eta\theta Cx - Cx - 2\eta Cx}{\theta C^2 - C^2\theta^2 + xy - \theta Cy - Cx + \theta Cx}\}. \end{aligned} \tag{3.62}$$

The differentiation of f with respect to θ is

$$\begin{aligned} \frac{\partial f}{\partial \theta} &= \frac{1}{2}\{\frac{C(\eta x^2 y - xC^2\theta^2 + 2Cxy - 2x\theta Cy - C^2 x)}{(\theta C^2 - C^2\theta^2 + xy - \theta Cy - Cy - Cx + \theta Cx)^2} \\ &\quad + \frac{C(-xy^2 - 2\eta x\theta Cy + 2x\theta C^2 + \eta\theta^2 C^2 y)}{(\theta C^2 - C^2\theta^2 + xy - \theta Cy - Cy - Cx + \theta Cx)^2}\}. \end{aligned} \tag{3.63}$$

We then set Equation 3.63 to be zero and solve for θ. The two solutions for θ are

$$\theta^* = \frac{-Cx + xy + \eta xy \pm \sqrt{\eta yx(C - x - y)^2}}{(\eta y - x)C}, \tag{3.64}$$

where

$$\theta_1^* = \frac{-Cx + xy + \eta xy + \sqrt{\eta yx(C - x - y)^2}}{(\eta y - x)C}, \tag{3.65}$$

$$\theta_2^* = \frac{-Cx + xy + \eta xy - \sqrt{\eta yx(C - x - y)^2}}{(\eta y - x)C}. \tag{3.66}$$

Now we are going to manipulate the solutions so that we may obtain a more intuitive equation about the relationship between x, y, and θ and to determine which root is an admissible solution. We will focus on θ_1^* first. The numerator can be

reduced as follows

$$\begin{aligned} numerator_1 &= -Cx + xy + \eta xy + \sqrt{\eta yx(C-x-y)^2} \\ &= x(y + \eta y - C) + [C - (x+y)]\sqrt{\eta xy} \\ &= C(\sqrt{\eta xy} - x) + xy + \eta xy - (x+y)\sqrt{\eta xy} \\ &= C\sqrt{x}(\sqrt{\eta y} - \sqrt{x}) + \sqrt{xy}(\sqrt{xy} + \eta\sqrt{xy} - x\sqrt{\eta} - y\sqrt{\eta}) \\ &= C\sqrt{x}(\sqrt{\eta y} - \sqrt{x}) + \sqrt{xy}(\sqrt{\eta y} - \sqrt{x})(\sqrt{\eta x} - \sqrt{y}). \end{aligned} \tag{3.67}$$

The denominator can be expanded to the following

$$\begin{aligned} denominator &= (\eta y - x)C \\ &= C(\sqrt{\eta y} + \sqrt{x}))(\sqrt{\eta y} - \sqrt{x}). \end{aligned} \tag{3.68}$$

Thus θ_1^* becomes

$$\begin{aligned} \theta_1^* &= \frac{C\sqrt{x}(\sqrt{\eta y} - \sqrt{x}) + \sqrt{xy}(\sqrt{\eta y} - \sqrt{x})(\sqrt{\eta x} - \sqrt{y})}{C(\sqrt{\eta y} + \sqrt{x})(\sqrt{\eta y} - \sqrt{x})} \\ &= \frac{C\sqrt{x} + \sqrt{xy}(\sqrt{\eta x} - \sqrt{y})}{(\sqrt{\eta y} + \sqrt{x})C}. \end{aligned} \tag{3.69}$$

Using the relationship defined in Equation 3.38, we can simplify θ^* as

$$\begin{aligned} \theta_1^* &= \frac{C\sqrt{x} + \sqrt{xy}(\sqrt{\eta x} - \sqrt{y})}{(\sqrt{\eta y} + \sqrt{x})C} \\ &= \frac{C\sqrt{x} + \sqrt{xy}(\sqrt{\eta x} - \sqrt{y})}{C(\sqrt{\eta y} + \sqrt{x})} \\ &= \frac{C + x\sqrt{\gamma}(\sqrt{\eta} - \sqrt{\gamma})}{C(\sqrt{\eta\gamma} + 1)} \\ &= \frac{1}{1 + \sqrt{\eta\gamma}} + \frac{x}{C}\frac{\sqrt{\gamma}}{1 + \sqrt{\eta\gamma}}(\sqrt{\eta} - \sqrt{\gamma}). \end{aligned} \tag{3.70}$$

A plot of the relationship between γ and θ_1^* is plotted in Figure 3-5. In this graph, $\frac{x}{C}$ is given as a constant and γ increases as $y = (\lambda_B L_B)$ increases.

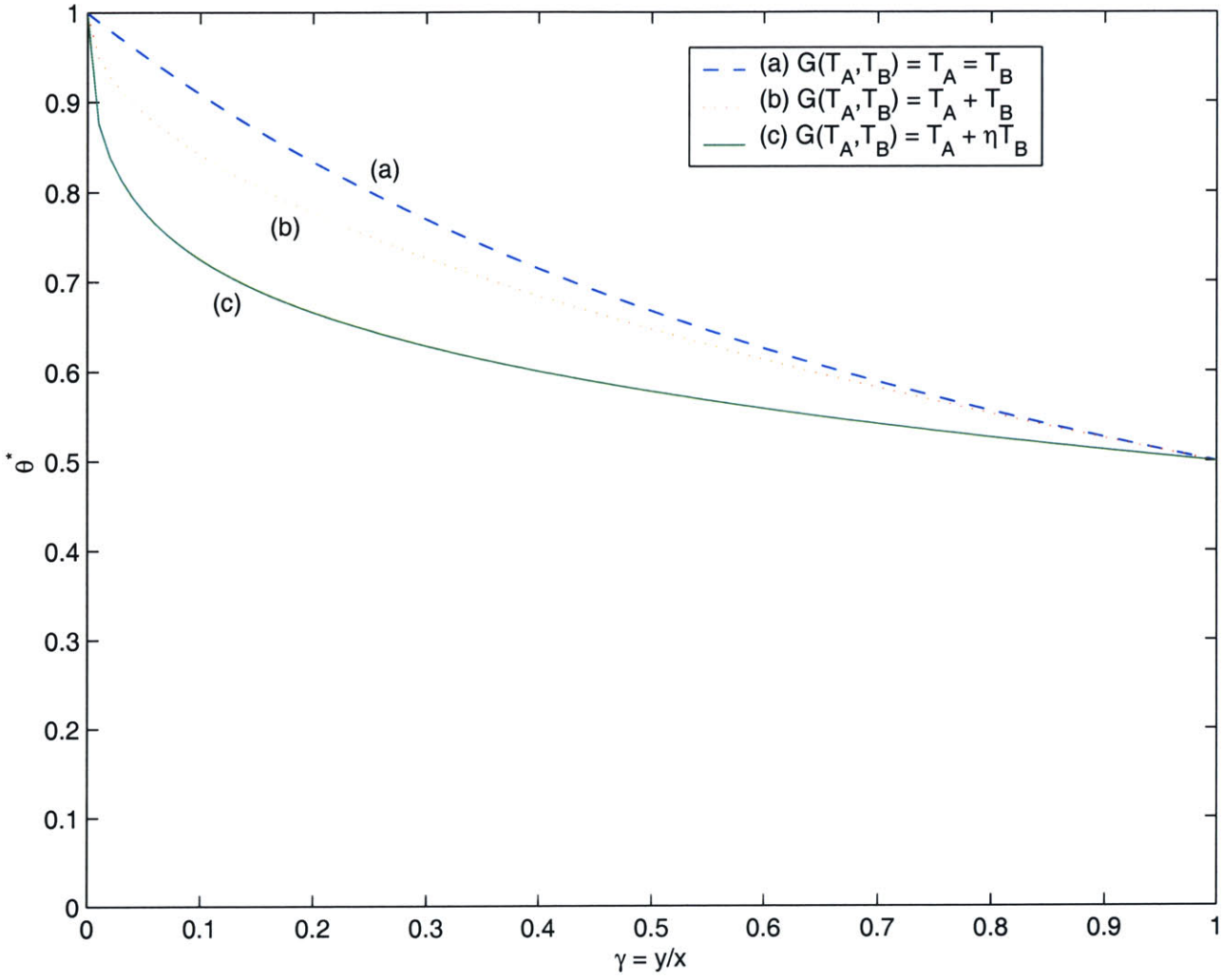


Figure 3-5: Relationship between γ and θ^* (III)

where $y < x$, (a)$\theta^* = \frac{x}{x+y}$, (b) $\theta^* = \frac{1}{1+\sqrt{\gamma}} + \frac{x\sqrt{\gamma}}{C}(\frac{1-\sqrt{\gamma}}{1+\sqrt{\gamma}})$
and (c) $\theta^* = \frac{1}{1+\sqrt{\eta\gamma}} + \frac{x}{C}\frac{\sqrt{\gamma}}{1+\sqrt{\eta\gamma}}(\sqrt{\eta} - \sqrt{\gamma})$.

Now if we plug θ^* back into the equations for the normalized delays for each class, we get the following

$$\begin{aligned}
T_{A_N}(\theta^*) &= \frac{\overline{X_A} + Q_A}{\frac{L_A}{C_A}} \\
&= \frac{1}{2}\frac{2C + 2x\sqrt{\eta\gamma} - 2y - x - x\sqrt{\eta\gamma}}{C + x\sqrt{\eta\gamma} - y - x - x\sqrt{\eta\gamma}} \\
&= \frac{2C + 2x\sqrt{\eta\gamma} - 2y - 2x - 2x\sqrt{\eta\gamma} + x + x\sqrt{\eta\gamma}}{2C + 2x\sqrt{\eta\gamma} - 2y - 2x - 2x\sqrt{\eta\gamma}} \\
&= 1 + \frac{1}{2}\frac{x + x\sqrt{\eta\gamma}}{C + x\sqrt{\eta\gamma} - y - x - x\sqrt{\eta\gamma}} \\
&= 1 + \frac{1}{2}\frac{1 + \sqrt{\eta\gamma}}{\frac{C}{x} + \sqrt{\eta\gamma} - (1 + \gamma + \sqrt{\eta\gamma})},
\end{aligned} \tag{3.71}$$

$$\begin{aligned}
T_{B_N}(\theta^*) &= \frac{\overline{X_B} + Q_B}{\frac{L_B}{C_B}} \\
&= \frac{1}{2}\frac{2C\sqrt{\eta\gamma} - 2x\sqrt{\eta\gamma} + y - y\sqrt{\eta\gamma}}{C\sqrt{\eta\gamma} - x\sqrt{\eta\gamma} - y\sqrt{\eta\gamma}} \\
&= \frac{2C\sqrt{\eta\gamma} - 2x\sqrt{\eta\gamma} - 2y\sqrt{\eta\gamma} + y + y\sqrt{\eta\gamma}}{2C\sqrt{\eta\gamma} - 2x\sqrt{\eta\gamma} - 2y\sqrt{\eta\gamma}} \\
&= 1 + \frac{y + y\sqrt{\eta\gamma}}{2C\sqrt{\eta\gamma} - 2x\sqrt{\eta\gamma} - 2y\sqrt{\eta\gamma}} \\
&= 1 + \frac{1}{2}\frac{1 + \sqrt{\eta\gamma}}{\sqrt{\eta\gamma}(\frac{C}{y} - \frac{1}{\gamma} - 1)}.
\end{aligned} \tag{3.72}$$

If we plug θ^* back into the equations for the non-normalized delays for each class, we get the following

$$\begin{aligned}
T_A(\theta^*) &= \overline{X}_A + Q_A \\
&= \frac{1}{2}\frac{L_A(1 + \sqrt{\eta\gamma})(2C + 2x\sqrt{\eta\gamma} - 2y - x - x\sqrt{\eta\gamma})}{(C + x\sqrt{\eta\gamma} - y)(C + x\sqrt{\eta\gamma} - y - x - x\sqrt{\eta\gamma})} \\
&= \frac{1}{2}\frac{L_A(1 + \sqrt{\eta\gamma})}{C + x\sqrt{\eta\gamma} - y}\left[2 + \frac{1 + \sqrt{\eta\gamma}}{\frac{C}{x} + \sqrt{\eta\gamma} - (1 + \gamma + 1))}\right],
\end{aligned} \tag{3.73}$$

$$
\begin{aligned}
T_B(\theta^*) &= \overline{X}_B + Q_B \\
&= \frac{1}{2}\frac{L_B(1+\sqrt{\eta\gamma})(2C\sqrt{\eta\gamma} - 2x\sqrt{\eta\gamma} + y - y\sqrt{\eta\gamma})}{(C\sqrt{\eta\gamma} - x\sqrt{\eta\gamma} + y)\sqrt{\eta\gamma}(C - x - y)} \\
&= \frac{1}{2}\frac{L_B(1+\sqrt{\eta\gamma})}{C\sqrt{\eta\gamma} - x\sqrt{\eta\gamma} + y}\left[2 + \frac{1+\sqrt{\eta\gamma}}{\sqrt{\eta\gamma}(\frac{C}{y} - \frac{1}{\gamma} - 1)}\right]. \qquad (3.74)
\end{aligned}
$$

Now when we continue to check on the second root in the same manner, we realize that the second root has the same property as the second root in Case 2. The second root is an inadmissible solution. Notice that Case 2 is a subset of Case 3, when $\eta = 1$.

3.4.4 Case 4

In this case, we look at optimizing θ when we sum the non-linear expected delays for transmitting messages from both Class A and Class B users with an adjustment factor on Class B users. The weight η indicates the importance of Class B users. Thus the relationship is shown as

$$
\begin{aligned}
G(T_{A_N}, T_{B_N}) &= (T_{A_N})^k + \eta(T_{B_N})^k \qquad (3.75) \\
f &= (\frac{1}{2}\frac{2\theta C - \lambda_A L_A}{\theta C - \lambda_A L_A})^k + \eta(\frac{1}{2}\frac{\lambda_B L_B + 2\theta C - 2C}{\lambda_B L_B + \theta C - C})^k \\
&= (\frac{1}{2})^k(\frac{-2\theta C + x}{-\theta C + x})^k + \eta(\frac{1}{2})^k(\frac{-2C + 2\theta C + y}{-C + \theta C + y})^k. \qquad (3.76)
\end{aligned}
$$

The differentiation of f with respect to θ is

$$
\begin{aligned}
\frac{\partial f}{\partial \theta} &= \frac{(\frac{1}{2})^k(\frac{-2\theta C+x}{-\theta C+x})^k k(-2\frac{C}{-\theta C+x} + \frac{(-2\theta C+x)C}{(-\theta C+x)^2})(-\theta C + x)}{-2\theta C + x} \\
&\quad + \frac{\eta(\frac{1}{2})^k(\frac{-2C+2\theta C+y}{-C+\theta C+y})^k k(2\frac{C}{-C+\theta C+y} - \frac{(-2C+2\theta C+y)C}{(-C+\theta C+y)^2})(-C + \theta C + y)}{-2C + 2\theta C + y}.
\end{aligned} \qquad (3.77)
$$

Setting Equation 3.77 equal to zero and solving for θ is a difficult task. It is left to the reader to use numerical analysis to solve the equation following the same technique used in the previous cases.

3.5 Delay Performance Figures

We are now prepared to analyze the performance of several situations. We will look at the case of 2 different user classes. Class A users will be sending very large messages while Class B users will be sending much smaller messages. We will use the equations that we have formulated in this chapter to plot the performance of Class B users. We assigned the auxiliary parameters with values typical of wireless communication systems. These values are summarized in Table 3.2. We also make the assumption that $y < x$ so that $\gamma = \frac{y}{x} = \frac{\lambda_B L_B}{\lambda_A L_A} \in [0, 1]$.

Parameter	Description	Value [Unit]
C	Total Channel Capacity	1 [Mb/sec]
L_A	Message Length for Class A	1 [Mb]
L_B	Message Length for Class B	1 [Kb]
η_1	Priority Factor	10
η_2	Priority Factor	∞

Table 3.2: Model Parameters for Evaluating Mixed Traffic Performance.

3.5.1 Performance Plots Set 1

In our first set of performance plots, we keep the rate of message arrivals for Class A constant and we show the delay that Class B undergoes for varying message arrival rates. Table 3.3 summarizes the variables that were chosen for each plot.

Ratio $\frac{x}{C} = \frac{\lambda_A L_A}{C}$	Figure
0.5	3-6
0.6	3-7
0.7	3-8
0.8	3-9
0.9	3-10

Table 3.3: Table of Mixed Traffic Performance Figures for Constant $\frac{x}{C}$.

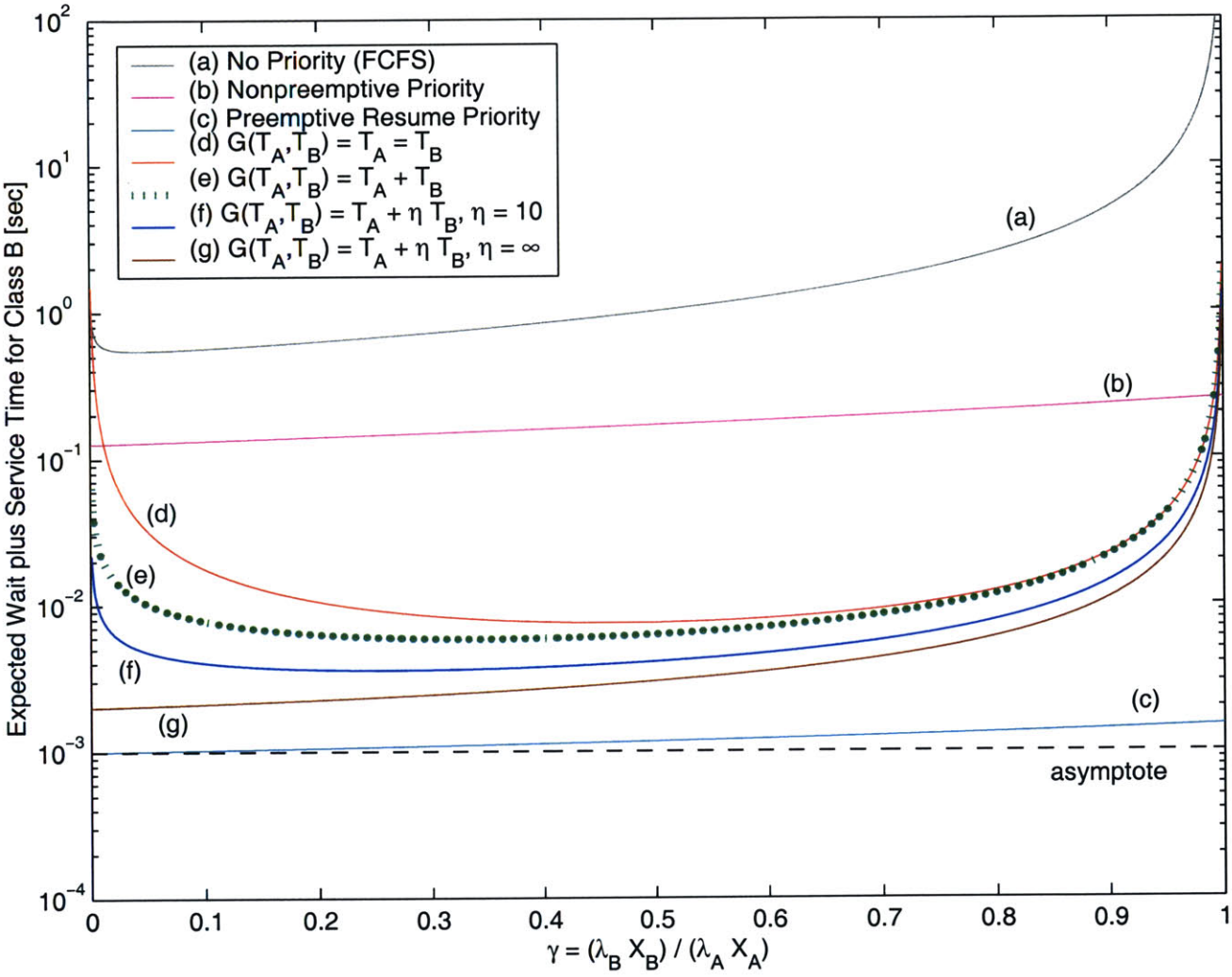


Figure 3-6: Performance Analysis 1 for Class B.
where $\frac{x}{C} = \frac{\lambda_A L_A}{C} = 0.5$

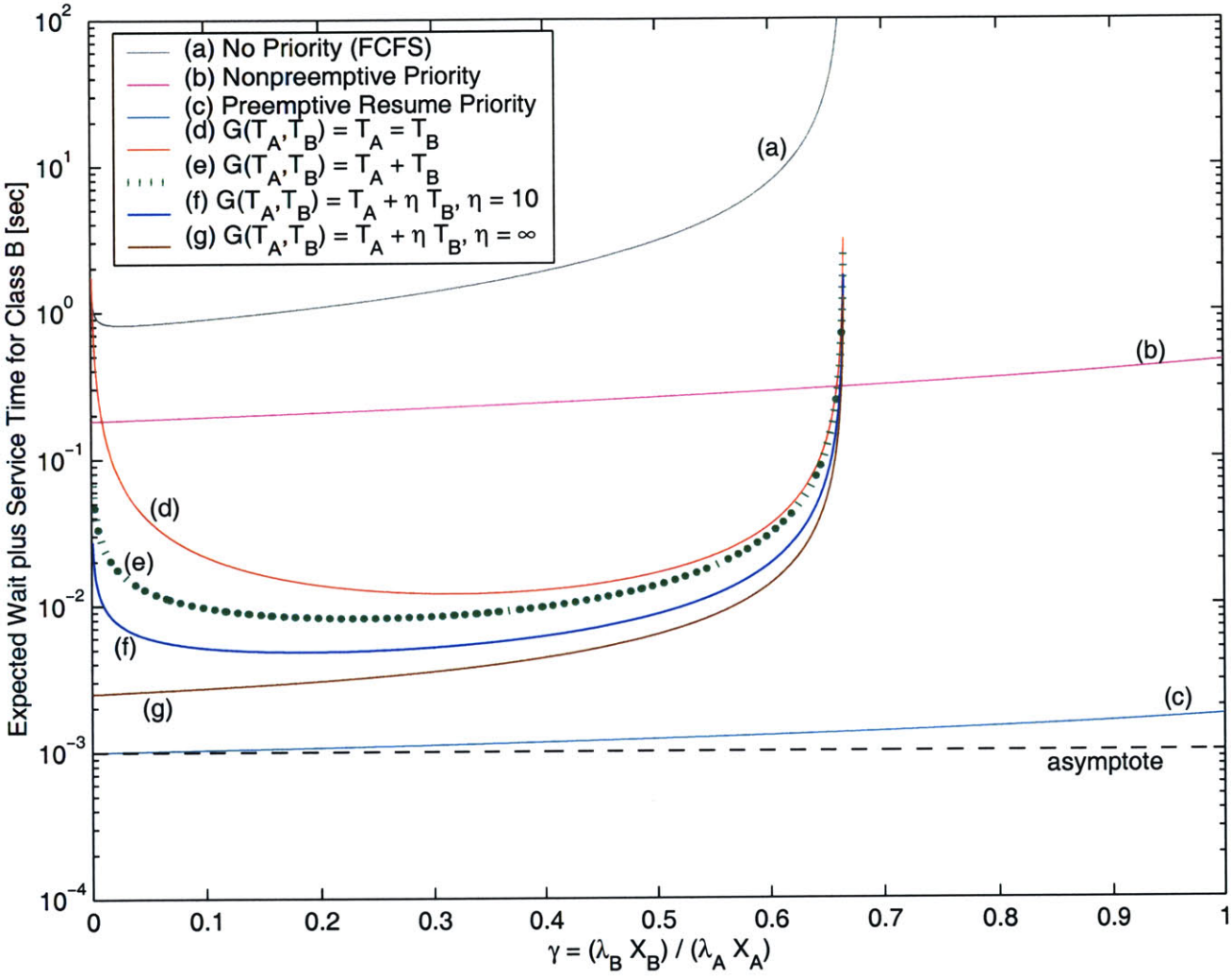


Figure 3-7: Performance Analysis 2 for Class B.
where $\frac{x}{C} = \frac{\lambda_A L_A}{C} = 0.6$

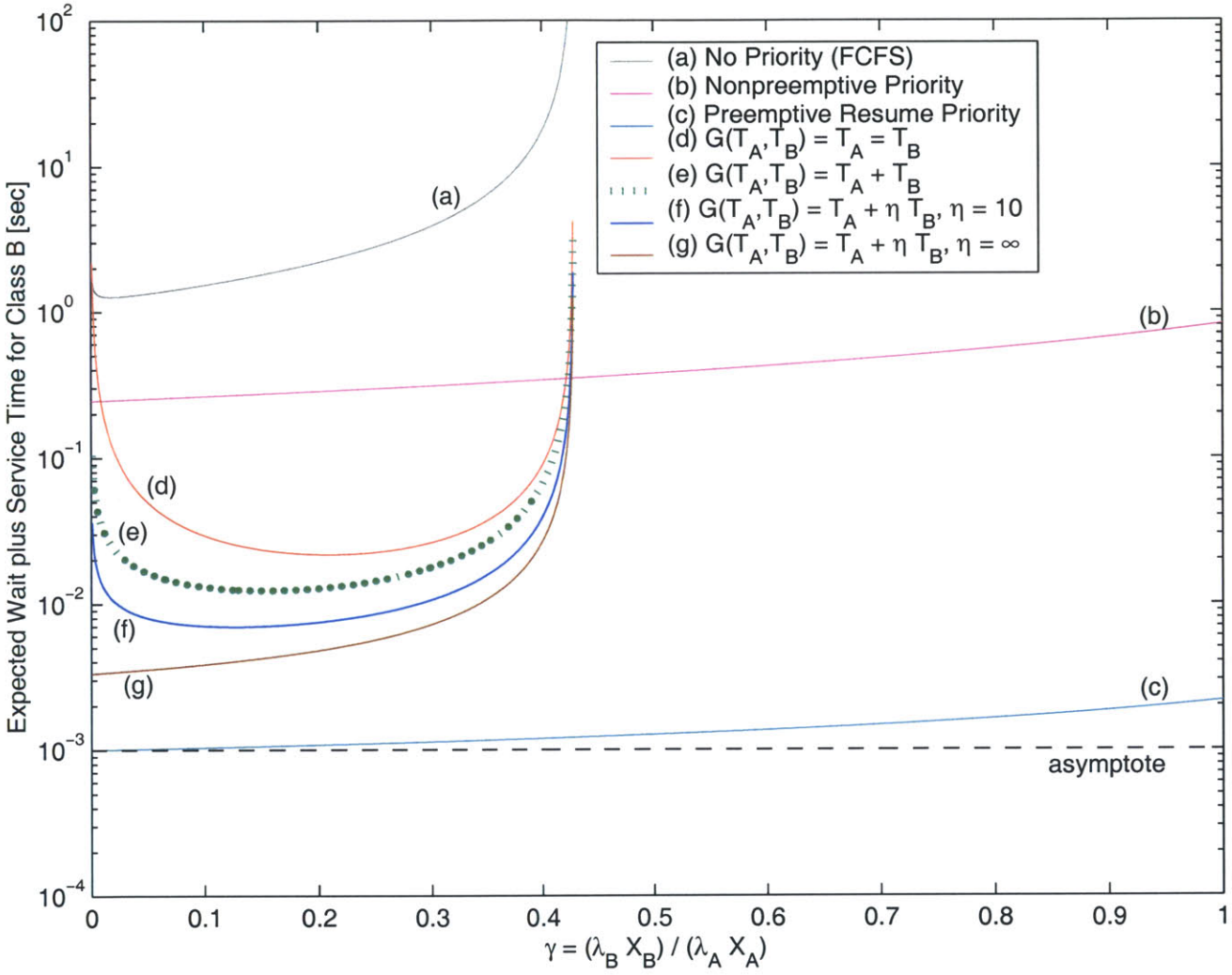


Figure 3-8: Performance Analysis 3 for Class B.
where $\frac{x}{C} = \frac{\lambda_A L_A}{C} = 0.7$

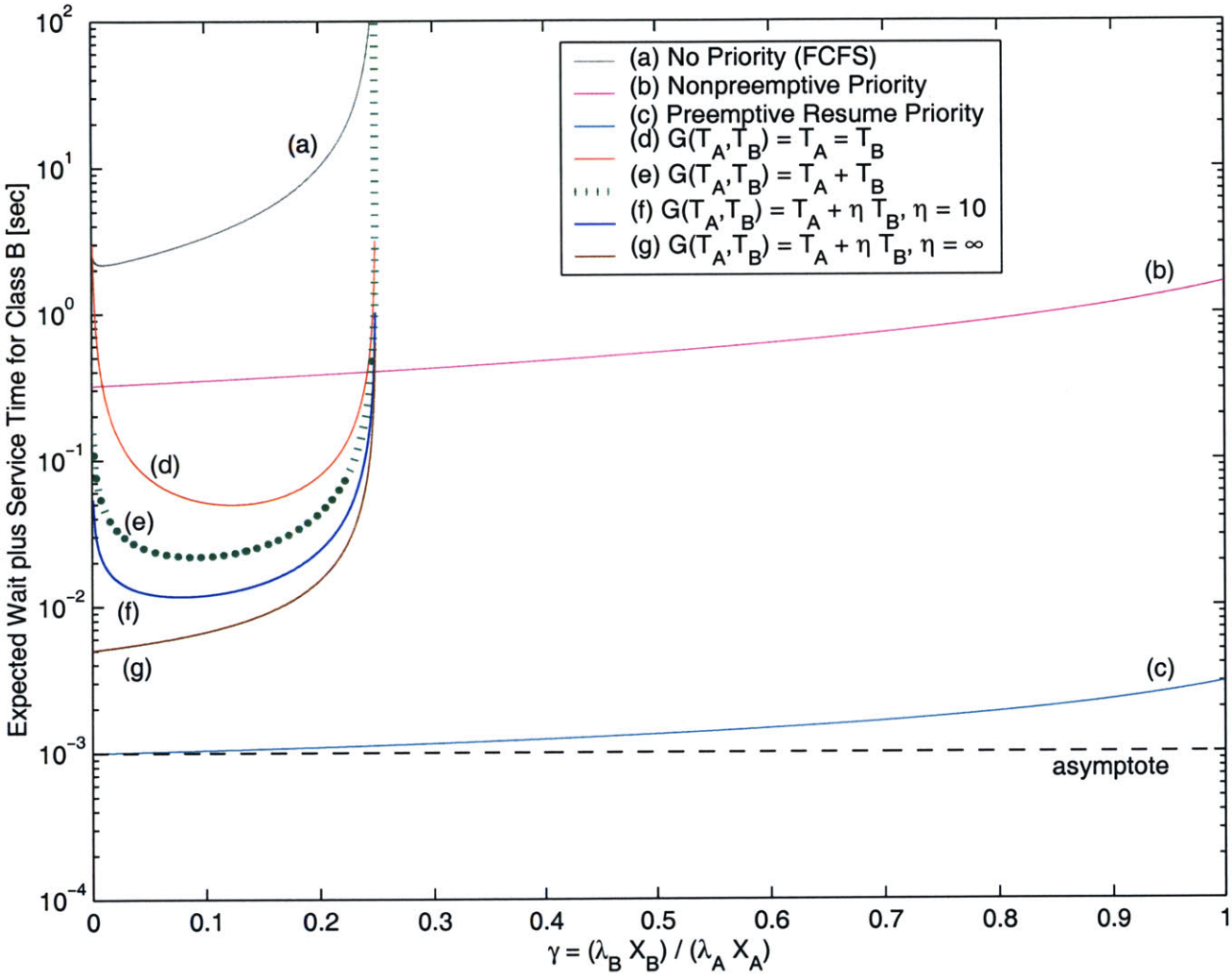


Figure 3-9: Performance Analysis 4 for Class B.
where $\frac{x}{C} = \frac{\lambda_A L_A}{C} = 0.8$

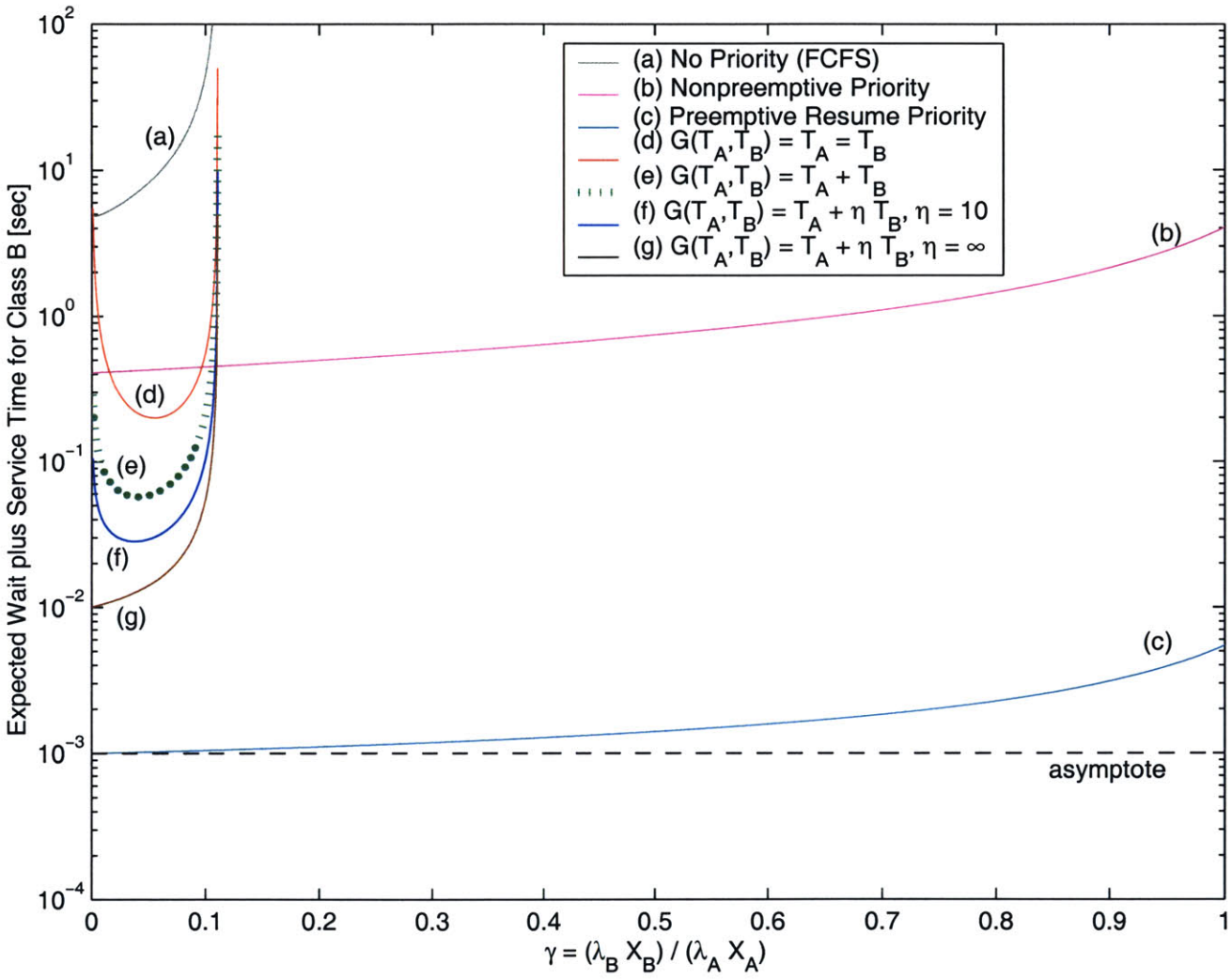


Figure 3-10: Performance Analysis 5 for Class B.
where $\frac{x}{C} = \frac{\lambda_A L_A}{C} = 0.9$

3.5.2 Performance Plots Set 2

In our second set of performance plots, we keep the rate of message arrivals for Class B constant and we show the delay that Class B undergoes for varying Class A message arrival rates. Table 3.4 summarizes the variables that were chosen for each plot.

Ratio $\frac{y}{C} = \frac{\lambda_B L_B}{C}$	Figure
0.1	3-11
0.2	3-12
0.3	3-13
0.4	3-14

Table 3.4: Table of Mixed Traffic Performance Figures for Constant $\frac{y}{C}$.

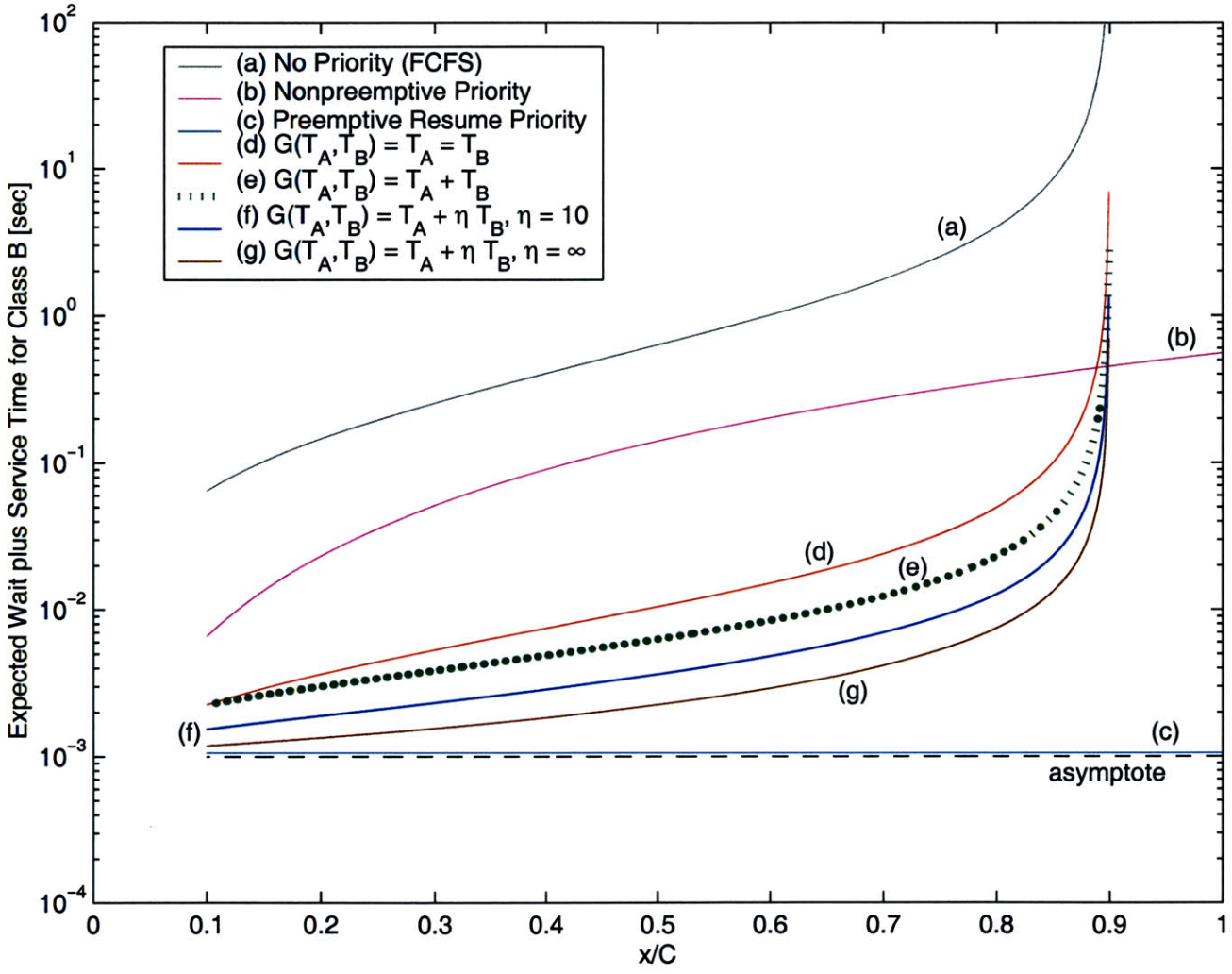


Figure 3-11: Performance Analysis 6 for Class B.
where $\frac{y}{C} = \frac{\lambda_B L_B}{C} = 0.1$

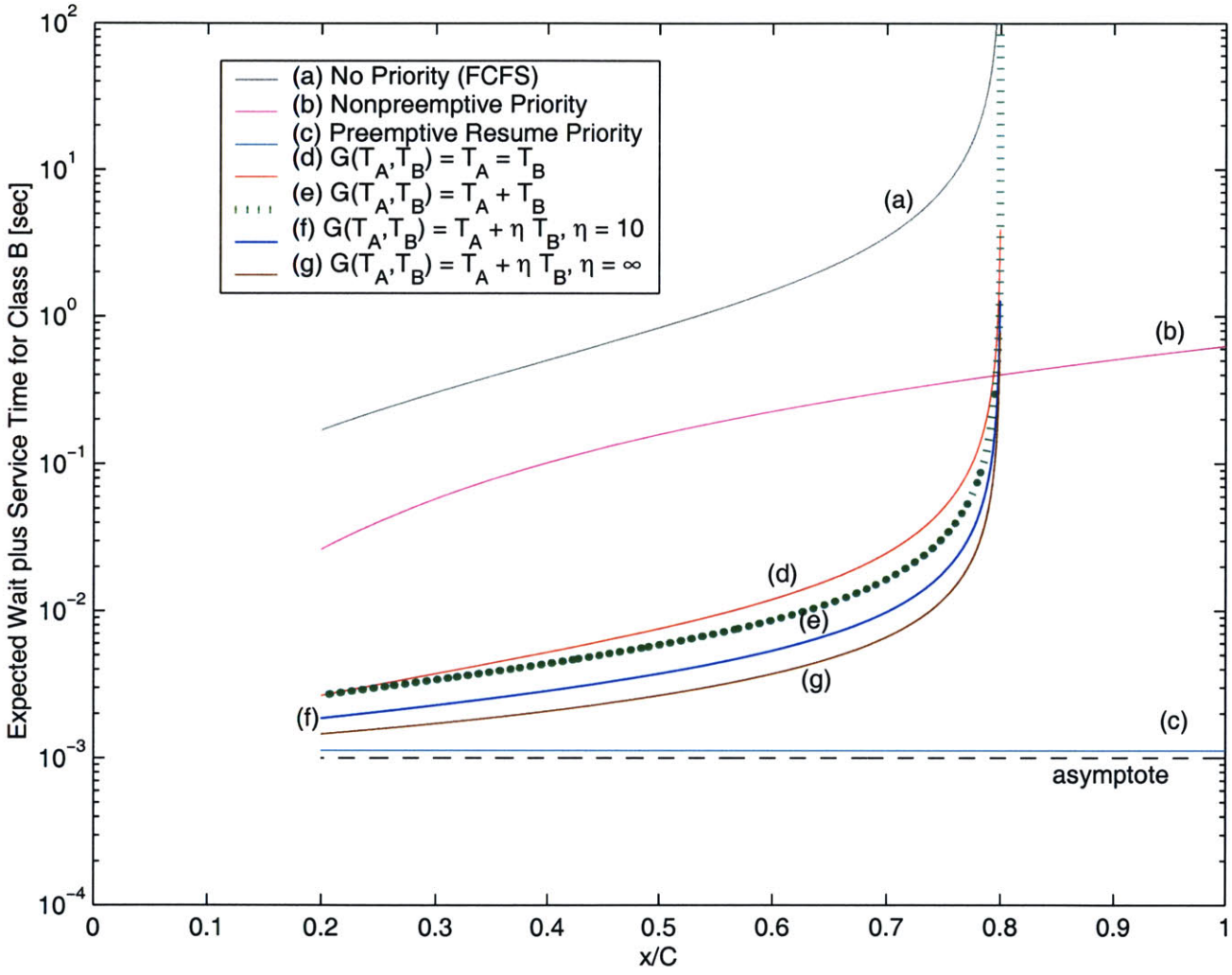


Figure 3-12: Performance Analysis 7 for Class B.
where $\frac{y}{C} = \frac{\lambda_B L_B}{C} = 0.2$

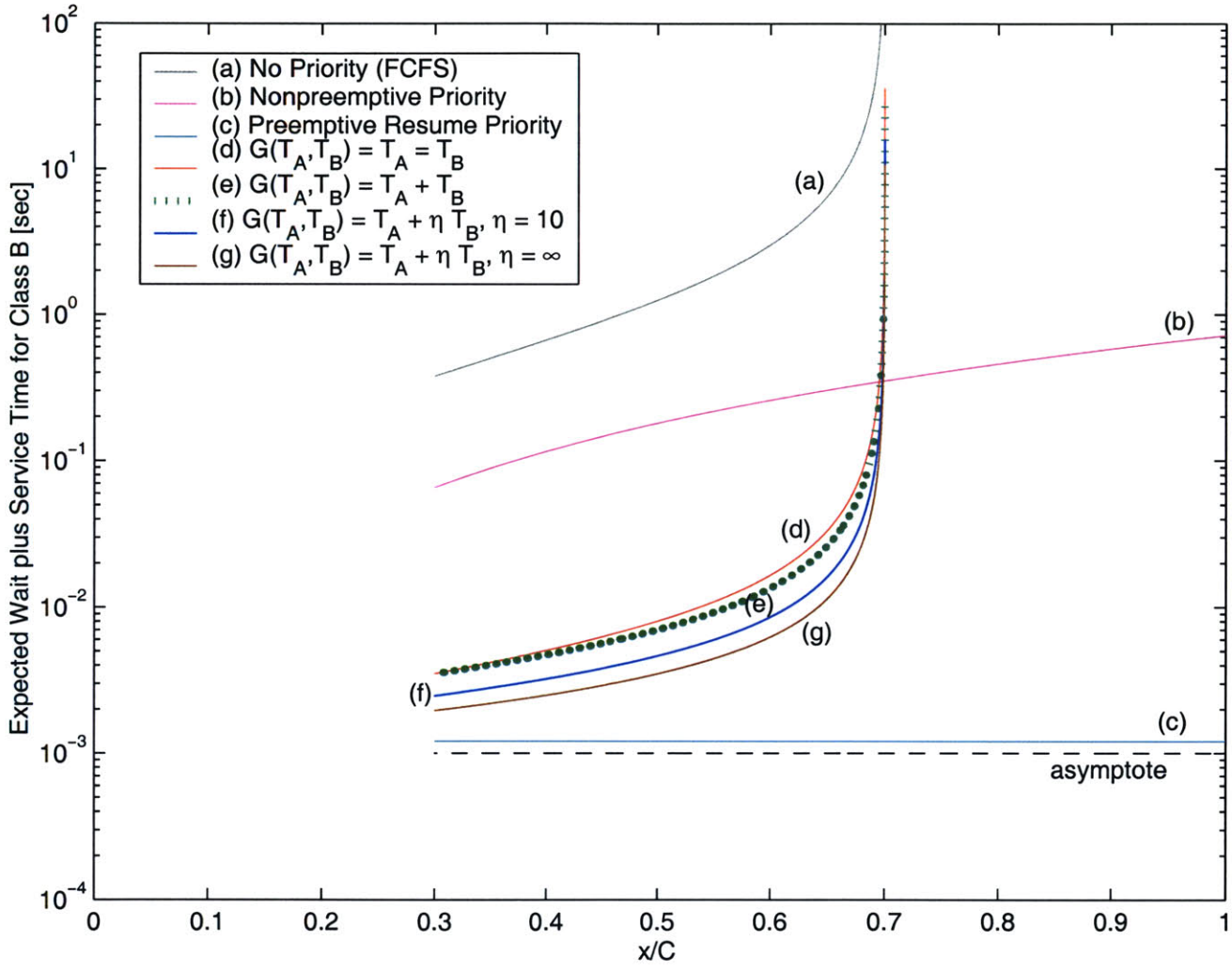


Figure 3-13: Performance Analysis 8 for Class B.
where $\frac{y}{C} = \frac{\lambda_B L_B}{C} = 0.3$

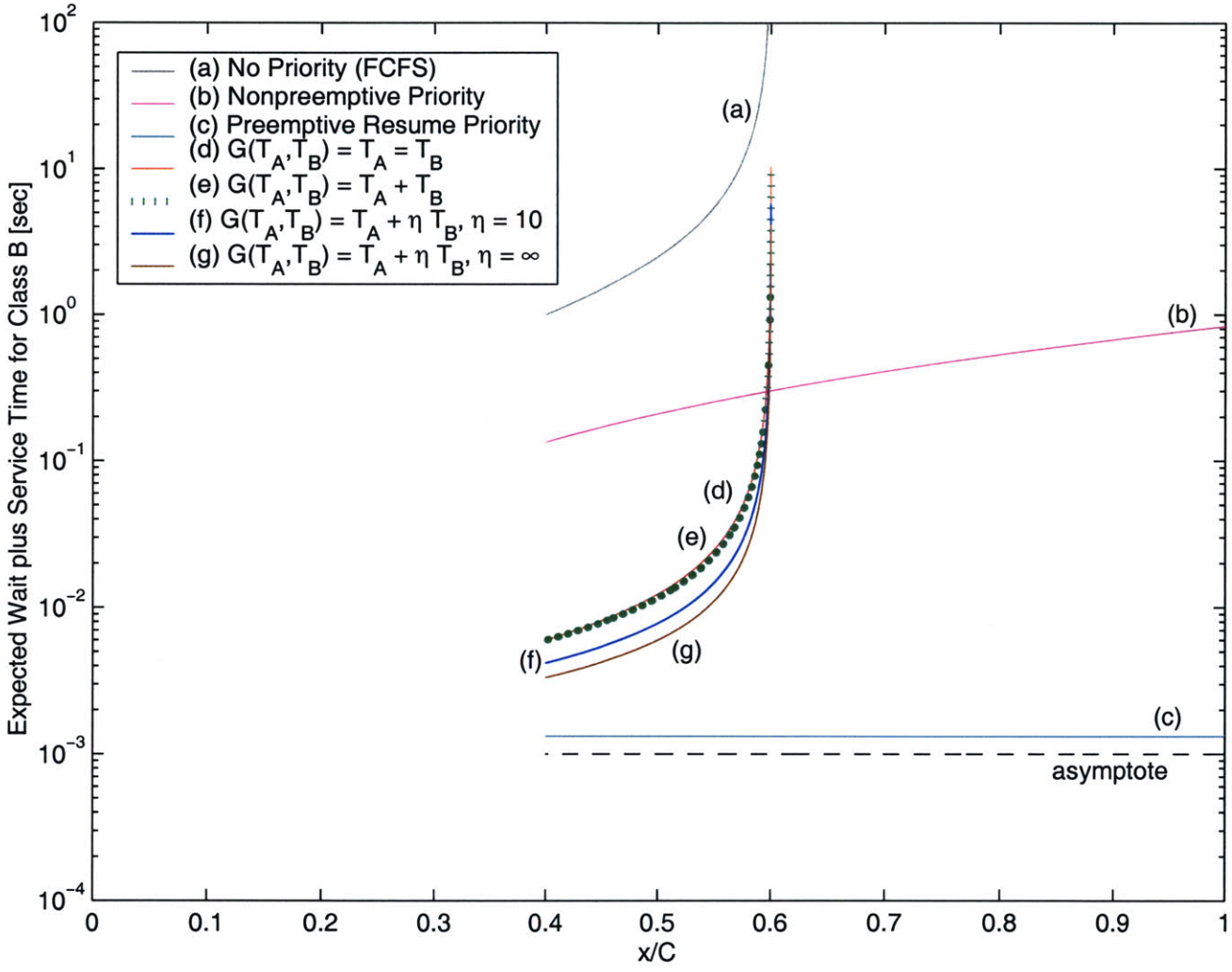


Figure 3-14: Performance Analysis 9 for Class B.
where $\frac{y}{C} = \frac{\lambda_B L_B}{C} = 0.4$

3.6 Interpretation of Results

We will summarize the results that can be observed by studying the delay performance charts and comparing the various priority and channel capacity allocation schemes.

3.6.1 Performance Plots Set 1

In our first set of performance plots (Figures 3-6 to 3-10), we keep the rate of message arrivals for Class A ($\frac{x}{C}$) constant and plot the time delay for Class B performance against the ratio $\gamma = \frac{y}{x} = \frac{\lambda_B \overline{X}_B}{\lambda_A \overline{X}_A}$. As γ increases from 0 to 1, we expect to see increasing delays for Class B users. We also plot a horizontal asymptote of $\frac{L_B}{C}$, the transmission time of a Class B message.

3.6.1.1 No Priority (FCFS)

$$\begin{aligned} T &= \frac{1}{2}\frac{2xC - x^2 - 4xy + y\lambda_A L_B + 2yC - y^2 + x\lambda_B L_A}{\lambda C(C - x - y)} \\ &= \frac{1}{2}\frac{2C - x - 4y + \gamma\lambda_A L_B + 2\gamma C - \gamma y + \lambda_B L_A}{x\lambda C(\frac{C}{x} - 1 - \gamma)}. \end{aligned} \tag{3.78}$$

We notice that having no priority is the worse scheme for Class B users. Users of Class B face the possibility of being stuck behind Class A users in the queue. In each of the figures, the vertical asymptote occurs at $\frac{x+y}{C} = 1$. FCFS also results in expected delay values approximately an order of magnitude greater than the nonpreemptive priority scheme.

3.6.1.2 Nonpreemptive Priority

$$\begin{aligned} T_B &= \frac{1}{2}\frac{2L_B C - yL_B + xL_A}{C(C - y)} \\ &= \frac{1}{2}\frac{2\frac{L_B C}{x} - \gamma L_B + L_A}{C(C - \gamma)}. \end{aligned} \tag{3.79}$$

With nonpreemptive priority, Class B users receive access to the channel first even if there are Class A messages in the queue. This priority technique provides a lower time delay than FCFS. The time delay for a Class B user, however, can still be

significant if a Class A user is being served upon arrival. Class B users must wait until the Class A transmission is completed before receiving access to the channel. The vertical asymptote here occurs at $\frac{y}{C} = 1$ since we continue to serve Class B users as long as they keep arriving in the queue.

3.6.1.3 Preemptive Resume Priority

$$\begin{aligned} T_B &= \frac{1}{2}\frac{L_B(2C-y)}{C(C-y)} \\ &= \frac{L_B}{C}\frac{2C-2y+y}{2C-2y} \\ &= \frac{L_B}{C}\left[1+\frac{1}{2}\frac{y}{C-y}\right] \\ &= \frac{L_B}{C}\left[1+\frac{1}{2}\frac{\gamma}{\frac{C}{x}-\gamma}\right]. \end{aligned} \tag{3.80}$$

A significant improvement in time performance can be seen with the use of preemptive resume priority. Here, Class B users are given access to the channel whenever it enters the system. They can interrupt the transmission of a Class A message. Therefore, the time delay is very close to the plotted asymptote which measures the transmission time of a Class B message on the channel. However, it is not identical to the horizontal asymptote because as an increasing number of Class B users arrive, the queueing delay will also increase. Again, the vertical asymptote occurs at $\frac{y}{C} = 1$ since we continue to serve Class B users as long as they keep arriving in the system. Since preemptive resume priority cannot be easily implemented for wireless broadcast systems, we turn to other schemes to improve the delay performance of Class B users.

3.6.1.4 Channel Capacity Allocation

The performance of our channel capacity allocation schemes fall within the extremes of nonpreemptive priority and preemptive resume priority. With these schemes, we notice that the delay performance curves increases at $\gamma \approx 0$. This is due to the fact that the optimum channel capacity allocation value (θ^*) is very high for Class A users. With a very small portion of channel capacity available to Class B, the time to transmit a message is significant. In these cases, the vertical asymptote occurs at $\frac{x+y}{C} = 1$.

3.6.1.4.1 $G(T_A, T_B) = T_A = T_B$

$$T_B = \frac{1}{2}\frac{1+\gamma}{\lambda_B \frac{C}{x}}\left[2 + \frac{1+\gamma}{\frac{C}{x} - (1+\gamma)}\right].$$

In this case, we equate the normalized delays for Class A and Class B users. We then solve for the optimum channel capacity allocation value (θ^*) and plug it back into the non-normalized delay for Class B.

3.6.1.4.2 $G(T_A, T_B) = T_A + T_B$

$$T_B = \frac{1}{2}\frac{L_B(1+\sqrt{\gamma})}{C - x - x\sqrt{\gamma}}\left[2 + \frac{1+\sqrt{\gamma}}{\frac{C}{y}\sqrt{\gamma} - \frac{1}{\sqrt{\gamma}} - \sqrt{\gamma}}\right].$$

In this case, we sum the normalized delays for Class A and Class B users. We then solve for the optimum channel capacity allocation value (θ^*) and plug it back into the non-normalized delay for Class B.

3.6.1.4.3 $G(T_A, T_B) = T_A + \eta T_B$

$$= \frac{1}{2}\frac{L_B(1+\sqrt{\eta\gamma})}{C\sqrt{\eta\gamma} - x\sqrt{\eta\gamma} + y}\left[2 + \frac{1+\sqrt{\eta\gamma}}{\sqrt{\eta\gamma}(\frac{C}{y} - \frac{1}{\gamma} - 1)}\right].$$

In this case, we sum the normalized delays for Class A and Class B users with an η multiplier to the time delay for Class B. The η factor is used to increase or decrease the priority of Class B messages. To increase the priority we use $\eta > 1$, thus to decrease the priority we use $\eta < 1$. We then solve for the optimum channel capacity allocation value (θ^*) and plug it back into the non-normalized delay for Class B.

We generate two curves for this case. First we plot the performance for Class B with $\eta = 10$. Then, we plot the performance for Class B with $\eta = \infty$. When $\eta = \infty$,

T_B reduces to

$$\begin{aligned} T_B &= \frac{1}{2}\frac{L_B(2C-2x-y)}{(C-x-y)(C-x)} \\ &= \frac{L_B}{C-x}\frac{2C-2x-2y+y}{2C-2x-2y} \\ &= \frac{L_B}{C-x}\left[1+\frac{y}{C-x-y}\right] \\ &= \frac{L_B}{C-x}\left[1+\frac{\gamma}{\frac{C}{x}-(1+\gamma)}\right]. \end{aligned} \tag{3.81}$$

3.6.2 Performance Plots Set 2

In our second set of performance plots (Figures 3-11 to 3-14), we keep $\frac{y}{C}$ constant and plot the time delay for Class B performance against the ratio $\frac{x}{C}$. As $\frac{x}{C}$ increases from 0 to 1, we expect to see increasing delays for Class B users. We also plot a horizontal asymptote of $\frac{L_B}{C}$, the transmission time of a Class B message. Due to our system constraint of fixing $\frac{y}{C}$, all the plots have a starting point at $\frac{y}{C}$.

3.6.2.1 No Priority (FCFS)

$$T = \frac{1}{2}\frac{2C-x-4y+\gamma\lambda_A L_B+2\gamma C-\gamma y+\lambda_B L_A}{x\lambda C(\frac{C}{x}-1-\gamma)}.$$

Again, we notice that having no priority is the worse scheme for Class B users. Users of Class B face the possibility of being stuck behind Class A users in the queue. Here, there is a vertical asymptote which occurs $\frac{x+y}{C}=1$. FCFS results in expected delay values approximately an order of magnitude greater than the nonpreemptive priority scheme.

3.6.2.2 Nonpreemptive Priority

$$T_B = \frac{1}{2}\frac{2\frac{L_B C}{x}-\gamma L_B+L_A}{C(C-\gamma)}.$$

With nonpreemptive priority, Class B users receive access to the channel first even if there are Class A messages in the queue. This priority techniques provides a lower time delay than FCFS. The time delay for a Class B user, however, can still be significant if a Class A user is being served upon arrival. Class B users must wait until the Class A transmission is completed before receiving access to the channel. A vertical asymptote occurs at $\frac{y}{C} = 1$, since we continue to serve Class B users as long as they keep arriving in the queue.

3.6.2.3 Preemptive Resume Priority

$$T_B = \frac{L_B}{C}\left[1 + \frac{1}{2}\frac{\gamma}{\frac{C}{x} - \gamma}\right].$$

A significant improvement in time performance can be seen with the use of preemptive resume priority. Here, Class B users are given access to the channel whenever it enters the system. They can interrupt the transmission of a Class A message. Therefore, the time delay is very close to the plotted asymptote which measures the transmission time of a Class B message on the channel. Again, there is a vertical asymptote at $\frac{y}{C} = 1$ since we continue to serve Class B users as long as they keep arriving in the system. Since preemptive resume priority cannot be easily implemented for wireless broadcast systems, we turn to other schemes to improve the delay performance of Class B users.

3.6.2.4 Channel Capacity Allocation

The performance of our channel capacity allocation schemes fall within the extremes of nonpreemptive priority and preemptive resume priority. In these cases, a vertical asymptote occurs at $\frac{x+y}{C} = 1$.

3.6.2.4.1 $G(T_A, T_B) = T_A = T_B$

$$T_B = \frac{1}{2}\frac{1+\gamma}{\lambda_B \frac{C}{x}}\left[2 + \frac{1+\gamma}{\frac{C}{x} - (1+\gamma)}\right].$$

In this case, we equate the normalized delays for Class A and Class B users. We then solve for the optimum channel capacity allocation value (θ^*) and plug it back into the non-normalized delay for Class B.

3.6.2.4.2 $G(T_A, T_B) = T_A + T_B$

$$T_B = \frac{1}{2}\frac{L_B(1+\sqrt{\gamma})}{C - x - x\sqrt{\gamma}}\left[2 + \frac{1+\sqrt{\gamma}}{\frac{C}{y}\sqrt{\gamma} - \frac{1}{\sqrt{\gamma}} - \sqrt{\gamma}}\right].$$

In this case, we sum the normalized delays for Class A and Class B users. We then solve for the optimum channel capacity allocation value (θ^*) and plug it back into the non-normalized delay for Class B.

3.6.2.4.3 $G(T_A, T_B) = T_A + \eta T_B$

$$T_B = \frac{1}{2}\frac{L_B(1+\sqrt{\eta\gamma})}{C\sqrt{\eta\gamma} - x\sqrt{\eta\gamma} + y}\left[2 + \frac{1+\sqrt{\eta\gamma}}{\sqrt{\eta\gamma}(\frac{C}{y} - \frac{1}{\gamma} - 1)}\right].$$

In this case, we sum the normalized delays for Class A and Class B users with an η multiplier to the time delay for Class B. The η factor is used to increase or decrease the priority of Class B messages. To increase the priority we use $\eta > 1$, thus to decrease the priority we use $\eta < 1$. We then solve for the optimum channel capacity allocation value (θ^*) and plug it back into the non-normalized delay for Class B.

We generate two curves for this case. First we plot the performance for Class B with $\eta = 10$. Then, we plot the performance for Class B with $\eta = \infty$. When $\eta = \infty$, T_B is

$$T_B = \frac{L_B}{C - x}\left[1 + \frac{\gamma}{\frac{C}{x} - (1+\gamma)}\right].$$

Note that when $\eta = \infty$, the performance curve is close to the preemptive resume priority priority for $\frac{x+y}{C} < 0.8$ before it starts to depart due to capacity limitations for Class B users.

3.7 Summary

We have introduced the topic of priority queueing and channel capacity allocation in this chapter. We narrowed the list of priority schemes and allocation schemes to model, compare, and analyze. Modeling allows us to make assumptions to create a model that is simple yet efficiently true to the real system so that the answers provided by the model have some credibility. In our model of mixed traffic, Class B users are those who have small messages to transmit. The figures show how the

different techniques affect the performance of Class B. In short, this thesis provides a methodology of determining resource allocation by incorporating system constraints and QoS priority. We should however be aware that increasing the performance of one class will come at a cost to the other classes of users.

Chapter 4

Conclusions

The communication systems today are getting better, with better voice quality, better security, more services and data capabilities, and an increase in capacity. The next generation of radio systems, to some extent, is simply an opportunity to update existing mobile radio systems as advances in technology and manufacturing occur. Currently the second generation standards will evolve further and form the basis for the third generation networks. Given the diverse needs and the pace of development in different regions, a single global approach does not seem adequate. The implementation of wireless broadband communication systems requires the following considerations: frequency allocation and selection, channel characterization, application and environment recognition, including technology development, air interface multiple access techniques, protocols and networks, systems development with efficient modulation, coding, and smart antenna techniques. The focus of this thesis is multiple access techniques for multichannel communication systems channel resource allocation for mixed traffic. Although the techniques developed here are used for networks with wireless channels, they are also applicable for satellite and optical networks.

The topic of multiple access communication has been developed with a focus on fixed access schemes and random access schemes. Examples of fixed access techniques are Frequency Division Multiple Access (FDMA) and Time Division Multiple Access (TDMA), in which each user is permanently assigned a fixed portion of the channel, either on a frequency basis (FDMA) or time basis (TDMA). These techniques are efficient for predictable streams of traffic, but in cases of bursty traffic, i.e., when the peak to average data rate is high, fixed allocation of channel capacity is wasteful. Random access techniques like pure ALOHA, or Slotted ALOHA (S-ALOHA), are introduced to cope with this type of short, high-rate traffic bursts occurring at random points in time. ALOHA communication is an uncoordinated protocol. It has low efficiency since packets can overlap causing collisions and lost data. Slotted ALOHA has an efficiency limited to $\frac{1}{e}$ (approximately 0.368) successful transmissions per slot,

and has stability problems if not controlled.

In our analysis in Chapter 2, we note that the S-ALOHA and TDMA schemes, independent of propagation delay, give best performance at light and heavy traffic loads, respectively. In the case of multichannel communication systems, we suggest allocating one subchannel as a reservation channel and the remaining subchannels as data channels. Implementing a reservation channel allows the communication system to be more fully utilized since channel resources are never given to non-active users. We note that given the same amount of capacity for data transmission, it is best to minimize the number of subchannels thereby creating subchannels with larger capacity rates. The ability to transmit data on larger transmission subchannels significantly reduce the total average delay for queueing and service.

In many multi-user communication systems, the channel allocation policy plays a key role in determining the overall system capacity, i.e., the maximum load that the system can carry for a given spectrum resource. The number of users are potentially unlimited in number which implies that resources must be efficiently assigned and utilized. The purpose of a channel allocation policy is, thus, to distribute the channel among users in such a way to satisfy Quality of Service (QoS) and achieve maximum channel utilization. Although an optimal solution can be determined analytically, in practice, the evaluation of an optimal solution is prevented by real-life complexities that cannot be modeled. Hence, the need for sub-optimal policies.

In our analysis in Chapter 3, we analyze channel capacity allocation schemes for wireless networks supporting mixed traffic. With the assumption of multiple classes of users, we analyzed systems that use priority schemes such as nonpreemptive priority and preemptive resume priority. We develop alternative schemes to distribute channel resources among the different user classes. The amount of channel capacity given to each user class is optimized according to different timing metrics. We can suggest schemes that are better than nonpreemptive priority technique and closely approximate the preemptive resume priority technique, which in many instances can be difficult to implement. Depending on the constraints of the communication system and the priority scheme that is to be given to user classes, Chapter 3 can provide a good understanding of the methodology of optimizing the allocation of channel capacity.

Appendix A

Queueing Theory

A.1 Overview

The basic methodological framework for analyzing network delay is queueing theory. To understand the nature and mechanism of delay in a network, it is often necessary to make simplistic assumptions. More realistic assumptions make the analysis extremely difficult. Queueing models have been developed to provide a basis for delay approximations and provide valuable qualitative results and insights.

We will focus on packet delay within a communication network. A packet undergoes delay that can be separated into the following delay components:

1. The processing delay, which is the time the packet is correctly received at the source and the time the packet is transmitted on the communication link.
2. The queueing delay, which is the time the packet is waiting in a queue for transmission and the time it begins to transmit.
3. The transmission duration, which is the time it takes to transmit the entire packet on the communication link.
4. The propagation delay, which is the time it takes for the last bit to be transmitted at the head node of the communication link and received at the tail node. Propagation delay proportional to the physical distance between the source and destination.

We start by introducing a standard nomenclature for single-station queues, i.e. queueing systems where users form a single queue. Such a queue is described as follows:

1. Arrival Process - We assume that users arrive one at a time, and the successive interarrival times are independent and identically distributed (iid). Thus the arrival process is a renewal process. It is described by the distribution of the interarrival times, represented by special symbols as follows:
   - M: Exponential M stands for memoryless, which is assumed to be a Poisson process, i.e., exponentially distributed interarrival times,
   - G: General G stands for a general distribution of interarrival times,
   - D: Deterministic D stands for a deterministic interarrival times,
2. Service Times - We assume that the service times of successive users are iid. They are represented by the same letters as the interarrival times.
3. Number of servers - Typically denoted by k. All the servers are assumed to be identical, and that any user can be served by any server.
4. Maximum number of users - Denoted by N. It includes the number of users in service. If an arriving user finds N users in the system, he/she is permanently lost. If capacity is not mentioned, it is assumed to be infinite.

As noticed by the contents of this thesis, we have used queueing models to develop expected delay analysis while investigating multiple access techniques and uplink architectures. What follows is a brief summary of the queueing delay models $Q_{M/M/k}(\cdot)$, $Q_{M/G/k}(\cdot)$, and $Q_{M/D/k}(\cdot)$. For complete derivations and a strong understanding of queueing theory, the reader should consult sources such as [4], [22], and [39].

A.2 M/M/k Queue

A.2.1 M/M/1

An M/M/1 model assumes (1) Poisson arrivals of rate λ, (2) independent and exponential service distribution with mean $\frac{1}{\mu}$[sec] and rate $\mu > \lambda$, and (3) a single server. The utilization factor ρ $(0 \leq \rho \leq 1)$ is defined as

$$\rho = \frac{\lambda}{\mu}. \tag{A.1}$$

The expected delay an arrival experiences upon entering the system is

$$Q_{M/M/1}(\lambda, \mu) = \frac{\rho}{\mu - \lambda}. \tag{A.2}$$

Thus the total expected delay an arrival experiences in the system is

$$\begin{aligned} T &= \frac{\rho}{\mu - \lambda} + \frac{1}{\mu} \\ &= \frac{1}{\lambda - \mu}. \end{aligned} \tag{A.3}$$

A.2.2 M/M/k

The M/M/k queuing system is identical to the M/M/1 system except that there are k servers, or channels of a transmission line. A user at the head of the queue is routed to any server that is available. The utilization factor ρ ($0 \leq \rho \leq 1$) is defined as

$$\rho = \frac{\lambda}{k\mu}. \tag{A.4}$$

The probability of n users in the system is

$$p_0 = \left[\sum_{k=0}^{k-1} \frac{(k\rho)^n}{n!} + \frac{(k\rho)^k}{k!(1-\rho)} \right]^{-1}. \tag{A.5}$$

The probability that an arrival has to wait in the queue is

$$P_Q = \frac{p_0 (k\rho)^k}{k!(1-\rho)}. \tag{A.6}$$

The expected delay an arrival experiences upon entering the system is

$$Q_{M/M/k}(\lambda, \mu, k) = \frac{\rho P_Q}{\lambda(1-\rho)}. \tag{A.7}$$

Thus the total expected delay an arrival experiences in the system is

$$T = \frac{\rho P_Q}{\lambda(1-\rho)} + \frac{1}{\mu}. \tag{A.8}$$

A.2.3 M/M/∞

In an M/M/∞ system, there is no waiting in queue. We assume that a new server is created immediately to handle a new arrival. Thus, the expected delay an arrival

experiences is zero and the model behaves as follows

$$Q_{M/M/\infty} = 0, \tag{A.9}$$

$$T = \frac{1}{\mu}. \tag{A.10}$$

A.3 M/G/k Queue

A.3.1 M/G/1 Queue

The M/G/1 model assumes (1) Poisson arrivals of rate λ, (2) a general service distribution which is iid service times with mean $\overline{X}$ and second moment $\overline{X^2} = \overline{X}^2$, and (3) a single server with infinite waiting room.

$$\overline{X} = E\{X\} = \text{Expected service time}, \tag{A.11}$$

$$\overline{X^2} = E\{X^2\} = \text{Second moment of service time}, \tag{A.12}$$

$$\rho = \lambda\overline{X}, \tag{A.13}$$

$$Q_{M/G/1}(\lambda, \overline{X}, \overline{X^2}) = \frac{\lambda\overline{X^2}}{2(1-\lambda\overline{X})}. \tag{A.14}$$

A.3.2 M/G/k Approximation

In this section we consider the M/G/k system in which users arrive at a Poisson rate λ and are served by any of k servers, each of whom has the service distribution G.

$$\overline{X} = E\{X\} = \text{Expected service time}, \tag{A.15}$$

$$\overline{X^2} = E\{X^2\} = \text{Second moment of service time}, \tag{A.16}$$

$$\rho = \frac{\lambda\overline{X}}{k}, \tag{A.17}$$

$$Q_{M/G/k}(\lambda, \overline{X}, \overline{X^2}, k) \approx \frac{\lambda^k \overline{X^2} (\overline{X})^{k-1}}{2(k-1)!(k-\lambda\overline{X})^2 (\sum_{n=0}^{k-1} \frac{(\lambda\overline{X})^n}{n!} + \frac{(\lambda\overline{X})^k}{(k-1)!(k-\lambda\overline{X})})}. \tag{A.18}$$

The M/G/k approximation is an exact solution if the service time G is exponential [39].

A.4 M/D/k Queue

When service times are identical for all arrivals, we have

$$\overline{X} = \frac{1}{\mu} \tag{A.19}$$

$$\overline{X^2} = \frac{1}{\mu^2} \tag{A.20}$$

A.4.1 M/D/1

Since the M/D/1 case yields the minimum possible value of $\overline{X^2}$ for given μ, it follows that the values of $Q_{M/D/1}$ and T for an M/D/1 queue are lower bounds to the corresponding quantities for an M/G/1 queue of the same λ and μ.

$$Q_{M/D/1}(\lambda, \mu) = Q_{M/G/1}(\lambda, \frac{1}{\mu}, \frac{1}{\mu^2}) \tag{A.21}$$

A.4.2 M/D/k

$$Q_{M/D/k}(\lambda, \mu, k) = Q_{M/G/k}(\lambda, \frac{1}{\mu}, \frac{1}{\mu^2}, k) \tag{A.22}$$

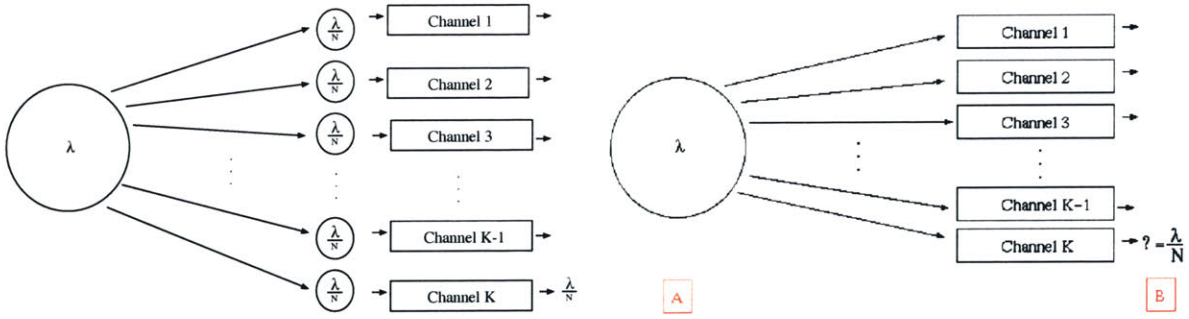
Appendix B

Proofs

B.1 Poisson Process

Case 1 is a multichannel system that assumes K channels each with equal capacity $\frac{C}{K}$ [bits/sec]. The user population is divided into K equally sized groups with each group accessing one channel, as shown in Figure B-1(a). Message arrivals are assumed to follow a Poisson distribution. The Poisson process is defined by a probabilistic description of the behavior of arrivals at points on a continuous time line

$$P[n \text{ arrivals, time } \tau] = \frac{(\lambda\tau)^n}{n!}e^{-\lambda\tau} \text{ where } n = 0, 1, 2, ... \text{ and } \tau > 0. \tag{B.1}$$



(a) Case 1.

(b) Case 2.

Figure B-1: Comparison between Two Poisson Cases.

Case 2 is a multichannel system that assumes K channels each with equal capacity $\frac{C}{K}$ [bits/sec]. Every user and for every packet transmission (or retransmission) selects randomly, uniformly, and independently of the past the channel over which this specific packet is to be transmitted.

The following is to illustrate that the arrivals from point A to point B in Figure B-1(b) is equal to the corresponding arrivals in Figure B-1(a)

$$
\begin{aligned}
P[\text{n arrivals at B, time } \tau] &= P[\text{n arrivals at A}]P[\text{all n chooses B}] \\
&\quad +P[\text{n+1 arrivals at A}]P[\text{n of these chooses B}] \\
&\quad +P[\text{n+2 arrivals at A}]P[\text{n of these choose B}] + ... \\
&= f(n)\frac{1}{K^n} + f(n+1)\frac{1}{K^n}\frac{K-1}{K}\binom{n+1}{n} \\
&\quad +f(n+2)\frac{1}{K^n}(\frac{K-1}{K})^2 + ... \\
&= \frac{1}{n!}\frac{1}{K^n}(\lambda\tau)^n e^{-\lambda\tau}[1 + \frac{\lambda\tau}{(n+1)!}\frac{K-1}{K}\frac{(n+1)!}{1!} \\
&\quad +\frac{(\lambda\tau)^2}{(n+2)!}(\frac{K-1}{K})^2\frac{(n+2)!}{2!} + ...] \\
&= \frac{1}{n!}\frac{1}{K^n}(\lambda\tau)e^{-\lambda\tau}[1 + \lambda\tau\frac{K-1}{K} + \frac{(\lambda\tau\frac{K-1}{K})^2}{2!} + ...] \\
&= \frac{1}{n!}\frac{1}{K^n}(\lambda\tau)^n e^{-\lambda\tau}e^{\lambda\tau\frac{K-1}{K}} \\
&= \frac{(\frac{\lambda}{K}\tau)^n}{n!}e^{-\frac{\lambda\tau}{K}} \qquad \text{(B.2)}
\end{aligned}
$$

where $f(n) = \frac{(\lambda\tau)^n}{n!}e^{-\lambda\tau}$. From Eq. B.2, we can see that the arrival rate at B is $\lambda = \frac{\lambda}{K}$.

B.2 Optimum Number of Channels

The proof of determining the optimal number of channels by dividing channel capacity can be found in [18]. It is found that to minimize overall expected message transmission delay, the optimal number of channels K is one.

Assume that we have K subchannels, each of capacity $\frac{C}{K}$ [bits/sec], as shown in

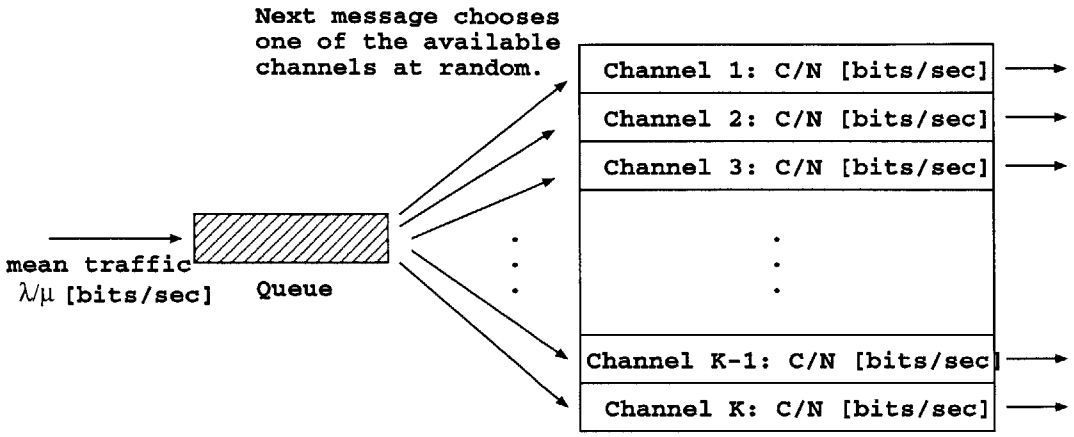


Figure B-2: Messages Transmitting on Multiple Channels.

Figure B-2. There is a queue with Poisson arrivals at an average arrival rate of λ messages per second. All message lengths are exponentially distributed with mean length $\frac{1}{\mu}$ [bits]. A first-come first-serve (FCFS) discipline is implemented on the queue, where messages at the head of the queue gain access to the first channel that becomes available. If more than one channel is available, the message chooses from this set randomly according to a uniform distribution.

Given values for λ, μ, and the total capacity of C [bits], the issue is to determine K, the total number of subchannels. The value of K should minimize T [sec], the expected delay spent in the system, i.e., the queueing delay plus transmission delay. Channel utilization is again defined as $\rho = \frac{\lambda}{\mu C}$ [bits/sec].

Theorem 1
The value of K which minimized T for all $0 \leq \rho < 1$ *is N=1.*

A system with more than one channel is non-optimum because the efficiency of a message is related to its transmitting rate. With one channel, we can transmit at a rate of C [bits/sec] whenever there are any messages in the system. If we have K channels ($K > 1$), there will be situations in which less than K channels are occupied, and we shall then be transmitting at a rate less than C [bits/sec].

This result implies that whenever possible one should design a multichannel system, where total capacity is fixed, with as few channels as the physical constraints of the network allow. The limiting case of one channel is optimum. Pragmatic engineering considerations however may drive the design of systems from this ideal optimum.

Bibliography

[1] Norman Abramson. The aloha system - another alternative for computer communications. *Proc. Fall Joint Comput. Conf., AFIPS Conf.*, page 37, 1970.

[2] Norman Abramson. Development of the ALOHANET. *IEEE Transactions on Information Theory*, 31:119–123, March 1985.

[3] Y. Lee B.D. Choi, D.I. Choi and D.K. Sung. Priority queueing system with fixed-length packet-train arrivals. *IEE Proceedings-Communications*, 145(5):331–336, October 1998.

[4] Dimitri Bertsekas and Robert Gallager. *Data Networks.* Prentice-Hall, Inc., 1987.

[5] Uyless Black. *Second Generation Mobile & Wireless Networks.* Prentice Hall PTR, 1999.

[6] John I. Capetanakis. Tree algorithms for packet broadcast channels. *IEEE Transactions on Information Theory*, IT-25(5):505–514, September 1979.

[7] Ki-Ho Cho and Hyunsoon Yoon. Design and analysis of a fair scheduling algorithm for qos guarantees in high-speed packet-switched networks. *1998 IEEE International Conference on Communications*, 3:1520–1525, 1998.

[8] J.S. Choi and C.K. Un. Delay performance of an input queueing packet wwitch with two priority classes. *IEE Proceedings-Communications*, 145(3):141–144, June 1998.

[9] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory.* John Wiley & Sons, Inc., 1991.

[10] Satish Damodaran and Krishna M. Sivalingam. Scheduling in wireless networks with multiple transmission channels. *1999 International Conference on Network Protocols*, pages 262–269, 1999.

[11] Juan Duan and Suresh Singh. Efficient utilization of multiple channels between two switches in atm networks. *1995 IEEE International Conference on Communications*, 3:1906–1911, 1995.

[12] Erol Gelenbe and Guy Pujolle. *Introduction to Queueing Networks.* John Wiley & Sons, Inc., 1998.

[13] Tri T. Ha. *Digital Satellite Communications.* McGraw-Hill, 1990.

[14] Gary N. Higginbottom. *Performance Evaluation of Communication Networks.* Artech House, Inc., 1998.

[15] Edward P.C. Kao and Sandra D. Wilson. Analysis of nonpreemptive priority queues with multiple servers and two priority classes. *European Journal of Operational Research*, 118(1):181–193, October 1999.

[16] Christopher Karpinsky. Uplink multiple access techniques for satellite communication systems. Master's thesis, Massachusetts Institute of Technology, 1998.

[17] Farooq Khan and Djamal Zeghlache. Multilevel channel assignment (mca): A performance analysis. In Jack M. Holtzman and Michele Zorzi, editors, *Advances in Wireless Communications*, pages 301–311. Kluwer Academic Publishers, 1998.

[18] Leonard Kleinrock. *Communication Nets: Stochastic Message Flow and Delay.* McGraw-Hill Book Company, 1964.

[19] Leonard Kleinrock. *Queueing Systems, Volume I: Theory.* John Wiley & Sons, Inc., 1976.

[20] Leonard Kleinrock. *Queueing Systems, Volume II: Computer Applications.* John Wiley & Sons, Inc., 1976.

[21] Leonard Kleinrock. On queueing problems in random-access communications. *IEEE Transactions on Information Theory*, IT-31(2):166–175, March 1985.

[22] V.G. Kulkarni. *Modeling, Analysis, Design, and Control of Stochastic Systems.* Springer-Verlag New York, Inc., 1999.

[23] F.L. Lo, T.S. Ng, and T.I. Yuk. Performance of multichannel CSMA networks. *International Conference on Information, Communications and Signal Processing ICICS '97*, September 9-12 1997.

[24] Marco Ajmone Marsan and Fabio Neri. A simulation study of delay in multichannel CSMA/CD protocols. *IEEE Transactions on Communcications*, 39(11):1590–1603, November 1991.

[25] Marco Ajmone Marsan and Daniele Roffinella. Multichannel local area network protocols. *IEEE Journal on Selected Areas in Communications*, SAC-1(5):885–897, November 1983.

[26] John J. Metzner. *Reliable Data Communications.* Academic Press, 1998.

[27] E. Modiano. Scheduling algorithms for message transmission over the bgs satellite system. Technical Report 1035, MIT Lincoln Laboratory, MIT Lincoln Laboratory, Lexington, MA, August 11 1997.

[28] E. Modiano. Scheduling packet transmissions in a multi-hop packet switched network based on message length. Technical Report 1036, MIT Lincoln Laboratory, MIT Lincoln Laboratory, Lexington, MA, June 30 1997.

[29] Jeannine Mosely and Pierre A. Humblet. A class of efficient contention resolution algorithms for multiple access channels. *IEEE Transmactions on Communications*, COM-33(2):145–150, February 1985.

[30] Andrew Muir and J.J. Garcia-Luna-Aceves. A channel access protocol for multihop wireless networks with multiple channels. *1998 IEEE International Conference on Communications*, 3:1617–1621, 1998.

[31] Tero Ojanpera and Ramjee Prasad. *Wideband CDMA for Third Generation Mobile Communications.* Artech House, 1998.

[32] E.R. Peterson. A dynamic programming model for assigning customers to priority service classes. *INFOR*, 36(4):238–246, November 1998.

[33] Ioannis E. Pountourakis and Efstathios D. Sykas. Multichannel multiple queue protocols. *European Transactsions on Telecommunications*, 7(6):553–563, November-December 1996.

[34] Ramjee Prasad. *Universal Wireless Personal Communications.* Artech House Publishers, 1998.

[35] D. Raychaudhuri. Performance analysis of random access packet-switched code division multiple access systems. *IEEE Transactions on Communcications*, 29:895–901, June 1981.

[36] Man Young Rhee. *CDMA Cellular Mobile Communications and Network Security.* Prentice Hall PTR, 1998.

[37] Raphael Rom and Moshe Sidi. *Multiple Access Protocols: Performance and Analysis.* Springer-Verlag, 1990.

[38] P.D. Roorda and V.C.M.Leung. Dynamic control of time slot assignment in multiaccess reservation protocols. *IEE PRoc.-Commun.*, 143(3):167–175, June 1996.

[39] Sheldon M. Ross. *Introduction to Probability Models.* Academic Press, 1989.

[40] Tarek N. Saadawi and Anthony Ephremides. Analysis, stability, and optimization of slotted ALOHA with a finite number of buffered users. *IEEE Transactions on Automatic Control*, AC-26(3):681–688, June 1981.

[41] Seiichi Sampei. *Applications of Digital Wireless Technologies to Global Wireless Communications.* Prentice Hall PTR, 1997.

[42] J. Scott Stadler. Protocol enhancement and packetized uplink multiple access for satellite networks. MIT Lincoln Laboratory, September 1998.

[43] William Webb. *The Complete Wireless Communications Professional: A Guide for Engineers and Managers.* Artech House Publishers, 1999.

[44] Gang Wu, Kaiji Mukumoto, and Akira Fukuda. An integrated voice and data transmission system with idle signal multiple access - dynamic analysis. *IEICE Trans. Commun.*, E76-B(11):1398–1407, November 1993.

[45] Jung-Shyr Wu, Jen-Kung Chung, and Kung-Hwa Lee. Performance evaulation for multichannel access schemes in CDMA cellular systems. *International Journal of Communication Systems*, 11:129–135, 1998.