INFORMATION TO USERS

This reproduction was made from a copy of a document sent to us for microfilming. While the most advanced technology has been used to photograph and reproduce this document, the quality of the reproduction is heavily dependent upon the quality of the material submitted.

The following explanation of techniques is provided to help clarify markings or notations which may appear on this reproduction.

1. The sign or “target” for pages apparently lacking from the document photographed is “Missing Page(s)”. If it was possible to obtain the missing page(s) or section, they are spliced into the film along with adjacent pages. This may have necessitated cutting through an image and duplicating adjacent pages to assure complete continuity.

2. When an image on the film is obliterated with a round black mark, it is an indication of either blurred copy because of movement during exposure, duplicate copy, or copyrighted materials that should not have been filmed. For blurred pages, a good image of the page can be found in the adjacent frame. If copyrighted materials were deleted, a target note will appear listing the pages in the adjacent frame.

3. When a map, drawing or chart, etc., is part of the material being photographed, a definite method of “sectioning” the material has been followed. It is customary to begin filming at the upper left hand corner of a large sheet and to continue from left to right in equal sections with small overlaps. If necessary, sectioning is continued again—beginning below the first row and continuing on until complete.

4. For illustrations that cannot be satisfactorily reproduced by xerographic means, photographic prints can be purchased at additional cost and inserted into your xerographic copy. These prints are available upon request from the Dissertations Customer Services Department.

5. Some pages in any document may have indistinct print. In all cases the best available copy has been filmed.
PLEASE NOTE:

In all cases this material has been filmed in the best possible way from the available copy. Problems encountered with this document have been identified here with a check mark √.

1. Glossy photographs or pages _____
2. Colored illustrations, paper or print _____
3. Photographs with dark background _____
4. Illustrations are poor copy _____
5. Pages with black marks, not original copy _____
6. Print shows through as there is text on both sides of page _____
7. Indistinct, broken or small print on several pages √
8. Print exceeds margin requirements _____
9. Tightly bound copy with print lost in spine _____
10. Computer printout pages with indistinct print _____
11. Page(s) __________ lacking when material received, and not available from school or author.
12. Page(s) __________ seem to be missing in numbering only as text follows.
13. Two pages numbered ________. Text follows.
14. Curling and wrinkled pages _____
15. Other ____________________________________________________________

University
Microfilms
International
RICE UNIVERSITY

Performance Analysis of an
Integrated Data Voice Link

by

Balakrishna R. Iyer

A THESIS SUBMITTED
IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE OF

Doctor of Philosophy

APPROVED, THESIS COMMITTEE :

J. Bartlett Sinclair
Assistant Professor of Computer Science in
the Department of Electrical Engineering
Chairman

Don H. Johnson
Associate Professor in the Department of
Electrical Engineering

Paul E. Pfeiffer
Professor of Mathematical Science

Houston, Texas

October, 1982
ABSTRACT

"Performance Analysis of an Integrated Data-Voice Channel"

by Balakrishna R. Iyer

We study the performance of a communication channel carrying data and voice traffic. The capacity of the channel is assumed to be insufficient for both types of traffic during periods of peak demand. Data is therefore buffered.

We model such a communications channel. Data is assumed to arrive at a constant fixed rate. Voice is generated by multiple sources that turn on and off randomly. The probability that voice occupies the maximum allowable capacity is the measure of voice performance and the distribution of data buffer lengths and data waiting times measures data performance. We obtain closed form expressions for the performance parameters. The 95th percentile of the probability distribution of buffer length and waiting time is used to produce design curves for an integrated data voice channel.

We investigate the effect of various system parameters on the performance for data and voice traffic.
ACKNOWLEDGEMENTS

I sincerely thank Bart Sinclair, my thesis advisor, for his guidance, interest, encouragement and inspiration during my stay at Rice University. I am grateful to Don Johnson and Paul Pfeiffer for serving on my thesis committee.

I express my gratitude to Robert Bartoscynski, John Dennis, Dan Friedman, Gene Golub, Eric Gross, Frank Jones, Linda Kaufmann, Debasis Mitra, Michael Pearlman, Michael Taksar and Fouad Tobagi for sharing their ideas with me.

I thank my friends Sergio Cabrera, Rick Covington, Stuart De Graaf, Deepak Dhandhere, Nirmala Ganapathy, C. K. Krishnan, Manoj Kumar, Kannan Moudgalya, Steve Preston, Ben Reaves, Mark Riggle, David Scheibner, Kent Spielman, Dilip Venkatachari, Neng Hui Wang, Owen Wilson and Michael Wheeler for making my graduate study enjoyable.

I dedicate this thesis to my parents who have made many untold sacrifices to give me a good education.

This research is supported by the National Science Foundation under grant MCS 80-04107.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chapter 1</td>
<td>Introduction</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>Terminology</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>Motivation</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>Scope</td>
<td>5</td>
</tr>
<tr>
<td>Chapter 2</td>
<td>Survey</td>
<td>6</td>
</tr>
<tr>
<td>1</td>
<td>Outline</td>
<td>6</td>
</tr>
<tr>
<td>2</td>
<td>Single Conversation Models</td>
<td>7</td>
</tr>
<tr>
<td>3</td>
<td>Multiple Conversation Models : Approximate Analysis</td>
<td>11</td>
</tr>
<tr>
<td>4</td>
<td>Multiple Conversation Models : Exact Analysis</td>
<td>20</td>
</tr>
<tr>
<td>5</td>
<td>Summary</td>
<td>25</td>
</tr>
<tr>
<td>Chapter 3</td>
<td>Modeling</td>
<td>35</td>
</tr>
<tr>
<td>1</td>
<td>Introduction</td>
<td>35</td>
</tr>
<tr>
<td>2</td>
<td>Voice Traffic</td>
<td>35</td>
</tr>
<tr>
<td>3</td>
<td>Data Traffic</td>
<td>38</td>
</tr>
<tr>
<td>4</td>
<td>The Link</td>
<td>38</td>
</tr>
<tr>
<td>Chapter 4</td>
<td>Performance Analysis for Voice Traffic</td>
<td>41</td>
</tr>
<tr>
<td>1</td>
<td>Introduction</td>
<td>41</td>
</tr>
<tr>
<td>2</td>
<td>Blocking Probability and Throughput</td>
<td>41</td>
</tr>
<tr>
<td>3</td>
<td>Scaling Effects</td>
<td>43</td>
</tr>
<tr>
<td>Chapter 5</td>
<td>Analysis for Data : Buffer Length</td>
<td>52</td>
</tr>
<tr>
<td>Chapter</td>
<td>Title</td>
<td>Page</td>
</tr>
<tr>
<td>---------</td>
<td>-----------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>1</td>
<td>Introduction</td>
<td>52</td>
</tr>
<tr>
<td>2</td>
<td>General Problem</td>
<td>53</td>
</tr>
<tr>
<td>3</td>
<td>System Dynamics</td>
<td>55</td>
</tr>
<tr>
<td>4</td>
<td>Algebraic Properties</td>
<td>57</td>
</tr>
<tr>
<td>5</td>
<td>Buffer Length Distribution</td>
<td>59</td>
</tr>
<tr>
<td>6</td>
<td>On Solving the Eigenvalue Problem</td>
<td>64</td>
</tr>
<tr>
<td>7</td>
<td>Moments of the Buffer Length Distribution</td>
<td>66</td>
</tr>
<tr>
<td>Chapter 6</td>
<td>Analysis for Data: Delay</td>
<td>69</td>
</tr>
<tr>
<td>1</td>
<td>Introduction</td>
<td>69</td>
</tr>
<tr>
<td>2</td>
<td>Theory</td>
<td>70</td>
</tr>
<tr>
<td>3</td>
<td>Application to Data Delay</td>
<td>79</td>
</tr>
<tr>
<td>4</td>
<td>Moments of the Data Delay</td>
<td>79</td>
</tr>
<tr>
<td>Chapter 7</td>
<td>Performance for Data and Conclusions</td>
<td>81</td>
</tr>
<tr>
<td>1</td>
<td>Performance Analysis</td>
<td>81</td>
</tr>
<tr>
<td>2</td>
<td>Conclusions</td>
<td>88</td>
</tr>
<tr>
<td>Chapter 8</td>
<td>Extensions and Open Problems</td>
<td>109</td>
</tr>
<tr>
<td>1</td>
<td>Introduction</td>
<td>109</td>
</tr>
<tr>
<td>2</td>
<td>Data Model Refinements</td>
<td>109</td>
</tr>
<tr>
<td>3</td>
<td>Voice Model Refinement</td>
<td>115</td>
</tr>
<tr>
<td>4</td>
<td>Link Model Refinements</td>
<td>116</td>
</tr>
</tbody>
</table>
CHAPTER 1

Introduction

1. Terminology

   In the past the predominant communications traffic between geographically separated points was voice. Due to today’s trend towards distributed computing, information sharing and video entertainment, we see the emergence of new traffic types like electronic mail, video graphics, file transfers, etc. In this thesis we divide these traffic types into two categories - voice and data. We study the performance of a point to point communications link that carries both types at the same time.

   Voice traffic is primarily generated by people talking to each other. A small portion of voice traffic is generated by people talking to machines and machines speaking to people. There is a delay between the generation of the voice signal by the speaker and its receipt by the hearer. The delay is due to the time taken by the signal to pass through several electrical components and the communications links. Since normal conversations are highly interactive the delay has to be kept small. It is usual to transmit voice traffic with the highest priority in the presence of other traffic types and subject it to the minimum buffering. There are several algorithms that reconstruct speech voice signals at the receiving end. Many of these algorithms do a good job of reconstruction in spite of loss of
small segments of the voice conversation. Also, people accept the fact that when they dial a telephone number all lines may be busy and they will have to dial again. Due to the statistical variation in the behavior of customers the probability of all the telephones subscribers wanting to talk at the same time is vanishingly small, and it is common practice to service a population with bandwidth less than what would be required for transmission of voice if all members of the population were talking simultaneously. Voice traffic arriving in excess of the maximum rate that can be supported by the link is lost. The probability that voice traffic is lost is called the blocking probability.

Data traffic is typically generated among computers and terminals. Unlike voice, data can tolerate delays but not loss of portions of messages. The link transmission capacity need not exceed the data arrival rate for all time, assuming sufficient buffer capacity, since by buffering we can transmit data, without any loss. We only require that the average link capacity for data exceed the average data arrival rate. Data arrive into a buffer and stay there until transmitted over the communications link. The amount of data in the buffer is defined as the buffer length. The time spent in the buffer by data is called the data delay.

Data and voice traffic may be transmitted on two separate links or on the same link. The advantage of the latter lies in the ability to increase the utilization of the link through the sharing of bandwidth. The statistical fluctuations in traffic make sharing
feasible. An integrated data-voice link is a communications link that carries both data and voice traffic. Bandwidth is dynamically shared by voice and data. Voice traffic has pre-emptive priority over data traffic. The link is a point to point transmission facility and the overhead of link access and multiplexing is assumed to be negligible.

2. Motivation

Deregulation of the long distance communications industry, the proliferation of cheap computer processors, advances in office automation and home information services and the trend towards distributed computing has increased the amount of data transmission. Voice (i.e., telephone) networks are extensive and one way to handle the increased data traffic is to use the available voice networks, perhaps with minor augmentation. Fears have been expressed in the literature [1] that one communications system carrying both voice and data is less cost-effective than two separate communications systems for voice and data. Decisions regarding the integration of data and voice have also to be made by entrepreneurs who are starting to offer communications services to compete with existing telephone networks. A cost/benefit [2] analysis of an integrated communications system requires the performance analysis of an integrated data voice link.

We have chosen the blocking probability as the performance parameter for voice traffic. When we discuss modeling issues in Chapter 3 we will show that blocking probability can be interpreted in two ways. It may denote the probability that a telephone customer
finds all lines on the link busy and cannot initiate a call. Alternatively, the blocking probability may denote the probability that a small segment of a conversation is lost. Whenever a segment of speech is lost, the speech reconstruction algorithms generate distortions over the phone. In either case distortions or busy lines cause customer dissatisfaction. Low blocking probabilities are critical to a voice network.

We measure data performance by the data delay. We want to know the average, standard deviation and 95th percentile of the delay. The 95th percentile is a particularly important design parameter since the engineering design of communication systems is done on a worst-case basis. The three performance parameters for data are obtained from the probability distribution function for the data delay. During communication between a terminal and a computer, data delay is a component of response time. During processing on a distributed system the delay is a measure of inter-processor communication cost. For mail and facsimile, data delays contribute to the transfer time of these messages from source to destination.

The cost of a communications facility is a function of the amount of data buffer required and we use the data buffer length as another performance parameter for the link. Again, we are interested in the average, standard deviation and 95 percentile of the buffer length distribution.
3. **Scope**

In this thesis we study the effect of data voice integration on data and voice performance. Since voice has priority over data, it is not affected by the presence of data. Data performance degrades with the presence of voice and we study this degradation. The cost of the communication system is affected by the amount of data buffer required. We develop models for the communications link and both data and voice traffic. The principal result is the derivation of explicit expressions for the blocking probability and the probability distributions for data buffer length and data delay. In computing delay we work under the assumption that the data buffer operates on a first-in first-out basis. From the distribution functions we obtain the average, standard deviation and 95th percentile. We study these performance measures for the integrated link as a function of link size and amounts of voice and data traffic. As a practical example, we examine the effect of integration on the widely used T1 carrier communications link.

For links with large capacities and random communications traffic the probability of the instantaneous traffic deviating from the average becomes small. In these cases there is little sharing of bandwidth between voice and data and little buffering and delay. For a range of intermediate link sizes, integration may involve considerable sharing and it is within this range that we focus our attention. We will also examine the effect on the various performance parameters due to scaling the system size.
CHAPTER 2

Survey

1. Outline

Our modeling and analysis work in this thesis has been preceded by other models and analyses of an integrated data voice link. We survey some of this previous work in this chapter. In one of the earliest efforts [3], Sherman modeled a link carrying a single conversation into which data is multiplexed. Sherman observed that the data buffer builds up and depletes alternately, taking on peak values in between. He obtained the asymptotic distribution for these peaks. This analysis is applicable to data voice links that support only one voice conversation.

Our focus is on links supporting multiple simultaneous conversations. Coviello and Vena [4] modeled such a link that uses a Time Division Multiple Access (TDMA) strategy to multiplex data among many voice conversations. Their model is also applicable to other data voice multiplexing strategies like Frequency Division Multiple Accessing (FDMA).

Occhiogrosso et. al. [5], Leon-Garcia et. al. [6] and Fischer and Harris [7] proposed similar models and carried out approximate analyses. Varshney [8] applied the approximations made by Occhiogrosso et. al., Leon-Garcia et. al. and Fischer and Harris to a model that explicitly considers the discretization of time caused
by the use of the TDMA scheme. Weinstein et. al. [9, 10] ran simulations of the TDMA scheme and estimated various performance parameters. These results exposed some of the shortcomings of the approximate analyses. Friedman [11] worked on the exact analysis of a model of the integrated data-voice link. His work is complemented by the exact analysis of a similar model by Shantikumar et. al. [12]. There are some computational problems associated with Shantikumar's analysis. These and related issues were examined by Feldman and Claybaugh [13].

The following discussion is divided into three sections. First, we describe the analysis of the multiplexing of a solitary voice conversation and data traffic. Then we indicate the various approximate analyses and simulation results for models based on the TDMA scheme. The last section contains a survey of recent work on exact analysis.

2. Single Conversation Models

Sherman [3] considered a communications link that carries voice traffic generated by a single conversation. He modeled a link that exploited pauses in a conversation by comparing an estimate of the instantaneous speech signal power to a threshold and gating speech on and off accordingly. Gaps are thereby introduced in voice transmission. Sherman distinguished between short and long pauses that give rise to gaps. The duration of the gaps are modeled by an i.i.d.\footnote{i.i.d.: independent and identically distributed}
random variable sequence \( \{T_i\} \). The value of each random variable is determined by making a random and independent choice of one among two exponential distributions with means \( \frac{1}{\lambda_1} \) and \( \frac{1}{\lambda_2} \) respectively. The first distribution is picked with probability \( A \), the probability of occurrence of a short pause (generally 0.95) and the second distribution with probability \( 1-A \), the probability of the occurrence of a long pause. The probability density of the random variable \( T_i \) is a hyperexponential

\[
p(t) = A \lambda_1 e^{-\lambda_1 t} + (1-A) \lambda_2 e^{-\lambda_2 t}.
\]

The speech signal between two gaps is usually referred to as a speech burst. Sherman assumed speech bursts to be i.i.d. exponentially with mean \( \frac{1}{\mu} \). Data arrives into a buffer at a fixed rate \( C_D \). The transmission rate that can be devoted to data is \( d(0) \) when no speech is being transmitted on the link and \( d(1) \) during transmission of a speech burst. We examine the interesting situation in which \( d(0) > C_D > d(1) \).

Sherman observed that data in the buffer builds up during speech bursts, peaks at the moment a speech gap starts and falls linearly until the occurrence of the next speech burst. The peak buffer length at the conclusion of the \( i \)’th speech burst is denoted as \( X_i \). \( \{X_i\} \) is a Markov process.

The analysis in [3] shows that the asymptotic density of \( \{X_i\} \) is exponential,
\[ p(x) = \lambda e^{-\lambda x}, \text{ where} \]

\[
\lambda = \frac{1}{2} \left\{ - \frac{\lambda_1 + \lambda_2}{d(0) - C_D} + \frac{\mu}{C_D - d(1)} + \right. \\
\left. \left[ \frac{\lambda_1 + \lambda_2}{d(0) - C_D} - \frac{\mu}{C_D - d(1)} \right]^2 + 4 \left[ \frac{\mu}{C_D - d(0)} \left[ \frac{(1-A)\lambda_1 + A\lambda_2}{d(0) - C_D} \right] - \frac{\lambda_1 \lambda_2}{(d(0) - C_D)^2} \right]^{\frac{1}{2}} \right\}.
\]

The requirement \( \lambda > 0 \) is the stability condition. From the expression for \( \lambda \) we can rewrite the stability condition as

\[
C_D < d(1) \frac{\lambda}{\lambda + \mu} + d(0) \frac{\mu}{\lambda + \mu}, \text{ where} \quad \frac{1}{\lambda} = A + \frac{1-A}{\lambda_1 + \lambda_2}.
\]

In other words, we require that the data arrival rate be less than the average transmission capacity left on the link after the transmission of voice.

The probability that the length of the buffer peaks exceed a certain length \( X_0 \) is \( e^{-\lambda X_0} \). Sherman plotted the average buffer length as a function of the data arrival rate \( C_D \), keeping the total link transmission capacity constant. He noticed the increase in buffer length with the increase in \( C_D \). There is a knee for this curve. A good integrated data-voice link (of the type modeled by Sherman) should be designed with parameters that put it in the region below the knee.

Let us set \( A = 1 \) in the above voice traffic model (this frees up the use of the variable \( \lambda_2 \)) and model data traffic as messages arriv-
ing according to a Poisson process of intensity $\lambda_2$ (we use $\lambda_1$ and $\mu_1$ for voice). Assume message lengths to be distributed exponentially with mean $\frac{1}{\mu_2}$. The transmission rate available for data is $d(0)$ when there is no speech transmission and $d(1)$ during speech transmission.

Denoting $X_t$ as the number of messages in the buffer at time $t$ and $I_t$ as the binary valued random variable indicating that speech is being transmitted ($I_t = 1$) or not ($I_t = 0$), we have a Markov process $(X_t, I_t)$ described by the chain in Figure 2.1.

Yechiali and Naor [14] obtained the steady state probability

$$p(x,i) = \lim_{t \to \infty} \Pr(X_t = x, I_t = i).$$

They proved that the cubic equation

$$g(z) = \frac{\lambda_2^2}{2} z^3 - (\lambda_2 (\lambda_1 + \mu_1 + d(1)\mu_2 + d(0)\mu_2 + \lambda_2)) \lambda_2 z^2 + \lambda_2 \mu_2 d(1) + \mu_2^2 d(0) d(1) + \lambda_2 \mu_2 d(0)$$

$$= d(0) d(1) \lambda_2^2 \mu_2^2,$$ has exactly one root in the open interval $(0,1)$.

Let $z_0$ be this root.

Now, let $\hat{\lambda} = \lambda_2$ and $\hat{\mu} = \frac{\mu_1 \mu_2 d(0) + \lambda_1 \mu_2 d(1)}{\lambda_1 + \mu_1}.$

Then

$$p(0,0) = \frac{\mu_1 (\hat{\mu} - \hat{\lambda}) z_0}{d(0) \mu_2 (1 - z_0)(d(1) \mu_2 - \lambda_2 z_0)}.$$
\[ p(0,1) = \frac{\lambda_1(\hat{\mu} - \hat{\lambda})z_0}{d(1)\mu_2(1 - z_0)(d(0)\mu_2 - \lambda_2z_0)} \]

\[ p(m,0) = p(m-1,0)\frac{\lambda_2}{d(0)\mu_2} + \left[ \sum_{j=0}^{m-1} p(j,0) \right] \frac{\lambda_1}{d(0)\mu_2} - \left[ \sum_{j=0}^{m-1} p(j,1) \right] \frac{\mu_1}{d(0)\mu_2} \]

\[ p(m,1) = p(m-1,1)\frac{\lambda_2}{d(1)\mu_2} + \left[ \sum_{j=0}^{m-1} p(j,1) \right] \frac{\mu_1}{d(1)\mu_2} - \left[ \sum_{j=0}^{m-1} p(j,0) \right] \frac{\lambda_1}{d(1)\mu_2} \]

Yechiali and Naor obtained asymptotic moments for \( X_t \) from these expressions. Using Little's result, we can also find the average data delay.

Yechiali and Naor's work can also be viewed as a specific case of Friedman's [11] analysis which we shall describe in a later section.

3. **Multiple Conversation Models: Approximate Analysis**

Coviello and Vena [4] describe a technique for multiplexing voice and data traffic on a communications link using a TDMA scheme. Time is divided into frames of fixed duration as in Figure 2.2. Each frame is further divided into slots. One slot is reserved in each successive frame for an on-going conversation. Multiple simultaneous conversations require the reservation of multiple slots in successive frames. Voice traffic will be transmitted only during these slots. The number of slots reserved for voice per frame varies dynamically with the number of on-going conversations. The slots that are not reserved for voice may be used for data transmission. Coviello and Vena coined the name Slotted Envelope Network (SENEN) to describe a
communications network based on this scheme. Usually a fixed number of slots in every frame will be dedicated exclusively to data traffic, thus assuring a minimum grade of service for data transmission. The rest of this section is a survey of the literature on the modeling and the approximate analysis of communications links based on the SENET concept.

Occhiogrosso et. al. [5] considered a SENET implementation in which no more than N voice conversations are allowed to take place simultaneously. All slots are of equal duration. At most N slots per frame can be reserved for voice conversations. The blocking probability is the probability that all these N slots are taken up by voice. In this SENET implementation an incoming voice call request, which arrives at an arbitrary instant in the current frame, is buffered until the start of the next frame while the communications facility ascertains the availability of a slot and reserves a free slot in the next frame. If N slots have already been reserved by voice the incoming request is "blocked" and lost. Under SENET, disconnection of a call is only effective at the end of a frame. When we increase frame duration we increase the time period over which slots are reserved for a conversation. This results in a higher blocking probability for all calls. Besides leading to high blocking probability, longer frames require greater storage space for buffering information until the start of the next frame. In long distance communication, voice traveling from one link to another would suffer an average delay of half a frame duration for each
crossover between links. If voice has to travel over several links this delay may constrain us to use an appreciably shorter frame. However, the frame duration cannot be reduced arbitrarily because of the overhead of management incurred at the end of each frame which results in inefficient link utilization due to the transmission of control bits. A trade-off has to be made in choosing frame length. Occhiogrosso et al. demonstrated that if the mean number of call requests and call disconnections per frame is much less than one, then the blocking probability may be well approximated by using the following model, that ignores the frame structure.

Voice calls arrive according to a Poisson process with intensity $\lambda$. A voice call requires transmission for an exponentially distributed amount of time with average $\frac{1}{\mu}$. The maximum number of simultaneous voice calls is $N$. The Markov chain in Figure 2.3 describes the voice process $I_t$, where $I_t$ is the number of calls at time $t$.

The blocking probability is the asymptotic probability of the process being in the state $N$ and is given by the Erlang B formula, 

$$\frac{\frac{\lambda}{\mu}^N}{\sum_{i=0}^{N} \frac{\lambda}{\mu}^i \frac{\lambda}{\mu}}.$$ 

Through the solution of the steady state probabilities for each state in the Markov chain we find the average number of slots occupied by voice. Let this number be $\bar{I}$ and let $C$ be the total number of slots in a frame. Occhiogrosso et al. assumed that every slot is available for data transmission with probability $\frac{C-I}{C}$, and that data messages are of a size that can be transmitted exactly in a
slot. They also assumed that data messages that arrive during a frame can be transmitted at the earliest during the next frame and derived an expression for the average data queue length and the average data delay as a function of the data arrival process.

Fischer and Harris [7] carried out an analysis for data delay assuming the number of voice calls at the beginning of any two different frames are independent random variables. In reality the two variables show a Markov dependence. Weinstein [9] pointed out, with the aid of simulation results, that Fischer and Harris' assumption led to a gross underestimate of data delay. This dependence was also ignored by Occhiogrosso et al. [5] and we suspect that their work also underestimates data delay appreciably.

Chang [15] and Weinstein [9, 10] separately proposed a two dimensional Markov chain model for the integrated data-voice link. They assumed data message lengths to be exponentially distributed and ignored the effects of the frame structure. Again voice calls arrive according to a Poisson process of intensity \( \lambda_1 \) requiring transmission for an exponentially distributed duration. The link is modeled as S+N servers where S of them are devoted to data exclusively. A data message can be serviced by at most one server. This corresponds to the situation in which a specific data message can take at most one slot per frame. Voice calls have pre-emptive priority over the remaining N servers. Each server is regarded as a voice sub-channel capable of handling a single conversation. Voice calls that arrive when all N sub-channels are occupied are lost. The loss probability
is given by the Erlang B formula.

Let us define $X_t$ as the number of data messages and $I_t$ as the number of voice calls in progress at time $t$. The Markov chain in Figure 2.4 describes the process $(X_t, I_t)$. We can write a probability balance equations for each state and solve for the steady state probability using a generating function technique. This requires finding the $S-1$ unique roots of a polynomial, which is a computationally difficult task. Chang worked out specific examples of this problem and suggested some methods to obtain an approximate solution.

Leon-Garcia et. al. [6] used a series of approximations to estimate the average data delay. Like other modelers they ignored the frame structure of SENET and modeled the voice process as a continuous time discrete state Markov process. Each transmitting voice source generates voice traffic at the same rate. Data traffic is assumed to arrive as packets with a mean arrival rate of $\lambda_2$ and mean length $\frac{1}{\mu_2}$. Let $X_t$ represent the number of data packets in the buffer at time $t$ and $I_t$ the number of voice conversations. Out of a total transmitting capacity of $B$, $r(i)$ is taken up to transmit voice and $d(i) = B - r(i)$ is available for the transmission of data packets (if any).

In the analysis the states taken by the process $I_t$ are classified into two sets $S_0$ and $S_1$,

$$S_0 = \{ k: 0 \leq k \leq N, \text{ where } \lambda_2 \geq d(N-k) \mu_2 \}$$
\( S_1 = \{ k: 0 \leq k \leq N, \text{ where } \lambda_2 < d(N-k) \mu_2 \} \).

\( S_0 \) is the set of states of the voice process for which the instantaneous data packet arrival rate is greater than or equal to the rate at which these packets may be transmitted on the link. \( S_1 \) contains the remaining states.

The asymptotic expected number of packets in the queue is

\[
\lim_{t \to \infty} E(X_t) = \lim_{t \to \infty} \left( \sum_{k=0}^{N} E(X_t | I_t = k) P(I_t = k) \right)
\]

\[
+ E(X_t | I_t \in S_0) P(I_t \in S_0) \right). \quad (2.1)
\]

The quantities \( \lim_{t \to \infty} P(I_t = k) \) and \( \lim_{t \to \infty} P(I_t \in S_0) \) are found by solving for the state probabilities of the voice Markov process. Leon-Garcia et. al. used estimates for the other terms.

Their estimate for \( \lim_{t \to \infty} E[X_t | I_t = k] \) is the expected number of data packets in queue for an \( M/M/1 \) queuing system that has an exponentially distributed service rate with mean \( \frac{1}{d(N-k)\mu_2} \) and Poisson arrivals with intensity \( \lambda_2 \).

In estimating \( \lim_{t \to \infty} E[X_t | I_t \in S_0] \) they assumed that \( r(i) \) is non-decreasing in \( i \). Hence, for some \( C, k > C \text{ iff } k \in S_0 \). They observed migrations of the process \( I_t \) about the value \( C \) as indicated in Figure 2.5a. \( I_t \) crosses the threshold \( C \) at times \( t_i \). Leon-Garcia et. al. approximated data as arriving continuously at a rate of \( \frac{\lambda_2}{\mu_2} \) bits per second and used the number of bits in the buffer to estimate
the number of data packets. The number of bits (regarded here as a continuous variable) varies typically as in Figure 2.5b. The rate of change in buffer length is random. In evaluating the estimate they replaced this variable growth and decay rate of the buffer by their respective constant average values. Then the variation of buffer length \( X_t \) looks typically like the plot in Figure 2.5c, showing peaks and valleys.

Leon-Garcia et al. obtained the first and second moment of the distribution of the intervals \((t_1 - t_0)\), \((t_2 - t_1)\), etc. and subsequently the first moment of the buffer peaks and valleys. This statistic of the buffer peaks and valleys was used to estimate upper and lower bounds for \( \lim_{t \to \infty} E[ X_t | I_t \in S_0 ] \).

These estimates are used in (2.1) to obtain estimates for upper and lower bounds for the average data buffer length and thereby the average data delay.

Leon-Garcia et al. compared their estimates to simulation results and found that the estimate of the upper bound matched well with their simulation results. A point of concern is that the estimated upper bound does in fact take values less than the simulation results motivating the need for more accurate analysis for the integrated data-voice link.

Varshney [8] modeled the integrated data-voice link, taking into consideration the frame structure imposed by SE/NET. As usual, voice arrivals are Poisson with intensity \( \lambda_v \). Each voice transmission
lasts for an exponentially distributed length of time with mean \( \frac{1}{\mu_1} \).

The analysis of the voice process follows well known methods in discrete state Markov process theory.

Varshney modeled data traffic to consist of packets that arrive according to a Poisson process of intensity \( \lambda_2 \) and data packet lengths to be distributed arbitrarily.

The link divides time into frames of length \( T \), each sub-divided into \( N+S \) slots. \( S \) slots are exclusively used by data. Voice has priority over the remaining \( N \) slots. Let us define

\[ I_j = \text{the number of conversations at the beginning of the } j^{th} \text{ frame} \]

\[ X_j = \text{the number of data packets in the queue at the beginning of the } j^{th} \text{ frame} \]

\[ \eta_j = \text{the number of data packets that arrive during the } j^{th} \text{ frame}. \]

\[
P[X_{j+1} = x, I_{j+1} = i] = \sum_{k=0}^{\infty} \sum_{m=0}^{N} P[X_{j+1} = x, I_{j+1} = i | X_j = k, I_j = m] \cdot P[X_j = k, I_j = m].
\]

Varshney showed that

\[
P[X_{j+1} = x, I_{j+1} = i] = \sum_{k=0}^{m} \sum_{m=0}^{N} P[\eta_{j+1} = k - [S+N-m]] \cdot P[I_{j+1} = i | I_j = m] \cdot P[X_j = k, I_j = m].
\]
Using a generating function technique he evaluated the first moment of the number of data packets in the queue. This method requires the explicit evaluation of the \((N+1)^{(N+\frac{S}{2})-1}\) complex roots of an analytic function, which turns out to be a computationally difficult task.

Varshney proposed three approximations. In the first he ignored the Markov dependence of \(I_{j+1}\) on \(I_j\) and considered them to be i.i.d. This is the same assumption made by Fischer and Harris [7] and leads to gross underestimates of the data delay.

The second approximation replaces the Markov process \(I_j\) by its time average. This approximation was also made by Occhiogrosso et al. [5] and fails to explain the excessive data delays of the actual link.

In the third approximate method Varshney analyzed for data delay separately for each value taken on by \(I_j\), treating \(I_j\) to be constant at that value. He found the asymptotic data delay for each value taken on by \(I_j\) and used the weighted average of the delay (weighted by the asymptotic state probabilities of \(I_j\)) as an estimate for the overall data delay. Leon-Garcia et. al. [6] made the same assumption, which is good if the voice process is varying slowly with respect to the variation of the data buffer length. Unfortunately this method is not applicable for those values of \(i\) for which the average data arrival rate into the system is greater than the average residual transmission capacity available on the link after transmission of voice, because under these condition the data queue is
unstable.

Weinstein et. al.'s simulation [10, 9] of the SENET data/voice link showed that excessively large data queues (and delays) occurred during periods of high voice utilization and these are cleared during subsequent low utilization periods. The cause of this behavior is the large disparity between the average voice call duration and the data packet length \( \frac{1}{\mu_2} \).

We suspect that Weinstein et. al.'s simulation results are biased because they considered links for which the utilization was around 0.9, which is rather high for many practical links that operate under lighter loading conditions. This provides a motivation to investigate the data-voice integration idea further, especially for links which carry a large amount of traffic exhibiting sufficient statistical variation for data voice multiplexing.

Weinstein et. al. also investigated flow control techniques to keep data delays at acceptable levels. The investigation of such flow control techniques and protocols to be built over them to ensure proper data transmission is a fertile area of research that we do not consider in this thesis.

4. Multiple Conversation Models: Exact Analysis

Friedman [11] modeled voice traffic at the level of speech bursts. He considered a communications link with total transmission capacity \( B \), servicing \( N \) simultaneous speakers. \( I_t \) of these \( N \) speakers are generating voice signals at time \( t \). \( N - I_t \) are temporarily
silent. Friedman argued that we can use the usual assumptions about speech bursts and interspersing gaps having exponentially distributed lengths. The \( N+1 \) state continuous time discrete state Markov process, described by the Markov chain in Figure 2.6, is a good model for voice traffic according to the experimental studies of Weinstein [16]. The arrival rate of a new speech burst is directly proportional to the number of silent voice sources. The communications link reserves at most \( C_v \) of its capacity for voice. Voice is transmitted at time \( t \) at the rate \( r(I_t) \). \( r(i) \) is monotone increasing in \( i \). The fraction of speech bursts lost is

\[
\frac{\sum_{i=0}^{N} i \cdot \text{Prob}[I_t = i]}{\sum_{i=0}^{N} \text{Prob}[I_t = i]}.
\]

Friedman treated this fraction as a performance parameter for the voice process. His work contains a detailed analysis of the voice process.

Data traffic consists of packets independently and exponentially distributed in length with mean \( \frac{1}{\mu_2} \), arriving according to a Poisson process of intensity \( \lambda_2 \). The data packets are transmitted on a FIFO basis at the rate \( B - r(I_t) \). Let \( d(i) = \mu_2 (B - r(i)) \) and \( X_t \) be the number of data packets in the system at time \( t \). The vector Markov process \( (X_t, I_t) \) is described by the Markov chain in Figure 2.7.

It can be shown that the asymptotic probability of being in each state is non-zero if the average data arrival rate is less than the average transmission capacity left unused by voice. Let us define the asymptotic probability
\[ p(k,i) = \lim_{t \to \infty} \text{Prob}[X_t = k \text{ and } I_t = i], \]

and the probability vector

\[ \mathbf{p}(k) = (p(k,0), p(k,1), \ldots, p(k,i), \ldots, p(k,N)). \]

The Markov process described by Figure 2.7 is said to have the matrix geometric property [17] if after setting \( \mathbf{p}(k) = \mathbf{0} \) (the vector of all 0's), for \( k < 0 \), and for finite positive integers \( q \) and \( r \),

\[
\text{if } k > 0, \text{ then } \mathbf{q}^T = \sum_{j=-r}^{q} \mathbf{p}^T(k+j) \mathbf{A}^j(\mathbf{e}), \tag{2.2}
\]

for some square matrices \( \mathbf{A}^j(\mathbf{e}) \) and non-singular \( \mathbf{A}(-r+1) \). If the Markov chain has the matrix geometric structure we can write \( \mathbf{p}^T(k) \) in the form

\[
\mathbf{p}^T(k) = \mathbf{p}^T(0) \mathbf{q}^k, \text{ for all } k > 0 \tag{2.3}
\]

Substituting (2.3) in (2.2) we get

\[
\mathbf{q}^T = \sum_{j=-r}^{q} \mathbf{p}^T(0) \mathbf{q}^{j} \mathbf{A}^j(\mathbf{e}).
\]

This suggests the following recursive formula for the computation of \( \mathbf{q} \)

\[
\mathbf{q} = \mathbf{A}^{-1}(-r+1)[-\mathbf{A}(-r) - \sum_{i=2}^{q+r} \mathbf{A}(-r+i) \mathbf{e}].
\]

A formal treatment of this matter is given by Neuts [17].

A concise representation of the balance equation of the process \((X_t, I_t)\) is given in (2.4).
For all $k > 0$, $\Omega^T = \pi^T(k-1)\lambda^- + \pi^T(k)\omega^- + \pi^T(k+1)\omega^+$, \hspace{1cm} (2.4)

where $\lambda^- = \lambda_2 \mathbb{I}$, and the identity matrix is of size $N+1 \times N+1$, $\omega^-$ is another $N+1 \times N+1$ diagonal matrix with $(\omega^-)_{ii} = d(i)$. $\omega^+$ is a tri-diagonal matrix with $(\omega^+)_{i,i} = -(N-i)\lambda_1 - d(i) - \lambda_2$, $(\omega^+)_{i-1,i} = (N-(i-1))\lambda_1$, $(\omega^+)_{i,i+1} = (i+1)\mu_1$. The Markov process exhibits the matrix geometric property and the solution of the steady state probabilities can be written as in (2.3). Now, we obtain an iterative procedure from (2.4) for the calculation of $\Omega$,

$$\Omega = \Omega^{-1}(-\lambda^- - \omega^- \Omega^2).$$

Theoretical questions about the convergence of the iterative procedures remain unanswered. Friedman worked out an alternative procedure to calculate $\Omega$ which assures convergence.

Once we know $\Omega$ we can evaluate $\pi^T(0)$ in (2.3) by observing that

$$\sum_{k=0}^{\infty} \pi(k,i) = \text{Prob}[X_t = i] = \pi_i, \text{ for } 0 \leq i \leq N.$$  

This can be written as a vector equation

$$\sum_{k=0}^{\infty} \pi(k) = \pi^T.$$  \hspace{1cm} (2.5)

From (2.3) and (2.5) and under appropriate conditions for $\Omega$ we may write

$$\pi^T(0)[\mathbb{I} - \Omega]^{-1} = \pi^T,$$

from which we get

$$\pi^T(0) = \pi^T[\mathbb{I} - \Omega].$$
Since we know the state probabilities explicitly we can obtain all the performance parameters for the data buffer length distribution. Friedman computed the Laplace transform for the time the data messages have to spend in the communications facility. He obtained all moments of this distribution. Inversion of the transform is extremely hard and we are not able to obtain the distribution of the data message waiting time explicitly for this model of the integrated data-voice link. In particular this means that it is not easy to obtain the 95th percentile of the data delay.

Shantikumar et. al. [12] modeled the integrated data-voice link as C identical servers. Each voice call is serviced by a single server. Voice calls arrive according to a Poisson process with intensity $\lambda_1$. Voice calls require service for an exponentially distributed length of time with an average of $\frac{1}{\mu_1}$. Voice calls are lost if they arrive when $N$ servers are servicing voice or $N' \ (N' < N)$ servers are servicing voice and there are $L$ data messages in the system. Voice calls can pre-empt data messages of their servers. Data messages seize a server on a FIFO basis according to the above discipline. Data message inter-arrival times and lengths are i.i.d with exponential distributions with averages $\frac{1}{\lambda_2}$ and $\frac{1}{\mu_2}$ respectively. Let $d(i) = \mu_2 (N-i)$, $(X_t, I_t)$ is a vector Markov process described by the Markov chain in Figure 2.8. Shantikumar et. al. observed that if we set $k = \max(C-N',L)$ and partition the states $p(i,j)$ according to whether $i < k$ or $i \geq k$, the partition of states with $i \geq k$ has a matrix geometric structure. An extension of the matrix geometric
theory assures us that \( k^T(i) = p^T(k-1) \odot k^i \), for \( i \geq k \). We solve for \( p(i) \) for \( i \leq k \) by using the balance equation for each of the states \( p(i,j) \) with \( i \leq k \) and the normalization criterion. The steady state probabilities for each state in the Markov chain are now used to compute buffer length statistics. Because of the large number of balance equations to be considered in this method we run into numerical difficulties. Feldman and Claybaugh [13] investigated some of these numerical difficulties for specific values of \( N' \) and \( L \).

5. Summary

We briefly described the exact analysis of the data buffer length for the case of a single speaker (voice source) [3]. The approximate analysis for models that consider multiple simultaneous conversations [4, 6–8] ignore the time correlation of the number of conversations. Weinsteins simulations [10] exposed the inadequacy of this assumption because the approximate analyses fail to explain the inordinately long delays incurred by data. Friedman's work [11] which was done independently of ours is an exact analysis of a model that explicitly considers this correlation. The models for the link and voice traffic used by Friedman is essentially identical to our models that will be described in Chapter 3. While Friedman regards data traffic to consist of Poisson arrivals of packets that are exponentially distributed in length, we assume that data traffic arrives at a constant rate. Friedman's approach yields the distribution of the buffer length and the Laplace transform of the delay—which is numerically hard to invert. Using our approach we will
obtain both the data buffer length and delay distributions as a finite sum of exponentials.
Markov Chain for the Process 

\[(X_t, I_t)\]

for the Single Speaker Model

Figure 2.1
Δ, X, O: each represents reservations for one voice conversation

Total Transmission Capacity: 8.4 Mbps  Frame Width: 10 ms  Slot Width: 1 ms

An Example of SENET

Figure 2.2
$N$: Population Size

Markov Chain for the Voice Process

$I_t$

Figure 2.3
Maximum Queue Length = 0

Markov Chain for the Process \( (X_t, I_t) \)

constructed by Chang and Weinstein

Figure 2.4
Voice Process

Figure 2.5a

Buffer Length

Figure 2.5b

Approximate Buffer Length

Figure 2.5c

Voice and Buffer Length Behavior. Approximation due to Leon-Garcia et al.
$N: \text{Maximum number of simultaneous speakers}$

Markov Chain for the Voice Process

$I_t$

Figure 2.6
$X_t$: Number of Data Packets

$I_t$: Number of Voice Conversations

Markov Chain for the Vector Process $(X_t, I_t)$ as per Friedman's Model

Illustrating the Matrix-Geometric Structure

Figure 2.7
\[ X_t : \text{Number of Data Packets} \quad I_t : \text{Number of Conversations} \]

Markov Chain for the Process \((X_t, I_t)\) as per the model by Shantikumar et. al.

Figure 2.8
CHAPTER 3

Modeling

1. Introduction

For the data voice problem we distinguish three different aspects of modeling - the voice traffic model, the data traffic model and the link model. We use a continuous time discrete state stationary Markov process to model the generation of voice traffic. Data traffic is modeled as a deterministic process. The link is modeled as a server with fixed service rate whose service is randomly shared by data and voice.

2. Voice Traffic

Depending on the transmission strategy selected for voice, the link may transmit signals during the entire duration of a telephone conversation or only while a speaker is actually talking. The voice model to be described in this chapter can be used to model voice traffic generated by use of either strategy. This model has been widely used [11, 18]. Related models for voice traffic were discussed in Chapter 2.

Let us consider the voice traffic generation process. At any time there is a pool of a fixed number, say K, of independent voice sources, each in one of two states - active or silent. Normally each voice source independently and asynchronously alternates between the
active and silent states. The active and silent periods are of random durations with exponential distributions. The two distributions are typically different, but common to all sources; also, the sources are mutually independent. The above 'normal' behavior is subject to blocking in the following sense. A maximum of $N (N \leq K)$ voice sources may be active at any time. Thus, if $N$ sources are active further transitions to the active state are blocked for the remaining $(K-N)$ sources. This corresponds to blocking additional requests when the maximum number of connections are already in use. $I_t$ is defined as the number of active sources at time $t$. In Figure 3.1 we have drawn the Markov chain describing the Markov process $I_t$. Let $\pi_i = \lim_{t \to \infty} \Pr[I_t = i]$. 

In a conventional telephone system a connection is established between a subscriber and the party that he successfully dials. During the entire conversation bandlimited speech is transmitted at a selected bit rate. Specifically, this transmission continues even during silences that typically punctuate conversations. $K$ is set equal to the total number of subscribers serviced by a link. We set the average active period $\frac{1}{\mu}$ equal to the average duration of a conversation (about 3 minutes) and the average silent period $\frac{1}{\lambda}$ to the average time between calls (about 30 minutes). Since all subscribers do not talk on the phone all the time, phone companies take advantage of the statistical behavior of subscriber demands for service by providing a total transmission capacity that is considerably less than that required to sustain conversations if all subscribers were
to speak at the same time. We set $N$ equal to the maximum number of voice conversations that can be serviced by the link simultaneously. As a result, there are periods during which the provided transmission capacity is saturated by on-going calls and a new call request from a subscriber cannot be serviced. The probability of this occurring for a particular call is the blocking probability. The asymptotic probability $\pi_N$ is the blocking probability.

In order to save bandwidth, some modern communication systems take advantage of the silences in conversations. As indicated in Chapter 2 a threshold detector generates signals to demarcate the silent intervals. The short periods of speech between silences are called speech bursts. Only speech bursts are transmitted by the communications system. Let $N$ be the number of conversations in progress. We assume that $N$ changes slowly and therefore we can analyze the system for each individual value of $N$ separately. We use $K = N$ voice sources to model the voice traffic. The average active period duration $\frac{1}{\mu}$ is set equal to the average speech burst duration and the average silent period duration $\frac{1}{\lambda}$ for a source is set equal to the average duration of the silence between two consecutive speech bursts. Usually there is an upper limit on the transmission capacity available for speech bursts. Assume that a maximum of $N_V$ speech bursts can be accommodated at this transmission rate. The statistical behavior of humans is exploited by setting $N_V < N$. When there are $N$ on-going conversations, we would like to know the fractional loss of speech bursts, given by $\frac{\sum_{i > N_V} i\pi_i}{\sum_{i=0}^{N} i\pi_i}$. (Typical values for $\frac{1}{\lambda}$ and $\frac{1}{\mu}$...
are 0.1 and 0.1 seconds [18], respectively.)

The state transition diagram for the speech burst process is obtained from Figure 3.1 by setting \( N = K \). We shall use the model in Figure 3.1 in the rest of the study and derive results for the speech burst model as a special case. In a later chapter we show how the two models above may be combined to form a more refined model for voice traffic.

3. Data Traffic

In Chapter 2 we saw several ways to model data traffic. We choose to model data traffic as the output of a single data source that generates data continuously at a fixed data rate, \( C_D \). In reality there would be variation in the data traffic rate. If this variation is small then \( C_D \) may be chosen as an upper bound for the data generation rate, and the moments obtained for the data buffer length and delay would be upper bounds of the actual moments. The data traffic variation is indeed small when the traffic is the output of a multiplexer of a large number of data sources.

In a later chapter we describe how the theory we develop may be extended to model variation in data traffic by having multiple data sources turning on and off randomly.

4. The Link

Our model for the link is motivated by the SENET [4] implementation of the TDMA scheme described in Chapter 2. We ignore the frame structure of SENET and assume that the link transmits voice at time t
at the rate of \( r(I_t) \). \( r(i) \) is non-decreasing in \( i \). By not restricting \( r(\cdot) \) to a linear function, our model encompasses a broad range of voice coding techniques. The total transmission capacity of the link is \( B \). Voice is always transmitted at a higher priority than data. The remaining transmission capacity \( B - r(I_t) = d(I_t) \) is used for data. Data enters the system through a FIFO buffer. When the input data rate \( C_D > d(I_t) \), the data in the buffer builds up. When \( C_D < d(I_t) \) the buffer is depleted. \( B - r(N) \) is the amount of bandwidth used exclusively for data. This assures data a minimum grade of service.
$K$: Population size  \hspace{1cm} N: Maximum number of simultaneous conversations

Markov Chain for the Voice Process

$I_t$

Figure 3.1
CHAPTER 4

Performance Analysis for Voice Traffic

1. Introduction

Performance parameters for voice traffic depend on the strategy used for voice transmission — whether transmission lasts for the entire duration of a conversation or only during occurrences of speech bursts. In this section we limit ourselves to voice performance analysis for conventional telephone systems in which transmission of voice signals continues between two speakers for the entire duration of a conversation. Friedman [11] carried out a detailed study of the voice performance statistics for a voice transmission system that transmits only speech bursts.

2. Blocking Probability and Throughput

$I_t$ (defined in the previous chapter) is the number of on-going conversations at time $t$. We choose time units so that the average duration of a conversation $\frac{1}{\mu}$ is one unit. The average (non-blocked) duration between calls is $\frac{1}{\lambda}$ time units. The asymptotic probability of there being $i$ on-going conversations is $\pi_i = \frac{\binom{k}{i}(\lambda)^i}{\sum_{j=0}^{N} \binom{k}{j}(\lambda)^j}$. $\pi_N$ is the asymptotic probability of the link carrying the maximum number of allowable conversations and also the asymptotic probability that an incoming call is blocked and lost.
The expected number of on-going conversations at time $t$ is $V_t = \sum_{i=0}^{N} i \Pr[I_t = i]$. The asymptotic expectation is $V = \sum_{i=0}^{N} i \pi_i$. $V$ can be considered as a measure of the average voice throughput.

We relate the asymptotic blocking probability $\pi_N$ and voice throughput $V$ by observing:

$$\sum_{i=1}^{N} \binom{K}{i} i \lambda^{i-1} = \sum_{i=1}^{N+1} \binom{K}{i} i \lambda^{i-1} - (K-N) \binom{K}{N} \lambda^N.$$  

$$= \sum_{j=0}^{N} \binom{K}{j} (K-j) \lambda^j - (K-N) \binom{K}{N} \lambda^N.$$  

$$= \sum_{i=0}^{N} \binom{K}{i} \lambda^i - \sum_{i=0}^{N} \binom{K}{i} i \lambda^i - (K-N) \binom{K}{N} \lambda^N.$$  

Dividing through by $\sum_{i=0}^{N} \binom{K}{i} \lambda^i$, and rearranging terms we get,

$$V = \frac{K\lambda}{1+\lambda} \left[ 1 - \frac{(K-N)}{K} \pi_N \right]. \quad (4.1)$$

In Figure 4.1a we plot the variation of blocking probability $\pi_N$ as a function of the maximum allowable number of simultaneous conversations for a fixed population of telephone users. There are 150 users in the population each of whom attempts to make a 3 minute call approximately every half hour ($K=150$ and $\lambda=0.1$). We find that the blocking probability falls quickly with increasing number of allowable conversations. The curve flattens out at $N=25$ as the blocking probability gets close to 0. There is a knee around $N = \frac{K\lambda}{1+\lambda} = 13.6$, which from equation (4.1) is the average throughput of voice in the absence of blocking. This curve illustrates a well known result that
a customer population of size $K$ can be served adequately (assuring a low blocking probability) by permitting a maximum of only $N$ ($< K$) conversations at any moment. By providing transmission capacity only to support 30 simultaneous conversations the phone company may engineer a link to serve 150 customers with a blocking probability value that is almost 0.

In Figure 4.1b we plot the voice throughput or the average number of simultaneous conversations as a function of the number of allowable conversations $N$. As expected voice throughput increases with increasing $N$. The rise is steep initially when increasing $N$ allows previously blocked calls to go through. There is a knee around $N = \frac{KA}{1 + \lambda}$ and the throughput approaches $\frac{KA}{1 + \lambda}$ beyond $N = 25$, where the blocking probability is so small that most calls go through.

3. Scaling Effects

Observing today's decreasing trend in communications cost we may expect greater communications traffic in the future. This motivates us to compare the performance of a communications link for a certain population size with the performance of a communications link that handles several times the population and has proportionally larger transmission capacity. We let $N$ and $K$ vary with the constraint that their ratio $\frac{N}{K} = \alpha$ remains constant. The behavior of each individual customer, and hence $\lambda$, remains the same. We let $N$ be the measure of the size of the voice system. The plot of the blocking probability and the voice throughput as a function of the system size appears in
Figure 4.2 for $\alpha = \frac{1}{6.5}$ and $\lambda = 0.1$. The blocking probability declines with size and the voice throughput increases linearly for large sizes.

In order to explain these trends, we compute the expressions for blocking probability and voice throughput in the asymptotic case for large values of $K$, assuming that the ratio $\frac{N}{K} = \alpha < 1$. The blocking probability

$$\pi_N = \frac{\binom{K}{N} \lambda^N}{\sum_{i=0}^{N} \binom{K}{i} \lambda^i \frac{(1+\lambda)^K}{(1+\lambda)^K}} = \frac{\binom{K}{N} \lambda^N}{\sum_{i=0}^{N} \binom{K}{i} \frac{\lambda^i}{(1+\lambda)^K}}.$$  \hspace{1cm} (4.2)

The denominator may be written in terms of the beta function $B(\cdot, \cdot)$ and an integral [19] as

$$\sum_{i=0}^{N} \binom{K}{i} \frac{\lambda^i}{(1+\lambda)^K} = \frac{1}{B(K-N,N+1)} \int_0^{1+\lambda} t^{K-N-1} (1-t)^N dt$$

$$= \frac{1}{B(K-N,N+1)} \int_0^{1+\lambda} t \exp[Kh(t)] dt,$$  \hspace{1cm} (4.3)

where

$$h(t) = (1-\alpha)\ln(t) + \alpha \ln(1-t).$$

To obtain the asymptotic value for the integral we need to know the location of the maxima of $h(t)$.

$$h'(t) = \frac{1-\alpha}{t} - \frac{\alpha}{1-t}$$

$$h''(t) = -\frac{1-\alpha}{t^2} - \frac{\alpha}{(1-t)^2}.$$
from which we conclude that \( h(t) \) has a unique maximum at \( t = 1 - \alpha \).

Three cases are examined according to the relative values of \( \frac{1}{1+\lambda} \) and \( 1 - \alpha \). Let \( \beta = \frac{1}{1+\lambda} \), for shorthand notation.

**Case 1:** \( \beta < 1 - \alpha \)

Figure 4.3a shows the general shape of \( h(t) \) and the locations of \( \beta \) and \( 1 - \alpha \). The maximum for \( h(t) \) occurs for \( t \) that lies on the right of the interval \( (0, \frac{1}{1+\lambda}] \). Application of Laplace's method [20] gives the following asymptotic approximation,

\[
\int_0^\beta \frac{e^{K t} h(t)}{h'(\beta)} e^{K h(\beta)} \to \frac{e^{K h(\beta)}}{h'(\beta)} e^{K h(\beta)}, \text{ for large } K. \tag{4.4}
\]

Using (4.4) in (4.3) we get

\[
\sum_{i=0}^{N} \frac{\lambda i}{(1+\lambda)^K} = \frac{1}{B(K-N,N+1)} \left( \frac{1}{1+\lambda} \right)^K \frac{\lambda^N}{K(1-\alpha(1+\lambda))} \tag{4.5}
\]

Substituting (4.5) in (4.2),

\[
\pi_N = 1 - \frac{N}{\lambda(K-N)}.
\]

(The two conditions \( K > N \) and \( \frac{1}{1+\lambda} < 1 - \alpha \) ensure that \( 0 < \pi_N < 1 \).) If we substitute this value of \( \pi_N \) in (4.1) we get the asymptotic voice throughput \( V_t = N \), indicating that the average number of simultaneous conversations is equal to the maximum allowed.

The asymptotic analysis reveals that under the condition \( \frac{1}{1+\lambda} < 1 - \alpha \), the blocking probability tends to a constant non-zero value as \( K \) becomes large. This is undesirable, especially if this constant value is unacceptably large. The above inequality may be alternately
written as \( N < \frac{K\lambda}{1+\lambda} \). In words, the maximum number of conversations permitted is less than the average number in the absence of blocking.

To avoid having this fixed non-zero asymptotic blocking probability the inequality should be reversed.

**Case 2 : \( \beta > 1 - \alpha \).**

Figure 4.3b shows the general shape of \( h(t) \) and the locations of \( \beta \) and \( 1 - \alpha \). The maxima for \( h(t) \) occurs for \( t \) in the interval \( (0, \frac{1}{1+\lambda}) \). Application of Laplace's method [20] gives the following asymptotic approximation,

\[
\beta \int_0^\infty g(t) e^{Kh(t)} dt = g(1-\alpha) \left( \frac{-2\pi}{Kh''(1-\alpha)} \right)^{1/2} e^{Kh(1-\alpha)}, \text{ for large } K. \quad (4.8)
\]

Using (4.8) in (4.3) we get

\[
\sum_{i=0}^{N} \left( \frac{K}{1+\lambda} \right)^i = \frac{1}{B(K-N, N+1)} \left( \frac{2\pi a}{K(1-\alpha)} \right)^{1/2} \left( 1 - \alpha \right)^{1-\alpha} \alpha. \quad (4.9)
\]

Substituting (4.9) in (4.2)

\[
\pi_N = \left( \frac{1}{2\pi a(1-\alpha)} \right)^{1/2} \frac{1}{\sqrt{K}} \left( \frac{1}{1+\lambda} \right) \alpha \left( \frac{1}{1+\lambda} \right) \left( \frac{1}{1-\alpha} \right)^{1-\alpha} \alpha K
\]

\[
= \left( \frac{1}{2\pi (1-\alpha)} \right)^{1/2} \frac{1}{\sqrt{N}} \left( \frac{1}{1+\lambda} \right) \alpha \left( \frac{1}{1+\lambda} \right) \left( \frac{1}{1-\alpha} \right)^{1-\alpha} \alpha K.
\]

For any fixed value of \( \alpha < 1 \) and variable \( \lambda \) the maximum value of \( \left( \frac{1}{1+\lambda} \right) \alpha \left( \frac{1}{1+\lambda} \right) \left( \frac{1}{1-\alpha} \right)^{1-\alpha} \) is 1 (when \( \frac{1}{\lambda} = \frac{1-\alpha}{a} \)). Hence the blocking probability approaches 0 at least at the rate \( O\left( \frac{1}{\sqrt{K}} \right) \) when \( \frac{1}{1+\lambda} > 1 - \alpha \), which can be rewritten as \( N > \frac{K\lambda}{1+\lambda} \). We see this trend in Figure 4.2
in which we chose the parameters of the system to satisfy the inequality. From (4.1) we conclude that voice throughput must approach \( \frac{K\lambda}{1+\lambda} \) asymptotically at the same rate, which is also reflected in Figure 4.2.

**Case 3**: \( \beta = 1 - \alpha \).

Figure 4.3c shows the general shape of \( h(t) \) and the locations of \( \beta \) and \( 1 - \alpha \). The maximum for \( h(t) \) occurs for \( t \) at the right endpoint of the interval \( (0, \frac{1}{1+\lambda}) \). Application of Laplace’s method [20] gives the following asymptotic approximation.

\[
\int_{0}^{\beta} g(t) e^{K h(t)} dt = g(\beta) \left( \frac{-\pi}{2K h''(\beta)} \right)^{1/2} e^{K h(\beta)}
\]  

(4.10)

Using (4.10) in (4.3) we get

\[
\sum_{i=0}^{N} \frac{\lambda^i}{(1+\lambda)^i} = \frac{1}{B(K-N,N+1)} \left( \frac{\pi\alpha}{2K(1-\alpha)} \right)^{1/2} \frac{1}{(1+\lambda)^{K+1}}.
\]  

(4.11)

Substituting (4.11) in (4.2)

\[
\pi_N = \left( \frac{2}{\pi\alpha(1-\alpha)} \right)^2 \frac{1}{\lambda^K} = \left( \frac{2}{\pi(1-\alpha)} \right)^2 \frac{1}{\lambda^K}.
\]

From this result we conclude that as long as the transmission capacity provided for voice is as least as much as that required to sustain the average number of conversations in the absence of blocking then we can expect at least an \( O(\frac{1}{\lambda^K}) \) decline in the blocking probability with increase in transmission capacity (measured by \( N \)) and the customer population (\( K \)), as long as their ratio is a con-
stant.
$K = 150$
$\lambda = 0.1$

**Figure 4.1a**

Blocking Probability Vs Maximum Conversations

**Figure 4.1b**

Average Number of Conversations Vs Maximum Conversations
Effect of Scaling on Voice Performance

Figure 4.2

\[ K = 6.5 \text{ N} \]
\[ \alpha = \frac{1}{6.5} \]
\[ \frac{\lambda}{\mu} = 0.1 \]
Case 1:

\[ h(t) \]

\[ 0 \quad \beta \quad 1-\alpha \quad t \]

Figure 4.3a

Case 2:

\[ h(t) \]

\[ 0 \quad 1-\alpha \quad \beta \quad t \]

Figure 4.3b

Case 3:

\[ h(t) \]

\[ 0 \quad 1-\alpha \quad \beta \quad t \]

Figure 4.3c

Relative positions of \( 1-\alpha \) and \( \beta \)

Figure 4.3
CHAPTER 5

Analysis for Data: Buffer Length

1. Introduction

The performance of an integrated data-voice link is measured by the asymptotic distribution of the data buffer and the data delay. In this chapter we develop a general theory for obtaining the buffer length distribution that is applicable to a wide range of problems of which the analysis of our model of the integrated link is only one. We examine the relevant algebraic properties of the equations governing the asymptotic probability distribution. The asymptotic buffer length distribution is a sum of exponential terms. The multiplicative coefficients for each of these terms and the exponential coefficients themselves have to be evaluated. Two methods to evaluate the multiplicative coefficients are discussed. Computation of the exponential coefficients entails the solution of a generalized eigenvalue problem. We include a description of algorithms we used to solve the eigenvalue problem. This approach was first considered in [21].

The random process $Y_t$, taking real non-negative values, describes the amount of data (number of bits) in the data buffer or the data buffer length at time $t$. We assume that we start observing the integrated link at time 0 at which time the data buffer is empty. As per our model for the integrated data-voice link, the data in the
buffer accumulates at a rate \( C_d - d(I_t) \) when \( C_d - d(I_t) > 0 \). A non-empty buffer depletes at the rate \( d(I_t) - C_d \) when the inequality is reversed. \( Y_t \) may be formally defined as follows,

\[
Y_t = \int_0^t [C_d - d(I_\tau)]^+ d\tau - \int_0^t [C_d - d(I_\tau)]^- I_{Y_t > 0} d\tau,
\]

where \([a]^+ = \max(a, 0)\), \([a]^− = \max(−a, 0)\) and \(I_{\{E\}}\) is the indicator function of the event \(E\). The first term accounts for buffer build-ups and the second for depletions.

2. General Problem

In this section we determine the asymptotic joint distribution of two random processes \(I_t\) and \(Y_t\). \(Y_t\) is a continuous non-negative valued random process that is related to \(I_t\) by a set of equations that govern their joint probability distribution. \(I_t\) is an irreducible discrete state continuous time stationary Markov process with a finite number of states \(\{0, 1, 2, \ldots, N\}\). Let \(\Pr[Y_t \leq y \text{ and } I_t = i] = P_i(t, y)\) and \(P(t, y) = (P_j(t, y))\) be a column matrix. The governing equations for the joint probability distribution are:

\[
\frac{\partial}{\partial t} P(t, y) + D \frac{\partial}{\partial y} P(t, y) = M \cdot P(t, y), \tag{5.1}
\]

for some non-singular matrix \(D\). In a subsequent chapter \(D\) will be constrained to be diagonal. We assume that \(Y_t\) has a unique asymptotic probability distribution and is bounded almost surely. For queuing processes such an assumption is synonymous with the assumption of system stability. \(P_i(t, \infty) = \Pr[I_t = i]\). If we let \(y \to \infty\) in equation (5.1) we get
\[
\frac{d}{dt} P(t, \infty) = M P(t, \infty).
\]

Hence, \( M \) is the rate generator matrix for the Markov process \( I_t \).

Specifically the \((i,j)\)th element \((M)_{i,j} = \frac{Pr[I_{t+At} = i | I_t = j]}{\Delta t}\), when \(i \neq j\) and \((M)_{i,i} = -\left( \sum_{j=0}^{N} (M)_{i,j} \right)\). The asymptotic distribution for \(Y_t\) may be defined in terms of

\[
F_i(y) = \lim_{t \to \infty} P_i(t, y).
\]

Let \( F(y) = (F_i(y)) \) be a column vector. From (5.1) we write

\[
D \frac{d}{dy} F(y) = M F(y). \tag{5.2}
\]

The formal solution for \( F(y) \) in (5.2) is given [22] by

\[
F(y) = \sum_{j=0}^{N} a_j e^{z_j y} \xi_j, \tag{5.3}
\]

where \((z_j, \xi_j)\) for \(j=0,1,2,\ldots,N\) are the eigenvalue and corresponding right eigenvector pairs for the eigenvalue problem

\[
z D \xi = M \xi. \tag{5.4}
\]

and the \(a_j\)'s are coefficients to be determined.

Let us denote \( F(0) = \mathbf{1} \), and define left eigenvectors \( \xi_j, j = 0,1,2,\ldots,N \), by the relation

\[
z \xi_j^T D = \xi_j^T M. \tag{5.5}
\]

We assume that all the eigenvalues are distinct.
Note that
\[ z_i \cdot \psi_{iD}^T \cdot \phi_j = \psi_{iM}^T \cdot \phi_j \]
\[ = z_j \cdot \psi_{iD}^T \cdot \phi_j . \]

Hence,
\[ \psi_{iD}^T \cdot \phi_j = 0 \text{ iff } z_i \neq z_j . \quad (5.6) \]

Using (5.6) in (5.3) we may solve for the \( a_j \)’s and write
\[ F(y) = \frac{N}{\sum_{j=0}^{\infty} \psi_{jD} \cdot \phi_j} \cdot \sum_{j=0}^{\infty} \psi_{jD} \cdot \phi_j . \quad (5.7) \]

The asymptotic distribution for \( Y_t \) is \( \lim_{t \to \infty} P[Y_t < y] = F_T(y) \cdot 1 \).

From the asymptotic distribution we can compute all moments for the buffer length. One solution method is to compute the eigenvalues \( z_j \) and the eigenvectors \( \phi_j \) and \( \psi_j \). If \( \phi \) is known we may use (5.7) to give us the asymptotic distribution of \( Y_t \). In many problems \( \phi \) is not known explicitly and we need techniques for the evaluation of the \( N+1 \) coefficients \( a_j \).

3. System Dynamics

The random processes \( I_t \) and \( Y_t \) describing the number of ongoing conversations and the length of the data buffer at time \( t \), respectively, satisfy the assumptions of the previous section. To show this we observe transitions in the interval \((t, t+\Delta t)\), and write for \( y \)

\[ \frac{1}{\Delta t} : \text{the N+1 length vector (1,1,1,...,1,1) of all 1's.} \]
\[ P_i(t+\Delta t, y) = (K-(i-1)) \lambda \Delta t \: P_{i-1}(t, y) + \]

\[ [1 - (K-i) \lambda \Delta t - i \Delta t] \: P_i(t, y + [d(i) - C_D] \Delta t) + \]

\[ (i+1) \Delta t \: P_{i+1}(t, y) + O(\Delta t^2) \quad \text{for } 0 < i < N, \]

\[ P_0(t+\Delta t, y) = [1-K \lambda \Delta t] \: P_0(t, y + [d(0) - C_D] \Delta t) + \]

\[ \Delta t \: P_1(t, y) + O(\Delta t^2) \]

\[ P_N(t+\Delta t, y) = [K-(N-1)] \lambda \Delta t \: P_{N-1}(t, y) + \]

\[ (1-N \lambda \Delta t) \: P_N(t, y + [d(N) - C_D] \Delta t) + O(\Delta t^2). \]

Rearranging terms, dividing through by \( \Delta t \) and taking the limit as \( t \to \infty \), we get

\[ \frac{\delta P_i}{\delta t}(t, y) + (C_D - d(i)) \frac{\delta P_i}{\delta y}(t, y) = (K-(i-1)) \lambda P_{i-1}(t, y) + \]

\[ -(K-i) \lambda \lambda P_i(t, y) + \]

\[ [i+1] P_{i+1}(t, y) \quad 0 < i < N \]

\[ \frac{\delta P_0}{\delta t}(t, y) + (C_D - d(0)) \frac{\delta P_0}{\delta y}(t, y) = -K \lambda P_0(t, y) + P_1(t, y) \]

\[ \frac{\delta P_N}{\delta t}(t, y) + (C_D - d(N)) \frac{\delta P_N}{\delta y}(t, y) = (K-(N-1)) \lambda P_{N-1}(t, y) \]
These equations can be put in the form of the matrix equation (5.1) by identifying $D$ and $M$ as described in Figures 5.1 and 5.2. In the next section we will refer to a theorem stating that the eigenvalues of the system described in (5.4) are distinct, for our choice of $D$ and $M$.

4. Algebraic Properties

Let us examine some of the structural properties of the matrices $D$ and $M$ before solving for the asymptotic probability distribution of the buffer length.

(a) $M$ is the generator matrix for the voice process $I_t$. By definition,

$$A^T M = 0^T 2.$$ 

The vector $\pi$ of asymptotic state probabilities for $I_t$ can be determined from the following equalities:

$$M \cdot \pi = 0 \quad \text{(Balance Equations)}$$

$$1^T \cdot \pi = 1 \quad \text{(Normalization Criterion)}$$

(b) Since the Markov process is a simple birth and death process the local balance equation between any two states is satisfied. Hence we can write

$$2 \ 0^T = (0,0,\ldots,0,0),$$

a row vector of $N+1$ elements, each of which is 0.
\[ \mathbf{M} \cdot \mathbf{\Pi} = \mathbf{\Pi} \cdot \mathbf{M}^T , \quad (5.9) \]

where \( \mathbf{\Pi} = \text{diag}(\pi_i) \) is an \((N+1)\times(N+1)\) diagonal matrix. We know that \( \pi_i \neq 0 \), for \( 0 \leq i \leq N \) and hence \( \mathbf{\Pi} \) is non-singular. Pre- and post-multiply both sides of equation (5.9) by \( \mathbf{\Pi}^{-1/2} \) to get

\[ \left( \mathbf{\Pi}^{-1/2} \right)^2 \mathbf{M} \cdot \mathbf{\Pi} \cdot \mathbf{\Pi}^{-1/2} = \left( \mathbf{\Pi}^{-1/2} \right)^2 \mathbf{M} \cdot \mathbf{T} \cdot \mathbf{\Pi} \cdot \mathbf{\Pi}^{-1/2} . \quad (5.10) \]

Let \( \mathbf{M}_s = \mathbf{\Pi}^{-1/2} \mathbf{M} \cdot \mathbf{\Pi}^{-1/2} \). The transposes of diagonal matrices \( \mathbf{\Pi}^{-1/2} \) and \( \mathbf{\Pi} \) are the matrices themselves. Equation (5.10) establishes that \( \mathbf{M}_s \) is symmetric. \( \mathbf{M}_s \) is also tridiagonal.

(c) We will show that \( \left( -\mathbf{M} \right) \) and \( \left( -\mathbf{M}_s \right) \) are both positive semi-definite and will refer to a theorem that proves that every proper minor of these matrices is non-zero. Since \( \mathbf{M}_s \) is obtained from \( \mathbf{M} \) through the application of a similarity transform they have the same real eigenvalues.

The sum of elements in each column of \( \left( -\mathbf{M} \right) \) is 0. Diagonal elements are positive and off-diagonal elements are non-positive. Gerschorn's theorem [23] states that the eigenvalues of \( \left( -\mathbf{M} \right) \) are contained in the union of \( N+1 \) circular regions in the complex plane, the \( i \)'th region centered at \( \left( -M_{i,i}, 0 \right) \) (a point on the positive real axis) and of radius \( -M_{i,i} \), for \( 0 \leq i \leq N \). Since the union does not contain any points in the negative half plane, all eigenvalues are non-negative. Therefore \( \left( -\mathbf{M} \right) \) is positive semi-definite. Since \( \left( -\mathbf{M}_s \right) \) and \( \left( -\mathbf{M} \right) \) have the same eigenvalues \( \left( -\mathbf{M}_s \right) \) is also positive.
A theorem in [24] asserts that every proper minor of \((-M_s)\) and \((-M_s)\) is positive.

(d) We now examine the signs of the eigenvalues of \(zD \Phi = M \Phi\). We have restricted ourselves to cases where \(D\) is non-singular. This is not a limitation on our analysis because if we do have to consider a link model for which \(D\) is non-singular this can be only because some diagonal element \(d_{i,i} = C_D - d(i) = 0\). We may analyze the delay and buffer length with \(C_D\) replaced everywhere by \(C_D + \Delta\) and obtain the required solution by letting \(\Delta \to 0\), since our data performance parameters are continuous with respect to \(C_D\).

Mitra [25] proved that, for non-singular matrices \(D\) and any arbitrary irreducible Markov generator matrix \(M\) all the \(N+1\) eigenvalues \(z_1\) of (5.4) are real and distinct, with one eigenvalue at 0. If \(\pi^T D \cdot \mathbf{1} < 0\), there are as many negative eigenvalues as there are positive terms in the main diagonal of \(D\). If \(\pi^T D \cdot \mathbf{1} > 0\), there are as many positive eigenvalues as there are negative terms in the main diagonal of \(D\).

5. Buffer Length Distribution

This section contains two procedures to obtain the distribution for the buffer length \(F(y)\), on the assumption that all the eigenvalues \(z_1\) and the right eigenvectors \(\Phi_1\) are known.

We first demonstrate a method to obtain left eigenvectors from right eigenvectors. Let us apply the similarity transformation
(5.10) to matrices on both sides of equations (5.4) and (5.5) to set up the following eigenvalue problem.

\[
z \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} \mathcal{D} \frac{1}{2} \end{pmatrix} \tilde{\mathbf{x}} = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} \mathcal{M} \frac{1}{2} \end{pmatrix} \tilde{\mathbf{x}}.
\]

(5.11)

Eigenvalues are preserved through a similarity transformation. The eigenvalues obtained from (5.11) are the same as the eigenvalues from (5.4) and (5.5). The eigenvectors \(\tilde{\mathbf{x}}_1\) and \(\tilde{\mathbf{x}}_1\) for (5.11) are related to \(\tilde{\mathbf{x}}_1\) and \(\tilde{\mathbf{x}}_1\) for (5.4) and (5.5) by

\[
\frac{1}{2} \tilde{\mathbf{x}}_1 = \tilde{\mathbf{x}}_1 \text{ and } \frac{1}{2} \tilde{\mathbf{x}}_1 = \tilde{\mathbf{x}}_1.
\]

(5.12)

Since \(\frac{1}{2} \mathcal{D} \frac{1}{2}\) and \(\frac{1}{2} \mathcal{M} \frac{1}{2}\) are symmetric, the left and right eigenvectors of (5.11) are identical. Using (5.12) we relate

\[
\tilde{\mathbf{x}} = \mathcal{D}^{-1} \tilde{\mathbf{x}}.
\]

(5.13)

\(\mathcal{D}^{-1}\) is a known diagonal non-singular matrix. From (5.13) all left eigenvectors may be expressed in terms of a right eigenvector.

Under the canonical condition that the average amount of transmission capacity required of the system is less than the maximum transmission capacity, the buffer length and the waiting time are bounded almost surely. This stability condition can be written concisely as \(\mathcal{T} \mathcal{D} \mathbf{1} < 0\). We assume that this stability condition is satisfied, and assume that \(\mathcal{D}\) has \(m\) positive terms along the diago-
nal. From (d) in the previous section we conclude that there must be m negative eigenvalues. We label the eigenvalues \( z_0 \) through \( z_N \) such that \( z_0 < z_1 < z_2 < z_3 \ldots \ldots < z_N \). Note that \( z_m = 0 \).

The positive eigenvalues \( z_{m+1} \) through \( z_N \) contribute unstable modes to the buffer length distribution function \( F(y) \) in (5.7). These modes cannot be excited. From (5.7) we conclude

\[
\psi_{i}^{T} \cdot D \cdot \phi = 0, \quad i = N, N+1, \ldots, m+1
\]  

(5.14)

From (5.7) and (5.14) we observe that only the eigenvalue \( z_m = 0 \) contributes to \( F(\omega) \), giving

\[
F(\omega) = \frac{\psi_{m}^{T} \cdot D \cdot \phi}{\psi_{m}^{T} \cdot D \cdot \phi_m}
\]  

(5.15)

Note that \( F(\omega) = \lim_{t \to \infty} \Pr[I_t = i] = \pi_i \). The normalization constraint requires

\[
1^{T} \cdot F(\omega) = 1
\]  

(5.16)

From (5.15) and (5.16)

\[
\psi_{m}^{T} \cdot D \cdot \phi = \frac{\psi_{m}^{T} \cdot D \cdot \phi_m}{1^{T} \cdot \phi_m}
\]  

(5.17)

If the number of on-going conversations \( I_t \) is such that the data input rate exceeds the data transmission rate \( (C_D > d(I_t)) \), then on balance the buffer content is increasing. In particular the buffer cannot be empty. \( C_D - d(I_t) \) is a term that appears in the diagonal of \( D \). \( d(i) \) is monotone non-increasing in \( i \). Exactly \( m \) of the
diagonal terms are positive. These are \( C_D - d(i) \) for \( i \in \{N, N-1, \ldots, N-m+1\} \). It follows that

\[
F_i(0) = 0, \quad N-m+1 \leq i \leq N. \quad (5.18)
\]

From the definition of \( f = F(0) \) and (5.18) we can write

\[
f^T = (f_1^T, 0^T), \quad (5.19)
\]

where \( f_1 \) is a vector of length \( N-m+1 \) and \( 0 \) is a vector of length \( m \).

We need \( N-m+1 \) independent equations to solve for \( f \). There are \( N-m+1 \) equations available from (5.14) and (5.17). These can be written concisely as

\[
\underline{A} f_1 = \begin{bmatrix}
\eta_T^D \phi_m^n \\
\eta_T^n \phi_m \\
1^T \phi_m
\end{bmatrix} \underline{u}, \quad (5.20)
\]

where the coefficients in (5.14) and (5.17) multiplying the unknown components of \( f \) are collected in \( \underline{A} \) and \( \underline{u}^T = (1, 0, 0, \ldots, 0) \) is a vector of length \( N-m+1 \).

Note, that in general \( \underline{A} \) is not a sparse matrix. We need to invert \( \underline{A} \) to find \( f_1 \). Using (5.19), (5.13) and (5.7) we write the asymptotic buffer distribution as

\[
F(y) = \sum_{i=0}^{m} \underbrace{z_i y (\eta_{-1} \phi_i^T D \phi_i \phi_i)}_{\underline{u}}.
\]

Anick et. al. [26] demonstrated another procedure to obtain the asymptotic distribution that avoids the inversion of the potentially dense matrix \( \underline{A} \). The buffer length distribution is written in the
form

\[ F(y) = F(0) + \sum_{i=0}^{m-1} a_i e^{z_i y} \]

(5.20)

We know that \( F(0) = \pi \) and \( F_i(0) = 0 \), \( N-m+1 \leq i \leq N \). From the tridiagonality of \( D^{-1} M \) we have

\[
\left[ (D^{-1} M)^j F(0) \right] = 0 \quad \text{for} \quad N-m+1+j \leq i
\]

(5.21)

where \( 0 \leq j \leq m-1 \).

Differentiating (5.2) w.r.t. \( t \) and using (5.21) we get

\[
\frac{d^j}{dy^j} F_N(y) \bigg|_{y=0} = 0 \quad 0 \leq j \leq m-1.
\]

(5.22)

We scale all eigenvalues so that \((\phi_i)_N = 1\). From (5.20)

\[ F_N(y) = \pi_N + \sum_{i=0}^{m-1} a_i e^{z_i y} \]

(5.23)

Substituting (5.23) in (5.22) we get

\[ V \bar{a} = -\pi_N u \]

where \( V \) is an \( m \times m \) Van der Monde's matrix with \((V)_{i,j} = z_j^i\) and \( a^T = [a_0, a_1, \ldots, a_{m-1}]\). \( V \) is non-singular because the eigenvalues \( z_i \) are distinct. The inverse of the Van der Monde matrix is well known.

We obtain
\[ a_j = -\pi_N \frac{\prod_{i=0}^{m-1} z_i}{z_i - z_j} \quad 0 \leq j \leq m-1. \]

\( F(y) \) is specified completely in terms of the eigenvalues and the eigenvectors.

6. **On Solving the Eigenvalue Problem**

We need the eigenvalues and the eigenvectors of the generalized eigenvalue problem \( z_0 \phi = M \phi \) to express the distribution function for the buffer length. We used two numerical techniques for this purpose.

In the first method [27] we make the observation that \( D^{-1}M \) is a lower Hessenberg matrix and use algorithms developed to find the eigenvalues and eigenvectors for such matrices.

In the second method [28] we observe that the rank of \( M \) is \( N \). We use a similarity transform to symmetrize \( M \) to \( M_S \). \( M_S \) is tridiagonal and in (d) of Section 2 we have shown that every proper minor of \( -M_S \) is non-zero. These are exactly the conditions that we need to apply a modified version of Crawford’s algorithm [29] to convert the eigenvalue problem \((-z)D \xi = (-M_S) \xi\) into an eigenvalue problem of the form

\[ (-z)T_D \xi = \begin{bmatrix} 0 & 0 \\ 0 & T \end{bmatrix} \begin{bmatrix} \xi \\ I_{N \times N} \end{bmatrix} \]

where \( T_D \) is a tridiagonal matrix. The time complexity of the conversion algorithm is \( O(N^2) \). \( T_D \) can be written as \( \begin{bmatrix} a & b^T \\ c & d \end{bmatrix} \) and \( \xi \) as
where \( \mathbf{c}^T = (c_1, 0, 0, \ldots, 0) \), \( \mathbf{b}^T = (b_1, 0, 0, \ldots, 0) \) and \( \mathbf{d} \) is tridiagonal. The eigenvalue problem can be written as

\[
\begin{bmatrix}
\mathbf{a}^T \\
\mathbf{c}^T
\end{bmatrix}
\begin{bmatrix}
\mathbf{x} \\
\mathbf{y}
\end{bmatrix} = \frac{1}{(-z)}
\begin{bmatrix}
0 & \mathbf{c}^T \\
\mathbf{d} & 0
\end{bmatrix}
\begin{bmatrix}
\mathbf{x} \\
\mathbf{y}
\end{bmatrix}.
\]

This gives \( \mathbf{x} = \frac{1}{\mathbf{a}^T} \mathbf{b}^T \mathbf{y} \) and

\[
\left( \frac{1}{\mathbf{a}^T} \mathbf{a} \mathbf{b}^T + \mathbf{d} \right) \mathbf{y} = \frac{1}{(-z)} \mathbf{y}.
\]

Except for \( (\mathbf{a} \mathbf{b}^T)_{1,1} \) all the elements of \( \mathbf{a} \mathbf{b}^T \) are 0. \( \left( \frac{1}{\mathbf{a}^T} \mathbf{a} \mathbf{b}^T + \mathbf{d} \right) \) is a tridiagonal system. We have deflated the original tridiagonal system of the zero eigenvalue. We use the QL algorithm to solve for the eigenvalues of this system. Having obtained the eigenvalues we use the inverse power method for finding the required eigenvectors using the original matrices \( \mathbf{N} \) and \( \mathbf{D} \).

For the special case \( K = N \) and \( d(i) = B - i \) (i.e. \( r(i) = i \)) Anick et. al. [26] devised a technique to obtain the \( N+1 \) eigenvalues by solving \( \left[ \frac{N+1}{2} \right] \) quadratic equations. From (5.20) we know

\[
\lim_{t \to \infty} \Pr[ Y_t \leq y ] = \frac{1}{T \cdot \Pi} + \sum_{i=0}^{m-1} a_i \frac{1}{T \cdot \phi_i}.
\]

Anick et. al. found an expression for \( \frac{1}{T \cdot \phi_i} \) involving the eigenvalue \( z_i \) and the external system parameters of the link. This special case corresponds to a link that does not block voice traffic and whose instantaneous transmission requirement for voice is linear in the number of on-going conversations.
7. Moments of the Buffer Length Distribution

From (5.20) we conclude that the asymptotic probability of the buffer length exceeding length \( y \), denoted by \( G(y) \), is

\[
G(y) = - \sum_{i=0}^{m-1} a_i e^{z_i y} \frac{1}{\lambda^{i} \theta_i}.
\]

The moments of the asymptotic buffer length distribution are

\[
\lim_{t \to \infty} E[X^j_t] = (-)^{j+1} \sum_{i=0}^{m-1} \frac{a_i j!}{z_i^j} \frac{1}{\lambda^{i} \theta_i}.
\]
\[ M = \begin{bmatrix}
-\lambda & i & 0 \\
\lambda & - (K-1) \lambda - 1 & 2 \\
0 & (K-1) \lambda & -(K-2) \lambda - 2 \\
\vdots & \ddots & \ddots \\
(K-i+1) \lambda & -(K-i) \lambda - i & i + 1 \\
\vdots & \ddots & N \\
(K-N+1) \lambda & -N
\end{bmatrix} \]

Figure 5.1
\[ D = \begin{bmatrix} C_0 - d(0) \\ C_0 - d(1) \\ \vdots \\ C_0 - d(N) \end{bmatrix} \text{ an } (N+1) \times (N+1) \text{ diagonal matrix.} \]

Figure 5.2
CHAPTER 6

Analysis for Data: Delay

1. Introduction

In this chapter we seek the asymptotic distribution for the data delay. We consider a FIFO buffer for data, restricting the applicability of our analysis to links with a single grade of data service.

The general method used involves conditioning the data delay on the buffer length seen by arriving data. The data delay is the time to deplete the buffer of that amount of data. We are able to obtain the Laplace transform of the delay whenever the buffer length distribution is a sum of exponentials and the data transmission rate varies only with the number of ongoing conversations. If, furthermore, whenever the buffer is non-empty the sum of the instantaneous data and voice transmission rate is constant and equal to the maximum transmission rate for the link, then the transform of the delay distribution is easily inverted to a sum of exponentials. The exponentials for the delay distribution and the buffer distribution are related in their exponential coefficients by a constant factor - the data arrival rate.

We define the data delay $W_t$ experienced by the data entering the buffer at time $t$ as
\[ W_t = w \text{ iff } \int_t^{t+w} d(I_{\tau}) d\tau = Y_t \]  

(6.1)

Data entering at \( t \) find \( Y_t \) amount of data already in the buffer and this is removed at the rate \( d(I_{\tau}) \), \( t \leq \tau \leq t+w \). In \( w \) units of time all the data that was in the buffer at time \( t \) is depleted. We are interested in the asymptotic distribution of \( W_t \) \( (\lim_{t \to \infty} W_t) \).

2. **Theory**

In this section we will prove two theorems concerning stochastic processes \( I_t, Y_t, \) and \( W_t \). \( I_t \) is a discrete-state, continuous-time, irreducible, stationary Markov process, taking on one of a finite number of states \( \{0, 1, \ldots, N\} \). \( M \) is the generator matrix for \( I_t \). If \( \pi_i = \lim_{t \to \infty} \Pr(I_t = i) \) and \( \pi = (\pi_i) \), then \( M \pi = 0 \). \( Y_t \) and \( W_t \) are continuous-time non-negative stochastic processes. \( Y_t \) and \( W_t \) are inter-related through \( I_t \) and a positive function \( g(\cdot) \) defined on \( \{0, 1, \ldots, N\} \), by the following equivalence

\[ W_t = w \iff \int_t^{t+w} g(I_{\tau}) d\tau = Y_t \]

\( G \) is defined as the \((N+1)x(N+1)\) diagonal matrix diag(\( g(i) \)). We define \( P(t,y) = (P_i(t,y)) \), where \( P_i(t,y) = \Pr(Y_t \leq y, I_t = i) \). The asymptotic joint distribution of \( I_t \) and \( Y_t \) is of the form

\[ F(y) = \lim_{t \to \infty} P(t,y) = F(\infty) + \sum_{j=1}^{\infty} a_j e^{z_j y} \Phi_j. \]

We assume that \( Y_t \) is bounded a.s. Note that \( F_i(\infty) = \lim_{t \to \infty} \Pr[I_t = i] \).
Under these assumptions we will establish the Laplace transform of the asymptotic joint distribution of $I_t, Y_t$ and $W_t$ through the following theorem.

**Theorem 1**

Define $W = \lim_{t \to \infty} W_t, Y = \lim_{t \to \infty} Y_t$ and $I_0 = \lim_{t \to \infty} I_t$.

We are assuming that when we observe the random processes at time $t = 0$, each has reached a steady state, independent of their starting states.

Let $\Gamma_i = E(e^{-\alpha W + \delta Y}I_{(I_0 = i)}), \Gamma = (\Gamma_i)$

and $k_i = E(e^{-\delta Y}I_{(I_0 = i)}), k = (k_i)$.

Then $\Gamma = k - \sum_{j=1}^{m} \left[ a_i z_j - \sum_{j=1}^{m} (\delta - z_j)G \right]^{-1} \cdot G_j a_j \frac{z_i}{\delta - z_j}$.

**Proof:**

The proof follows from the two lemmas that we shall first state below and then prove.

**Lemma 1.1**

If $\psi_i(\alpha, \delta) = \int_{w=0}^{\infty} e^{-\alpha w} E[e^{-\delta g(I_{(I_0 = i)})}I_{(I_0 = i)}] dw$

then $\Gamma_i = k_i - \sum_{j=1}^{m} \psi_i(\alpha, \delta - z_j) a_j \frac{z_i}{\delta - z_j}$.

$I_{[E]}$ denotes the indicator function of the event $E$. 
Lemma 1.2

If \( \mathbf{a} = (a_i) \) then

\[
\mathbf{a} = (a_k - M + 5 \varepsilon)^{-1} \cdot 1
\]

To prove the theorem, we write the following vector equality on the basis of Lemma 1.1

\[
\mathbf{a} = k - a \sum_{j=1}^{M} \mathbf{a}^T (a, \delta - z_j) \cdot \mathbf{a}_j \cdot 1 \cdot 1 \cdot s_j \frac{z_j}{\delta - z_j}
\]

Now we apply Lemma 1.2

\[
\mathbf{a} = k - a \sum_{j=1}^{M} 1 \cdot 1 \cdot 1 \cdot (a_k - M + 5 \varepsilon)^{-1} \cdot \mathbf{a}_j \cdot s_j \frac{z_j}{\delta - z_j}
\]

\[
= k - a \sum_{j=1}^{M} (a_k - M + 5 \varepsilon)^{-1} \cdot \mathbf{a}_j \cdot s_j \frac{z_j}{\delta - z_j}
\]

We will now turn to proving the lemmas.

Proof of Lemma 1.1

The joint probability distribution of \( W, Y \) and \( I_0 \) may be decomposed into an absolutely continuous part described by a joint density function \( f_{WYI_0} \) and a singular part that consists of point mass concentrations. Then

\[
\int_{w=0}^{\infty} \int_{y=0}^{\infty} e^{-a w-\delta y} f_{WYI_0}(w, y, i) \, dy \, dw
\]

\[
+ \sum_{(w_k, y_k)} e^{-a w_k-\delta y_k} P \left[ (w, Y, I_0) = (w_k, y_k, i) \right]
\]
\[
= \int_{y=0}^{\infty} \int_{w=0}^{\infty} e^{-aw-\delta y} f_{W|Y,I_0} (w, y, i) \, dy \, dw + P[Y=0, I_0 = i]
\]

since only \((w_k, y_k, i) = (0, 0, i)\) has non-zero probability mass and the events 
\([Y=0, I_0 = i] = [Y=0, W=0 \text{ and } I_0 = i]\).

\[
\Gamma_i = P[Y=0, I_0 = i] + \\
\int_{w=0}^{\infty} \int_{y=0}^{\infty} e^{-aw-\delta y} f_{W|Y,I_0} (w, y, i) f_{Y|I_0} (y, i) \, dy \, dw.
\]

We have conditioned on \(Y\) and \(I_0\) and assumed that at \(t=0\), the system is in steady state. Hence, \(Y=Y_0\) and \(W=W_0\).

\[
\Gamma_i = P[Y=0, I_0 = i] + \\
\int_{w=0}^{\infty} \int_{y=0}^{\infty} e^{-aw-\delta y} \frac{d}{dw} F_{W|(bW|Y,I_0)} (w | y, i) \cdot \\
\frac{d}{dy} F_{Y|I_0} (y, i) \, dy \, dw \quad \text{a.s.}
\]

where

\[
F_{W|(bW|Y,I_0)} (w | y, i) = Pr[W \leq w | (bW, Y = y, I_0 = i)]
\]

and 
\(F_{Y|I_0} (y, i) = Pr[Y \leq y \text{ and } I_0 = i]\).

\[
\Gamma_i = P[Y=0, I_0 = i] + \\
\alpha \int_{y=0}^{\infty} \int_{w=0}^{\infty} e^{-aw-\delta y} F_{W|(bW|Y,I_0)} (w | y, i).
\]
\[
\frac{d}{dy} F_{Y|I_{0}}(y, i) \, dw \, dy
\]

We have changed the order of integration, integrated by parts w.r.t. the \( w \) variable, and used the fact that

\[ F_{W|Y_{I_{0}}} (0|y, i) = 0 \text{ for } y > 0. \]

\[
\Gamma_{i} = p[Y = \omega, I_{0} = i] + \int_{y=0}^{\infty} e^{-by} \frac{d}{dy} F_{Y|I_{0}}(y, i) \, dy
\]

\[
- a \sum_{j=1}^{w} \int_{y=0}^{\infty} \int_{w=0}^{\infty} e^{-aw-\delta_{j}y} \Pr\{ W \geq w | (bvY = y, I_{0} = i) \} \cdot \\
\frac{d}{dy} F_{Y|I_{0}}(y, i) \, dw \, dy. \quad \text{as.s.}
\]

We have rewritten \( F_{W|Y_{I_{0}}}(y, i) \) as 1 -

\[ \Pr[W \geq w | Y = y, I_{0} = i]. \]

\[
\Gamma_{i} = E(e^{-\delta_{j}Y} \cdot I_{\{I_{0} = i\}}) - \\
\sum_{j=1}^{w} \int_{y=0}^{\infty} \int_{w=0}^{\infty} e^{-aw-(\delta_{j}z_{j})y}.
\]

\[ \cdot \Pr\{ \int_{0}^{y} g(I_{t}) \, dt \leq y | I_{0} = i \} \cdot a_{j} z_{j} (A_{j})_{i} dy \, dw \quad (6.2) \]

where we have changed the order of integration and used the definition of \( W_{t} \) and the explicit expression of the distribution \( F_{Y|I_{0}}(y, i) \).

For the sake of conciseness we define \( U(w) = \int_{0}^{w} g(I_{t}) \, dt \).
\[
\int_{w=0}^{\infty} \int_{y=0}^{\infty} e^{-\hat{\alpha}w - \hat{\beta}y} \Pr \left[ U(w) \leq y \mid (bvI_0 = i) \right] dy \, dw
\]

\[
= \int_{w=0}^{\infty} \int_{y=0}^{\infty} e^{-\hat{\alpha}w - \hat{\beta}y} \Pr \left[ U(w) \leq y \mid (bvI_0 = i) \right] dy \, dw
\]

since \( \Pr \left[ U(w) \leq y \mid (bvI_0 = i) \right] \) for \( w > 0 \),
is right continuous at \( y = 0 \)

\[
= \int_{w=0}^{\infty} e^{-\hat{\alpha}w} \left[ \int_{y=0}^{\infty} e^{-\hat{\beta}y} \Pr \left[ U(w) \leq y \mid (bvI_0 = i) \right] dy \right] dw \quad \text{a.s.}
\]

since \( \Pr \left[ U(w) \leq y \mid (bvI_0 = i) \right] \) is continuous

on \( y \in [0, \infty) \) for \( w > 0 \)

\[
= \int_{w=0}^{\infty} e^{-\hat{\alpha}w} E \left[ \frac{e^{-\hat{\beta}U(w)}}{\hat{\delta}} \mid (bvI_0 = i) \right] dw
\]

\[
= \int_{w=0}^{\infty} e^{-\hat{\alpha}w} E \left[ \frac{e^{-\hat{\beta}U(w)}}{\hat{\delta}} \mid (bvI_0 = i) \right] dw \quad \text{a.s.}
\]

since \( E \left[ \frac{e^{-\hat{\beta}U(w)}}{\hat{\delta}} \mid I_0 = i \right] \) is continuous

on \( \omega \in [0, \infty) \).

\[
R_i = \frac{1}{\hat{\delta}} \varphi_i (\hat{\alpha}, \hat{\beta}) \quad \text{.} \quad (6.3)
\]

Using (6.3) in (6.2) we get
\[ \Gamma_i = k_i - \hat{\alpha} \sum_{j=1}^{M} \omega_i(\hat{\alpha}, \delta - z_j) a_j \frac{z_j}{\delta - z_j} (g_j)_i \]

**Proof for Lemma 1.2**

This proof is due to Puri [30]. Visualize a binary valued random process \( Z_t \) taking values 0 and 1 with 0 being an absorbing state. The process \( Z_t \) is defined by

\[
\text{Pr} \left[ Z_{t+\Delta t} = 0 | (bV_t = 1), \{I_{t} = i(\tau), 0 \leq \tau < \infty\} \right] = \delta g(i(t)) \Delta t + O(\Delta t),
\]

where \( g(.) \) is positive. The vector process \((I_t, Z_t)\) is Markovian. \( Z_t \) is a death process for which the death rate is \( g(I_t) \). The probability of survival is written as

\[
\text{Pr}[Z_t = 1 | Z_0 = 1, \{I_\tau = i(\tau), 0 \leq \tau < \infty\}] = \exp \left( - \int_0^t g(i(\tau)) \, d\tau \right).
\]

Let

\[
\tilde{P}_{ik}(t) = \text{Pr}[I_t = k, Z_t = 1 | (bV_0 = 1, Z_0 = 1)].
\]

\[
= E \left[ \exp \left( - \int_0^t g(I_\tau) \, d\tau \right) I_{\{I_t = k\}} | I_0 = i, Z_0 = 1 \right].
\]

Let us define \( p_{ij} = m_{ji} \), the transition rates for the process \( I_t \).

The backward Chapman-Kolmogorov equations for the vector process \((I_t, Z_t)\) give us

\[
\tilde{P}_{ik}(w) = \delta_{ik} e^{-\delta g(i)w} + \sum_{j=0}^{N} \sum_{x=0}^{w} \left( p_{ij} - \delta g(i) \right) e^{-\delta g(i)x} p_{ij} \tilde{P}_{jk}(w-x) \, dx.
\]

Taking the Laplace transform of each side, we obtain
\[
(a - p_{ij} + \delta g(i)) \tilde{\pi}_{ik}(a) = \delta_{ik} + \sum_{j=0}^{N} p_{ij} \tilde{\pi}_{jk}(a),
\]
where \( \tilde{\pi}_{ik}(a) \) is the Laplace transform of \( \tilde{p}_{ik}(w) \). Let
\[
\tilde{\pi} = (\tilde{\pi}_{ik}),
\]
be an \((N+1) \times (N+1)\) matrix. We write the above equation for \( i = 0, N \) compactly in matrix notation as
\[
\left( \alpha I - M^T + \delta G \right) \tilde{\pi} = \tilde{I}.
\]
(6.4)
We may also choose to work with the forward Chapman-Kolmogorov equations, which yield
\[
\tilde{p}_{ik}(w) = \delta_{ik} e^{(p_{kk} - \delta g(k))w} + \sum_{j=0}^{N} \int_{0}^{w} e^{(p_{kk} - \delta g(k))x} p_{jk} \tilde{p}_{ij}(w-x) \, dx.
\]
Taking Laplace transforms, we get
\[
\tilde{\pi} \left( \alpha I - M^T + \delta G \right) = \tilde{I}.
\]
(6.5)
Puri [30] observes that for positive \( \alpha \) and \( \delta \) the matrix \( \alpha I - M^T + \delta G \) is diagonally dominant and its determinant is nonzero. Therefore the unique solution for \( \tilde{\pi}(a) \) in (6.4) and (6.5) is
\[
(\alpha I - M^T + \delta G)^{-1}.
\]
Recognizing that
\[
\Pi(a, \delta) = \tilde{\pi}(a) \times I
\]
establishes Lemma 1.2.
By relating $G$ to the scalars $z_i$ and vectors $\phi_i$ as per Theorem 2 below, we are able to invert the Laplace transform.

**Theorem 2**

We consider processes $I_t$, $Y_t$ and $W_t$ that satisfy the assumptions of Theorem 1. Let us use the terminology defined in Theorem 1. The matrix $G$ is constrained to satisfy the following equality

$$G = -D + rI,$$

for some constant $r$, and diagonal matrix $D$.

$(z_j, \phi_j)$ for $j, 0 \leq j \leq m$, are required to be eigenvalues and associated right eigenvectors of $zD \phi = M \phi$. We define

$$\hat{F}_i(w) = P[W \leq w, I_0 = i],$$

and

$$\hat{F}(w) = (\hat{F}_i(w)).$$

Under the above assumptions

$$\hat{F}(w) = \mathbb{F}(\infty) + \sum_{j=1}^{m} a_j e^{-z_j w} \phi_j.$$

**Proof of Theorem 2**

From Theorem 1 we have,

$$E[e^{-aW_{I_0=i}}] = \lim_{\delta \to 0} \mathbb{F}(0, \delta)$$

$$\mathbb{F}(0, \delta) = (E(I_{I_0 = i})) + a \sum_{j=1}^{m} a_j \left[ (a - rz_j) \mathbb{I} - (M - rz_j D) \right]^{-1} \cdot \phi_j$$

$$= \pi + a \sum_{j=1}^{m} \frac{a_j}{a - rz_j} \cdot \phi_j.$$
By treating this as a Laplace–Stieltjes transform and taking the inverse, we have our main result (we use the fact that $\mathbb{F}(\omega) = \pi$)

$$\hat{\mathbb{F}}(w) = \mathbb{F}(\omega) + \sum_{j=1}^{m} a_j e^{jw} \mathbb{H}_j$$

3. Application to Data Delay

Equation (6.1) defined the data delay for our integrated data-voice link. The stochastic processes $I_t$, $Y_t$, and $W_t$ satisfy the assumptions of Theorem 1 with $g(\cdot) = d(\cdot) \geq 0$. The Laplace transform of the asymptotic joint distribution of the number of ongoing conversations ($I_t$) and the data delay is obtained from Theorem 1. Since

$$d(i) = -(C_D - d(i)) + C_D$$

we may write the matrix $\mathbb{G} = -\mathbb{D} + C_D \mathbb{I}$ ($\mathbb{D}$ is defined in Chapter 5). The eigenvalues and vectors $(z_i, \mathbb{E}_i)$ are obtained from solving the eigenvalue problem in (5.4) and (5.5). Applying Theorem 2 we obtain the asymptotic distribution for data delay.

$$\hat{\mathbb{F}}(w) = \mathbb{F}(\omega) + \sum_{i=0}^{m-1} a_i e^{ziC_DW} \mathbb{E}_i,$$  \hspace{1cm} (6.6)

in the notation of Chapter 5.

4. Moments of the Data Delay

From (6.6) we obtain the asymptotic probability of the delay exceeding $w$
\[ G(w) = \lim_{t \to \infty} \Pr[ W_t > w ] = - \sum_{i=0}^{m-1} a_i e^{-z_i C_D w} \frac{1}{\lambda} T_{\tilde{e}_i} \cdot \frac{1}{\lambda} T_{\tilde{e}_i}. \]

We can write the asymptotic moments for the delay from (6.6) as

\[ \lim_{t \to \infty} E(W_t^j) = (-)^{j+1} \sum_{i=0}^{m-1} \frac{a_i}{(z_i C_D)^j} \frac{1}{\lambda} T_{\tilde{e}_i} \cdot \frac{1}{\lambda} T_{\tilde{e}_i}. \]

Note that ratio of the \( j \)'th moments of the buffer length and delay

\[ \lim_{t \to \infty} \frac{E(X_t^j)}{E(W_t^j)} = \frac{C_j^D}{C_D}. \]

For \( j = 1 \) the ratio of first moments is consistent

\[ \lim_{t \to \infty} \frac{E(X_t)}{E(W_t)} = \frac{C_1^D}{C_D} \]

with Little's result.
CHAPTER 7

Performance for Data and Conclusions

1. Performance Analysis

In Chapters 5 and 6 we gave explicit expressions for the asymptotic distributions of the buffer length and data delay as functions of external parameters, such as the total transmission capacity of the link, data arrival rate $C_D$, $\frac{\lambda}{\mu}$ ratio for voice, population of voice users $K$, the maximum number of allowable conversations on the link, $N$ and the functional relationship $d(t)$ between the number of ongoing conversations and the data transmission rate. In this chapter we use the two asymptotic distributions to explore the effects of data voice integration on data delay and buffer length.

The following assumptions were made in carrying out the studies reported in this chapter.

(1) The transmission capacity required for voice is directly proportional to the number of voice conversations in progress. ($r(i) = k \cdot i$, we select units for transmission capacity so $k=1$)

(2) The remaining transmission capacity is available for data. If there is any data in the buffer it is transmitted at the maximum possible rate, thus saturating the link ($d(i) = B - r(i) = B - i$).
(3) The data arrival rate is a constant \( C_D \).

Specifically we look at the behavior of the data statistics for a link with identical voice traffic but varying data traffic loads and for a link with fixed data traffic but varying voice traffic loads. A common voice link (T1 carrier) modified to incorporate data voice integration is the subject of the next study. The size of the link and the density of communications traffic has a direct impact on performance and this is examined at the end of this section.

First, we consider a communications link of 3.2 Mbs transmission capacity (\( B = 3.2 \text{ Mbs} \)) divided into 50 channels each of 64 Kbs transmission capacity. Each conversation (active voice source) requires the use of one channel \( (r(i) = 64 \cdot 1 \text{ Kbs}) \).

Figure 7.1 refers to a situation in which data is transmitted at a fixed rate \( C_D \), on the link. We have chosen three separate cases with \( C_D = 30.5, 32.5 \) and 34.5 channels. The link serves 150 voice customers, each attempting to make a 3 minute call approximately every half hour \( (K = 150, \frac{\lambda}{\mu} = 0.1) \). the number of channels usable by voice, \( N \), is varied. Figure 7.1 contains plots of the average, standard deviation and the 95th percentile of the data delay and buffer length as a function of \( N \), the number of voice channels. The rate at which data enters the buffer is \( C_D \), the maximum voice rate is \( r(N) \). When \( C_D + r(N) < B \) data experiences no delay because it is never obstructed by voice. When \( C_D + r(N) \) exceeds \( B \), the delay and the buffer length statistics initially increase rapidly, flattening out at about \( N = 25 \). This is because voice traffic rarely requires more
than 25 channels of transmission capacity as reflected by the blocking probability curve in Figure 4.1a.

Curves of the type shown in Figure 7.1 can be used for design purposes in situations in which we are constrained to use a specific link (B), carry a specific amount of data traffic (Cp) and service a given number of customers K, and assure a specific performance for data in terms of either the average, standard deviation and/or the 95th percentile of the delay and/or buffer length. The curves give us the maximum number of allowable voice conversations N. Referring to curves of the sort plotted in Figure 4.1a we can find the blocking probability.

The statistics in the flat portion of the curve can be computed by assuming that the blocking probability is 0. The eigenvalues and eigenvectors for the analysis in the absence of blocking can be found with much lesser computational difficulty than in the method outlined in this thesis by the procedure described in [26].

In Figure 7.2 we consider a fixed voice user population size (K = 150) serviced by a fixed number of voice channels (the cases N = 19, 21, and 23 will be examined) from an integrated data voice link (B = 50 channels, channel rate = 64 Kbs). Once again we assume each voice user attempts a call of average 3 minutes duration approximately every half hour (λ = 0.1). Each successful call uses up one 64 Kbs channel. The statistics for the data delay and the data buffer length are plotted as a function of the data arrival rate in Figure 7.2. The data suffers an exponentially increasing delay
penalty with increasing data traffic rate. A similar exponential increase of buffer length is also observed.

The following study helps determine the data rate that can be supported on an existing voice system when improvements are made in voice transmission. We consider a practical system, the Bell Systems T1 carrier, which usually carries 24 voice conversations (channels) at 64 Kbs each. Advances in the field of voice processing and transmission indicate that the voice rate may be reduced to 16 Kbs. The T1 may then be partitioned into 96 channels of 16 Kbs each. If only 24 of these reduced rate channels are used for voice (N=24), then we could transmit data on the remaining 72 channels without data suffering any delay.

Figure 7.3 describes the behavior of data delay and buffer when the data rate exceeds the rate that can be supported by 72 sub-channels. We examine this problem for a spectrum of voice population sizes (K = 24, 48, 72, 96, 120 and 150).

The following general remarks can be made from the T1-carrier studies reported in Figure 7.3. The 95th percentile is flatter near the origin than either the average or the standard deviation, but has a sharper knee and rises more quickly. In regions before the knee the 95th percentile is 0, indicating that most of the time data suffers no delay. The average and standard deviation of the delay distribution are non-zero and significant, however, indicating that when data experiences delay this delay is appreciable. Similar remarks can be made about the buffer length. The standard deviation
starts to rise before the average, but has a less pronounced knee and
does not rise as sharply as the average. Not only does the average
delay increase with the data rate but the spread of this delay dis-
tribution also increases. For this reason knowledge of the average
data delay as well as its standard deviation and 95th percentile is
required to measure performance for data. Tables 7.1 to 7.6 present
the delay statistics as a function of the number of channels of data
(each channel has a transmission capacity of 16 Kbs). In Table 7.7
we summarize the maximum number of extra channels (over 72) that we
can use for data for which the 95th percentiles of delay and buffer
length are still 0. Depending on the voice traffic load we see an
approximate savings between 4% and 30% of transmission capacity due
to integration.

Finally, let the channel capacity B, the number of circuits N,
the number of customers K and the rate of arrival of data traffic C_D
vary with the constraint that they bear a fixed ratio to each other.
By looking at the blocking probability and the data delay and buffer
statistics we can study the effect of scaling system size. Figure
7.4 shows the results of this study, using N as the representative of
system size. We talked about the variation of blocking probability
and voice throughput in Chapter 4.

The data buffer length statistics rise initially with system
size. The increase may be explained from the definition of the
buffer length
\[ Y_t = \int_0^t [(C_D + I_\tau - B)^+ \, dt - \int_0^t 1_{\{Y_\tau > 0\}} (C_D + I_\tau - B)^- \, dt]. \]

When we scale \( K, C_D, B \) and \( N \) upwards the constant terms in the integrand scale up, \( I_\tau \) can take a larger range of values permitting larger values for the buffer build up and depletion rates. Since the system handles proportionally greater traffic it is reasonable to expect an increase in buffer length statistics.

The subsequent decrease of the buffer length statistics occurs because the probability that \( C_D + I_\tau - B > 0 \) decreases with upward scaling. We carry out asymptotic analysis to establish this.

In Chapter 4 we found that when \( \frac{KA}{1+\lambda} > N \) voice traffic asymptotically saturates the transmission capacity allocated to voice and effective sharing of bandwidth between voice and data is not possible.

We consider the practical case \( \frac{KA}{1+\lambda} < N \). Under this condition the asymptotic voice throughput is \( \frac{KA}{1+\lambda} \). For asymptotic system stability we require \( B - C_D > \frac{KA}{1+\lambda} \).

We can now write

\[ \lim_{t \to \infty} \Pr[I_t > B-C_D] = \frac{\sum_{i=C+1}^{N} (\frac{K}{\lambda})^i}{\sum_{i=0}^{N} (\frac{K}{\lambda})^i}. \]

We set \( C = \lfloor B-C_D \rfloor - 1 \).
\[
\lim_{t \to \infty} \Pr[I_t > B - C_D] = 1 - \frac{\sum_{i=0}^{C} \binom{K}{i} \lambda^i}{\sum_{i=0}^{N} \binom{K}{i} \lambda^i}
\]

\[
= 1 - \frac{\frac{1}{B(K-N, N+1)} \int_0^\frac{1}{1+\lambda} Ke^{Kg(t)} dt}{\frac{1}{B(K-C, C+1)} \int_0^\frac{1}{1+\lambda} Ke^{K(h(t))} dt}.
\]

We let \(h(t) = (1-\alpha)\ln t + \alpha \ln (1-t)\) and \(g(t) = (1-\beta)\ln t + \beta \ln (1-t)\), where \(\alpha = \frac{N}{K}\) and \(\beta = \frac{C}{K}\). We used expressions for the summations from [19]. Since \(\frac{K\lambda}{1+\lambda}\) is smaller than both \(N\) and \(B - C_D\) (and hence \(C\)), the unique maxima for \(h(t)\) and \(g(t)\) lie in the interval of integration. Applying Laplace's method [20] to the integrals we obtain the asymptotic probability that data and voice compete for transmission capacity as \(N\) becomes large.

\[
\lim_{t \to \infty} \Pr[I_t > B - C_D] = 0.
\]

The peak of the standard deviation for buffer length is separated from the peaks for the average and 95th percentile. This separation remains unexplained.

The three statistics for data delay fall monotonically with increasing system size despite the initial increase in buffer length. This is because even though the buffer length increases the data transmission rate is being scaled up and hence the buffer gets cleared at a faster rate.
2. Conclusions

The principal contributions of this thesis are the development of a mathematical model of a data voice link and the derivation of closed form expressions for the blocking probability, the asymptotic data delay and the buffer length distributions.

When voice traffic generation is modeled by a Markov process and data traffic assumed to arrive at a constant rate for a link with fixed transmission capacity, with priority to voice, then the joint distribution of buffer length and voice traffic is governed by a system of partial differential equations (in time and buffer length) with constant coefficients. The asymptotic joint distribution is determined as the solution of a system of ordinary differential equations with constant coefficients.

The solution for the asymptotic distribution is a finite sum of exponential terms. To obtain these terms we solve an eigenvalue problem for a square matrix whose size equaled the number of states for the voice Markov process. The non-positive eigenvalues obtained are the exponential coefficients. The multiplicative coefficients are products of components of the right eigenvectors of the square matrix, a known function of the negative eigenvalues and the blocking probability.

We proved two general theorems for stochastic processes motivated by the need to find the asymptotic delay distribution. The first theorem gives the Laplace transform of the asymptotic delay distribution and the second gives the inverse of this transform. The
asymptotic delay distribution is identical to the asymptotic buffer length distribution except for the exponential coefficients which are obtained by scaling the exponential coefficients of the buffer length distribution by the data arrival rate. We conjecture that the two theorems will be applicable to stochastic processes modeling other phenomena.

Using the analytical results, we were able to compute the first moments, standard deviations and 95th percentiles of the two distributions. In doing so, we had to develop computer programs that generated the eigenvalues and eigenvectors efficiently. We used these statistics as the basis of our investigation into the effects of integration on data and voice transmission.

The expressions for the asymptotic state probabilities for the voice process are well known in the literature. Specifically we exhibited the formula for the asymptotic blocking probability – a statistic important for understanding the effects of integration. According to our model voice traffic is transmitted with pre-emptive priority. Hence the analysis for the voice process does not involve the data process.

When the number of allowable conversations increases the blocking probability falls to 0. We illustrated this fact with an example where voice service, with very low blocking probability, is provided to a voice customer population on the assumption that it was unlikely for more than 20% of the population to want service at the same time. We exhibited a relation between the blocking probability and the
voice throughput (average number of simultaneous conversations). An asymptotic analysis that considers a linear scaling of the number of customers and maximum allowable conversations reveals that depending on the choice of parameters the blocking probability either goes to 0 as the inverse of the square root of the scaling factor or it approaches a fixed non-zero value. In the latter case, the average asymptotic throughput approaches the maximum allowable number of conversations. This implies that asymptotically data will not be able to effectively share bandwidth with voice.

If, when the blocking probability asymptotically approaches 0, we let the total transmission capacity of the link and the data arrival rate increase linearly according to the same scale, the buffer length increases initially with upward scaling and then decreases as the size of the system continues to grow. The increase is attributable to the increase in the amount of data being handled by the link and the decrease to the decreasing probability of voice and data competing for transmission capacity with scaling.

The data delay decreases monotonically to 0. In the region where the buffer length decreases this is not surprising. In the initial region the scaling of transmission rate nullifies the effect of increasing buffer size.

We conclude that in large systems the data and voice traffic streams tend to be deterministic and the statistical variation in the transmission requirement of either stream is too small a fraction of the entire transmission capacity for an appreciable savings in
bandwidth.

We found some encouraging results for an intermediate range of transmission capacities. When the blocking probability for voice tends to 0 the delay statistics approach constant values. In the case where we are interested in the effects of variation in the maximum number of allowable voice conversations these constant values may be computed by assuming that there is no blocking of voice.

When we increase the amount of data traffic on a link with a fixed amount of voice traffic there is an exponential increase in the delay penalty for data and a similar increase in the buffer length statistics. We showed that we can save between 3% and 40% in transmission capacity on an integrated data–voice T1 carrier. The actual amount of savings is a function of the traffic intensities for data and voice.

The intermediate range of transmission capacities where integration results in savings of bandwidth corresponds closely to the capacities used for communication in small offices—a commercially viable market.

In the T1 example, even when the 95th percentile of the asymptotic delay and buffer length distribution is 0, the averages and the standard deviations may be appreciable. Hence the probability mass beyond the 95th percentile is diffused, with some small portion of the data suffering large delays. We suggest that data traffic be classified into two types (e.g. interactive and background) and transmitted on a priority basis. The higher priority class will not
be affected by the mass beyond the 95th percentile. The lower priority class will be subject to the large delays due to the mass in the tail of the distribution.
Delay  Buffer Length

\[ \lambda = 0.1 \]
\[ \frac{1}{\mu} = 180 \text{s} \]
\[ B = 50 \text{ ch.} \]

ch. rate = 64 Kbs

Data Performance Vs Maximum Conversations Allwd.

Figure 7.1
Data Performance Vs Data Rate

Figure 7.2
Data Performance for a T1 Carrier

Figure 7.3a
\[
\frac{\lambda}{\mu} = 0.1 \\
\frac{1}{\mu} = 180 \text{ s} \\
K = 48 \\
N = 24 \\
B = 96 \text{ ch.} \\
\text{ch. rate} = 16 \text{ Kbs}
\]

Data Performance for a T1 Carrier

Figure 7.3b
AVG.  

STD.  
DEV.  

95th  
PILE  

Delay  

Buffer Length  

\[
\frac{\lambda}{\mu} = 0.1 \\
\frac{1}{\mu} = 180 \text{s} \\
K = 72 \\
N = 24 \\
B = 96 \text{ ch.} \\
\text{ch. rate} = 16 \text{ Kbs}  
\]

Data Performance for a T1 Carrier  

Figure 7.3c
Data Performance for a T1 Carrier

Figure 7.3d
Data Performance for a T1 Carrier

Figure 7.3e
Data Performance for a T1 Carrier

Figure 7.3f
Effect of Scaling on Integration

Figure 7.4
<table>
<thead>
<tr>
<th></th>
<th>Performed Chns</th>
<th>88</th>
<th>89</th>
<th>90</th>
<th>91</th>
<th>92</th>
</tr>
</thead>
<tbody>
<tr>
<td>Delay (mS) Avg.</td>
<td>0.2</td>
<td>1.4</td>
<td>9.1</td>
<td>53.7</td>
<td>307</td>
<td></td>
</tr>
<tr>
<td>Delay (mS) Stdv.</td>
<td>15.7</td>
<td>54</td>
<td>142</td>
<td>408</td>
<td>1194</td>
<td></td>
</tr>
<tr>
<td>Delay 95%</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>5.8</td>
<td>1953</td>
<td></td>
</tr>
<tr>
<td>Buffer (Kb) Avg.</td>
<td>0.3</td>
<td>2.0</td>
<td>13.1</td>
<td>78.2</td>
<td>448.5</td>
<td></td>
</tr>
<tr>
<td>Buffer (Kb) Stdv.</td>
<td>22.1</td>
<td>88.8</td>
<td>205</td>
<td>598</td>
<td>1758</td>
<td></td>
</tr>
<tr>
<td>Buffer 95%</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>8.4</td>
<td>2878</td>
<td></td>
</tr>
</tbody>
</table>

K = 24  N = 24  λ = 0.1
Average Call Duration = 3 minutes
Transmission rate of each Channel = 16 Kbs
B = 96 channels

Performance Characteristics of a T1 Carrier
Table 7.1
<table>
<thead>
<tr>
<th># of data channels</th>
<th>84</th>
<th>85</th>
<th>88</th>
<th>87</th>
<th>88</th>
<th>89</th>
</tr>
</thead>
<tbody>
<tr>
<td>DELAY (mS)</td>
<td>0.4</td>
<td>1.5</td>
<td>5.8</td>
<td>26.0</td>
<td>97.3</td>
<td>354.8</td>
</tr>
<tr>
<td>Stdv.</td>
<td>24.3</td>
<td>57.0</td>
<td>130</td>
<td>293</td>
<td>879</td>
<td>1457</td>
</tr>
<tr>
<td>95%</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>218.7</td>
<td>2237</td>
</tr>
<tr>
<td>BUFSIZE (Kb)</td>
<td>0.5</td>
<td>2.1</td>
<td>9.1</td>
<td>36.2</td>
<td>137.0</td>
<td>505.3</td>
</tr>
<tr>
<td>Stdv.</td>
<td>32.7</td>
<td>77.5</td>
<td>179.5</td>
<td>407.7</td>
<td>916.2</td>
<td>2075</td>
</tr>
<tr>
<td>95%</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>304.9</td>
<td>3188</td>
</tr>
</tbody>
</table>

\[ K = 48 \quad N = 24 \quad \lambda = 0.1 \]

Average Call Duration = 3 minutes

Transmission rate of each Channel = 18 Kbs

B = 96 channels

Performance Characteristics of a T1 Carrier

Table 7.2
<table>
<thead>
<tr>
<th>Delay (ms)</th>
<th>Avg.</th>
<th>Stdv.</th>
<th>95%</th>
<th>Avg.</th>
<th>Stdv.</th>
<th>95%</th>
<th>Avg.</th>
<th>Stdv.</th>
<th>95%</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.3</td>
<td>22.8</td>
<td>0</td>
<td>1.0</td>
<td>47.5</td>
<td>0</td>
<td>3.5</td>
<td>97.0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>0.3</td>
<td>22.8</td>
<td>0</td>
<td>1.0</td>
<td>47.5</td>
<td>0</td>
<td>3.5</td>
<td>97.0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>0.3</td>
<td>22.8</td>
<td>0</td>
<td>1.0</td>
<td>47.5</td>
<td>0</td>
<td>3.5</td>
<td>97.0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>0.3</td>
<td>22.8</td>
<td>0</td>
<td>1.0</td>
<td>47.5</td>
<td>0</td>
<td>3.5</td>
<td>97.0</td>
<td>0</td>
</tr>
</tbody>
</table>

$K = 72 \quad N = 24 \quad \lambda = 0.1$

Average Call Duration = 3 minutes
Transmission rate of each Channel = 18 Kbs
$B = 96$ channels

Performance Characteristics of a T1 Carrier

Table 7.3
<table>
<thead>
<tr>
<th># of data channels</th>
<th>77</th>
<th>78</th>
<th>79</th>
<th>80</th>
<th>81</th>
<th>82</th>
<th>83</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Delay (mS)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Avg.</td>
<td>0.5</td>
<td>1.7</td>
<td>5.1</td>
<td>14.8</td>
<td>40.5</td>
<td>108.4</td>
<td>283.4</td>
</tr>
<tr>
<td>Stdv.</td>
<td>33.0</td>
<td>85.6</td>
<td>125.5</td>
<td>234.8</td>
<td>431.5</td>
<td>784.8</td>
<td>1424</td>
</tr>
<tr>
<td>95%</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>131.2</td>
<td>1474</td>
</tr>
<tr>
<td><strong>Buffer (kB)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Avg.</td>
<td>0.7</td>
<td>2.1</td>
<td>8.4</td>
<td>18.7</td>
<td>52.5</td>
<td>142.2</td>
<td>378.4</td>
</tr>
<tr>
<td>Stdv.</td>
<td>40.8</td>
<td>81.8</td>
<td>158.7</td>
<td>300.4</td>
<td>559.1</td>
<td>1030</td>
<td>1892</td>
</tr>
<tr>
<td>95%</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>172.1</td>
<td>1857</td>
</tr>
</tbody>
</table>

\[ K = 96 \quad N = 24 \quad \lambda = 0.1 \]

Average Call Duration = 3 minutes

Transmission rate of each Channel = 18 Kbs

B = 96 channels

Performance Characteristics of a T1 Carrier

Table 7.4
### Performance Characteristics of a T1 Carrier

**Table 7.5**

<table>
<thead>
<tr>
<th># of data chnnls</th>
<th>74</th>
<th>75</th>
<th>78</th>
<th>77</th>
<th>78</th>
<th>78</th>
<th>80</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>DELAY (mS)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Avg.</td>
<td>0.4</td>
<td>1.5</td>
<td>5.0</td>
<td>14.2</td>
<td>38.0</td>
<td>88.0</td>
<td>234.5</td>
</tr>
<tr>
<td>Stdv.</td>
<td>17.3</td>
<td>48.0</td>
<td>107.8</td>
<td>215.9</td>
<td>405.3</td>
<td>730.8</td>
<td>1288</td>
</tr>
<tr>
<td>95%</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>38.2</td>
<td>1005</td>
</tr>
<tr>
<td><strong>BUFFER (Kb)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Avg.</td>
<td>0.4</td>
<td>1.8</td>
<td>8.0</td>
<td>17.6</td>
<td>47.5</td>
<td>121.8</td>
<td>300.1</td>
</tr>
<tr>
<td>Stdv.</td>
<td>20.5</td>
<td>57.5</td>
<td>130.8</td>
<td>268.0</td>
<td>505.8</td>
<td>923.7</td>
<td>1649</td>
</tr>
<tr>
<td>95%</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>45.8</td>
<td>1287</td>
</tr>
</tbody>
</table>

**K = 120  N = 24  λ = 0.1**

Average Call Duration = 3 minutes

Transmission rate of each Channel = 18 Kbps

B = 96 channels
<table>
<thead>
<tr>
<th>Performance</th>
<th>73</th>
<th>74</th>
<th>75</th>
<th>76</th>
<th>77</th>
</tr>
</thead>
<tbody>
<tr>
<td>DEL (mS)</td>
<td>1.1</td>
<td>8.0</td>
<td>31.2</td>
<td>84.3</td>
<td>247.8</td>
</tr>
<tr>
<td>Stdv. 85%</td>
<td>20.2</td>
<td>93.5</td>
<td>282.1</td>
<td>588.0</td>
<td>1170</td>
</tr>
<tr>
<td>LAY</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>170.8</td>
<td>1417</td>
</tr>
<tr>
<td>BUFLN (Kb)</td>
<td>1.3</td>
<td>9.4</td>
<td>37.5</td>
<td>114.7</td>
<td>305.0</td>
</tr>
<tr>
<td>Stdv. 95%</td>
<td>23.8</td>
<td>110.8</td>
<td>314.7</td>
<td>714.8</td>
<td>1441</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>207.8</td>
<td>1748</td>
</tr>
</tbody>
</table>

\[ K = 150 \quad N = 24 \quad \lambda = 0.1 \]

Average Call Duration = 3 minutes

Transmission rate of each Channel = 18 Kbs

B = 98 channels

Performance Characteristics of a T1 Carrier

Table 7.8
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>24</td>
<td>18</td>
<td>9.1</td>
<td>142</td>
<td>13.1</td>
<td>205</td>
</tr>
<tr>
<td>48</td>
<td>15</td>
<td>28.0</td>
<td>293</td>
<td>36.2</td>
<td>408</td>
</tr>
<tr>
<td>72</td>
<td>12</td>
<td>36.5</td>
<td>384</td>
<td>48.9</td>
<td>384</td>
</tr>
<tr>
<td>96</td>
<td>9</td>
<td>40.5</td>
<td>432</td>
<td>52.5</td>
<td>558</td>
</tr>
<tr>
<td>120</td>
<td>6</td>
<td>38.0</td>
<td>405</td>
<td>47.5</td>
<td>506</td>
</tr>
<tr>
<td>150</td>
<td>3</td>
<td>31.2</td>
<td>282</td>
<td>37.5</td>
<td>315</td>
</tr>
</tbody>
</table>

Total Number of Channels = 96  
Channel Rate = 96 Kbs  
Number of Voice Channels = 24  
Number of Data Channels = 72 + Number of Extra Channels  
Bandwidth Savings on a T1 Carrier  
Table 7.7
CHAPTER 8

Extensions and Open Problems

1. Introduction

In previous chapters we modeled and analyzed the performance of the integrated data-voice link. Here we suggest some generalizations and extensions to the model proposed in Chapter 3 and investigate the effect of such modifications on the analysis of Chapters 4, 5 and 6.

In Chapter 3 we separated the modeling effort into three parts - data traffic model, voice traffic model and link model. We propose extensions to the model resulting from modification of each of these parts. These extensions change the governing equations for the asymptotic probabilities for the data process and lead to open problems.

2. Data Model Refinements

The main criticism of the data traffic model in Chapter 3 is that the data arrival rate is constant. In many practical situations the data traffic arrival rate varies with time. When this variation is significant the upper bounds obtained by replacing the time varying data rate by its time invariant maximum are not expected to be tight. Friedman [11] modeled this variation by assuming that data traffic is generated by the Poisson arrival of messages exponentially distributed in length.
Another way to model the variation in the data arrival rate is to assume that data traffic is generated by a single data source that turns on and off alternately—the on and off times being exponentially distributed with means $\frac{1}{\mu_2}$ and $\frac{1}{\lambda_2}$, respectively (we will use $\lambda_1$ and $\mu_1$ for the voice sources). When the data source is on it generates data that feeds into the data buffer at the rate $C_D$. $r(i)$ is the transmission capacity used up by $i$ simultaneous voice conversations. A non-empty buffer is depleted at the maximum possible rate $d(i) = B - r(i)$. $N$ is the maximum number of voice conversations possible simultaneously.

$Y_t$ is the length of the data buffer at time $t$. $I_t$ is the number of voice conversations. Let $V_t$ be a binary valued stochastic process that takes value 1 when the data source is on and 0 when it is off.

Formally

$$Y_t = \int_0^t [C_D - d(I_\tau)]^+ d\tau - \int_0^t I\{Y_\tau > 0\} [V_\tau C_D - d(I_\tau)]^- d\tau$$

We define

$$P_i^1(t, y) = \Pr[ Y_t < y, I_t = i, V_t = 1]; \quad P_i^0(t, y) = (P_i^1(t, y));$$

$$P_i^0(t, y) = \Pr[ Y_t < y, I_t = i, V_t = 0]; \quad P_i^1(t, y) = (P_i^0(t, y));$$

Observing events in the time interval $(t, t+\Delta t)$ we may write

$$P_i^1(t+\Delta t, y) = (K-(i-1))\lambda_1 P_i^{1}(t, y)\Delta t$$
\[ + (1 - \lambda_1 (K-i)\Delta t - i\mu_1 \Delta t - \mu_2 \Delta t)P_i^1(t,y-(C_d-d(i))\Delta t) \]

\[ + (i+1)\mu_1 P_{i+1}^1(t,y)\Delta t + \lambda_2 P_i^0(t,y)\Delta t. \]

\[ P_i^0(t+\Delta t,y) = (K-(i-1))\lambda_1 P_{i-1}^0(t,y)\Delta t \]

\[ + (1 - \lambda_1 (K-i)\Delta t - i\mu_1 \Delta t - \lambda_2 \Delta t)P_i^0(t,y+d(i)\Delta t) \]

\[ + (i+1)\mu_1 P_{i+1}^0(t,y)\Delta t + \mu_2 P_i^1(t,y)\Delta t. \]

Consider these equations for \(0 \leq i \leq N\), rearrange terms and take the limit as \(\Delta t \to 0\) to give

\[ \frac{\mathcal{D}}{\mathcal{D}t} P^1(t,y) + \frac{\mathcal{D}}{\mathcal{D}y} P^1(t,y) = (\mathbb{M} - \mu_2 \mathbb{I})P^1(t,y) + \lambda_2 P^0(t,y), \]

\[ \text{and} \]

\[ \frac{\mathcal{D}}{\mathcal{D}t} P^0(t,y) + \mathbb{D} \frac{\mathcal{D}}{\mathcal{D}y} P^0(t,y) = (\mathbb{M} - \lambda_2 \mathbb{I})P^0(t,y) + \mu_2 P^1(t,y). \]

\[ \mathbb{M} \text{ and } \mathbb{D} \text{ are defined in Chapter 5.} \]

Let \(\mathcal{P}(t,y) = \begin{bmatrix} P^0(t,y) \\ P^1(t,y) \end{bmatrix}\), then

\[ \frac{\mathcal{D}}{\mathcal{D}t} \mathcal{P}(t,y) + \mathbb{D} \frac{\mathcal{D}}{\mathcal{D}y} \mathcal{P}(t,y) = \mathbb{M} \mathcal{P}(t,y), \]

where \(\mathbb{D} \text{ and } \mathbb{M} \text{ are defined in Figure 8.1.} \)

If we let the random process \(I_t\) in Section 2, Chapter 5 stand for the vector process \((I_t, V_t)\) of this section then the problem of finding the asymptotic distribution of \(\mathcal{P}(t,y)\) satisfies the constraints of Section 2, Chapter 5 and the solution is a sum of exponentials. We need to solve the eigenvalue problem.
\[
\hat{z} \hat{D} \hat{A} = \hat{M} \hat{A}.
\]

Unfortunately, \( \hat{M} \) is no longer tridiagonal. By reordering the elements of \( \hat{P}(t, y) \) we can make \( \hat{M} \) banded with bandwidth 5, reducing the numerical work to solve the eigenvalue problem.

\( \hat{M} \) is the generator matrix for the Markov process \((I_t, V_t)\). We can define the waiting time \( W_t \) by

\[
W_t = w \text{ iff } \int_{t}^{t+w} g((I_{\tau}, V_{\tau}))d\tau = Y_t.
\]

where \( g((i, v)) = B - r(i) > 0 \). The theory in Section 6.2 can be applied to give the Laplace transform of the asymptotic delay distribution. Finding the inverse transform is an open problem.

The single data source on-off model described above is not good for modeling small variations in the data traffic arrival rate. To improve the model in this respect we can consider data traffic as being generated by \( S \) identical and independent sources. Each data source turns on and off alternately. The on and off times are exponentially distributed with means \( \frac{1}{\mu_2} \) and \( \frac{1}{\lambda_2} \), respectively. The Markov process \( J_t \) represents the number of data sources that are on at time \( t \) and takes values between 0 and \( S \) inclusive. The instantaneous data generation rate is \( c(J_t) \). \( I_t \) is the Markov process describing the number of on-going voice conversations at time \( t \). The buffer length \( Y_t \) is now defined as

\[
Y_t = \int_0^t \left[ c(J_{\tau}) + r(I_{\tau}) - B \right]^+d\tau
\]
\[ - \int_0^t \mathbb{1}_{\{Y_t > 0\}} \left[ c(J_t) + r(I_t) - B \right] dt. \]

The first term accounts for buffer build-ups and the second term for buffer depletions.

We define

\[ P_i^j = \Pr( Y_t < y, I_t = i, J_t = j ) ; \quad P_i^j(t,y) = (P_i^j(t,y)). \]

Observing events over the time interval \((t, t+\Delta t)\) we get

\[ P_i^j(t+\Delta t, y) = (K-(i-1))\eta_1 P_i^{i-1}(t,y) \Delta t \]

\[ + (1 - \eta_1(K-i)\Delta t - \mu_1\Delta t - \lambda_2(S-j)\Delta t - j\mu_2\Delta t)P_i^j(t, y - (C(j) - d(i))\Delta t) \]

\[ + (i+1)\mu_1 P_i^{i+1}(t,y) \Delta t + (S-(j-1))\lambda_2 P_i^{j-1}(t,y) \Delta t \]

\[ + (j+1)\mu_2 P_i^{j+1}(t,y) \Delta t. \]

Rearranging terms, dividing by \(\Delta t\) for such equations and taking the limit as \(\Delta t \to 0\) we get the following vector relation

\[ \frac{\delta}{\delta t} P^j(t,y) + D \frac{\delta}{\delta y^T} P^j(t,y) = (M - \lambda_2(S-j)I - j\mu_2 I) P^j(t,y) \]

\[ + (S-(j-1))\lambda_2 P_i^{j-1}(t,y) + (j+1)\mu_2 P_i^{j+1}, \]

where \(D = \text{diag}(C(j) - d(i))\).

Defining \( \hat{P}^T = \left( \left( \hat{P}^0 \right)^T \ldots \left( \hat{P}^j \right)^T \ldots \left( \hat{P}^S \right)^T \right) \) we can write

\[ \frac{\delta}{\delta t} P(t,y) + \hat{D} \frac{\delta}{\delta y} P(t,y) = \hat{M} \hat{P}(t,y), \]

where \(\hat{D}\) and \(\hat{M}\) are defined in Figure 8.2.
Using the theory in Chapter 5 we may express the buffer length distribution as a sum of exponentials. The theory in Section 2, Chapter 6 is used to evaluate the transform of the data delay. Inversion of this transform is an open problem.

One major shortcoming of the method outlined above is in finding the eigenvalues and eigenvectors of the problem

\[ \hat{z}^A \hat{D} \hat{\Psi} = \hat{M} \hat{\Psi}. \]

\( \hat{D} \) and \( \hat{M} \) are \((N+1) \times (M+1)\) matrices. The use of a numerical technique for finding the eigenvalues and vectors of such large matrices is non-trivial.

In all the data models that we considered so far the data traffic arrival process is independent of the delay experienced by previously transmitted data. In many real applications the data arrival rate fluctuates as a function of the anticipated or previously experienced delay. Let us go back to the model described in Chapter 3 and make the following modification. Since the length of the data queue is a good indication of the data delay let the data arrival process be a dependent on the data buffer process. The data arrival rate is \( c(Y_t) \). Defining

\[ P_i(t,y) = \text{Pr}[ Y_t < y, I_t = i ], \]

we may write

\[ P_i(t+\Delta t) = (K-(i-1))\lambda P_i(t,y) + (1-\lambda(K-i)\Delta t - i\mu\Delta t)P_i(t,y-(c(y)-d(i))\Delta t) \]

\[ + (i+1)\mu P_{i+1}(t,y). \]
From this we write

\[
\frac{\delta P}{\delta t} + D(y) \frac{\delta P}{\delta y} = M \cdot P
\]

where \( D(y) = \text{diag}(c(y) - d(i)) \). Hence

\[
\lim_{t \to \infty} P(t, y) = e^{\frac{\chi}{D} - 1(y) M \cdot dy}
\]

\[\lim_{t \to \infty} P(t, 0) = \lim_{t \to \infty} P(t, 0)\]

We have not investigated any further into this model.

3. Voice Model Refinement

In Chapter 3 voice traffic generation was modeled by a Markov process representing the number of on-going conversations. We may define a conversation to be the duration of an entire telephone call (thus modeling most intra-city telephone links) or the duration of a speech burst. When we choose to model voice at the level of speech bursts we assume \( K = N \) to be the number of on-going calls and \( I_t \) to be the number of speech bursts that are to be transmitted at time \( t \). A good discussion of the statistics of the speech burst process is available in [11]. The number of on-going conversations \( K \) is not a constant in time. Friedman's [11] treatment assumes that \( K \) is slowly varying with respect to speech burst durations. Under this assumption, we can carry out the analysis for each value of \( K \) separately, treating \( K \) as a constant.

In this section we propose that the random variable pair \((K_t, I_t)\) be used to denote the voice process. \( K_t \) is the number of on-going calls and \( I_t \) denotes the number of speech bursts to be transmitted at
time $t$. Speech bursts and the silent periods are exponentially distributed with means $\frac{1}{\mu_1}$ and $\frac{1}{\lambda_1}$, respectively. A two dimensional Markov chain may be drawn to represent the voice process. The theory discussed in Chapters 5 and 6 is unchanged except that $I_t$ is replaced by the vector process $(K_t, I_t)$. We can then find the asymptotic distribution for delay and buffer length. The size of the matrices involved in the eigenvalue problem is equal to the number of states for the voice process, which has increased by an order of magnitude over the voice model considered in Chapter 3. A reference to this approach can be found in [18, 8].

4. Link Model Refinements

In considering the data buffer to be a FIFO buffer we have restricted our analysis to links with only one grade of service. In this section we suggest consideration of two FIFO data buffers with one buffer being depicted at a higher priority. We consider two streams of data at the rates $C_{D1}$ and $C_{D2}$ feeding into buffers $B1$ and $B2$. Buffer $B2$ is depleted only when buffer $B1$ is empty. The definition and the analysis of the sum of the lengths of $B1$ and $B2$ is the same as the analysis in Chapter 5 and 6. We refrain from giving the formal definitions for the delay and for the buffer lengths for each buffer separately. Since the two grades of service may represent interactive and background file transfers, a practical situation, we consider it worthwhile to suggest the pursuit of this prioritized data types model.
\[ \hat{D} = \begin{bmatrix} D - C_D I & 0 \\ 0 & D \end{bmatrix} \]

\[ \hat{M} = \begin{bmatrix} M - \lambda_2 I & \mu_2 I \\ \lambda_2 I & M - \mu_2 I \end{bmatrix} \]

Figure 8.1
\[ \hat{D} = \begin{bmatrix} D^0 & D^1 \\ & D^S \end{bmatrix} \]

\[ \hat{M} = \begin{bmatrix} \lambda_2(S-j-1) \mathbb{I} & M - \lambda_2(S-j) \mathbb{I} - j \mu_2 \mathbb{I} & (j+1) \mu_2 \mathbb{I} \\ \vdots & \ddots & \ddots \\ \vdots & \ddots & \ddots \end{bmatrix} \]

Figure 8.2
Bibliography


[27] E. Gross, Bell Labs, Murray Hill, N.J., private communications.

